Learning Generalizable Visual Representations Towards Novel Viewpoints, Scenes and Vocabularies

Xiaoyu Zhu

CMU-LTI-24-019

December 2024

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15123

Thesis Committee:

Alexander Hauptmann (Chair)	Carnegie Mellon University
Teruko Mitamura	Carnegie Mellon University
Yonatan Bisk	Carnegie Mellon University
Junwei Liang	Hong Kong University of Science and Technology

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Language and Information Technology.

Keywords: Model Generalization, Representation Learning, Multimodal Learning, Visual Perception

Abstract

Deep learning has made significant progress to analyze an unprecedented amount of rich visual information from the real world to enable applications such as robotics, surveillance, and public safety monitoring. The successful deployment of deep learning techniques highly relies on the availability of large-scale domain-specific annotated data. However, these constraints are unlikely to be met in many real-world scenarios. In practice, various domain gaps exist between the training and test data. Test data are typically drawn from out-of-domain distributions, encompassing novel viewpoints, varied noise conditions, and diverse scenes. In addition to the diversity in visual representations, deep learning models trained on fixed, closed-set labels may not meet the query requirements of arbitrary text prompts from users. Additionally, novel vocabularies may not be accessible during training. To enable the deployment of a robust visual perception system, learning generalized feature representations during training is crucial.

In this thesis, with the goal of developing systems which can generalize to novel viewpoints, scenes and vocabularies, we explore different representation learning methods based on Siamese learning, masked visual modeling, and generatively pre-training. This thesis consists of three parts. The first part conducts robust semantic instance segmentation for videos and 3D data. We aim to learn feature representations that are invariant to various viewpoints and noise conditions via Siamese learning. We propose to leverage temporal consistency for videos and spatial consistency for 3D volumetric images, such that the learned feature representations have strong generalization ability. In the second part, we tackle the problem of human action analysis, which requires the model to learn from dynamic cues. We propose representation learning techniques based on masked visual modeling, such that the model can learn better spatial-temporal context. We also exploit both RGB videos and 3D human meshes for robust multi-modal action analysis. Finally, in the third part, we leverage generatively pre-trained vision-language models and develop systems that can handle novel vocabularies and text prompts. Our final goal is to build a robust system that can generalize to novel viewpoints, scenes, and vocabularies.

Thanks to my advisor, Alex, the best teacher in the world [165], who taught me how to face challenges in research and life. I couldn't be more fortunate to be Alex's student. Thanks to Junwei and Bernie, who patiently supervised me over the past years since I had very little knowledge about AI. Thanks to Teruko and Yonatan, for the valuable suggestions for this thesis and inspiration for long-term scientific goals. Thanks to my friends at CMU, Meta, Google and Argo AI, without whom I won't have been able to survive this journey. Thanks to my parents for supporting every decision I made.

Contents

Chapter 1

Introduction

1.1 Motivation of Research

Deep learning has made significant progress in analyzing an unprecedented amount of rich visual information from the real world. This progress enables applications such as robotics, surveillance, and public safety monitoring. The successful deployment of deep learning techniques highly relies on the availability of large-scale domain-specific annotated data [103, 332]. However, these constraints are unlikely to be met in many real-world scenarios. In practice, various domain gaps exist between the training and test data. Test data are typically drawn from out-of-domain distributions, encompassing novel viewpoints [159, 167, 349], varied noise conditions [224, 305, 348], and diverse scenes [350, 352]. In addition to the diversity in visual representations, deep learning models [62, 243] trained on fixed, closed-set labels may not meet the query requirements of arbitrary text prompts from users, and novel vocabularies may not be accessible during training. To enable the deployment of robust visual perception system, it is crucial for the system to learn generalized feature representations towards novel viewpoints, scenes and vocabularies during the training stage.

However, it is challenging to develop systems that can achieve robust performance given unexpected diversities in both vision and language spaces for the following reasons:

Collecting sufficient in-domain data sometimes is unfeasible. To perform scene understanding and human action analysis, it is necessary to collect sufficient and well-annotated data to train a robust system. However, collecting such data is not feasible in most cases, especially in complex, rare or violent scenarios.

Fine-grained annotations for videos and 3D data are costly. Comparing to images,

videos and 3D data involve an extra dimension, which makes the label annotation process time-consuming and expensive. Based on the published statistics [116, 204, 270], it takes about 114 s to annotate a 3D instance in a fully manual manner [270], and 30s if extra assistance of a 3D object detector is available [116]. For 3D biomedical images, annotating all structures on one tomogram takes about a month by a structural biology expert [348]. These extensive annotation efforts significantly hinder the performance of perception models.

The lack of paired text-video, text-3D data comparing to text-image data. There are large-scale image-caption pairs available in the Internet to train vision-language foundation models. For example, LAION-5B is a dataset of 5,85 billion CLIP-filtered image-text pairs [252]. The largest dataset with paired text and 3D data is Objaverse-XL [61], which contains around 10M 3D Objects. The paired text-3D dataset is 500x smaller than the text-image dataset.

Considering the aforementioned challenges, we investigate how to design effective models that can learn generalizable feature representations in both vision and language space. Our path is orthogonal to simply scaling the data and model size. The first part of this thesis conducts robust semantic instance segmentation for videos and 3D data. We aim to learn feature representations that are invariant to various viewpoints and noise conditions via Siamese learning. We propose to leverage temporal consistency for videos and spatial consistency for 3D volumetric images, such that the learned feature representations have strong generalization ability. In the second part, we tackle the problem of human action analysis, which requires the model to learn from dynamic cues. We propose representation learning techniques based on masked visual modeling, such that the model can learn better spatial-temporal context. We also exploit both RGB videos and 3D human meshes for robust multi-modal action analysis. Finally, in the third part, we leverage generatively pre-trained vision-language models and develop systems that can handle novel vocabularies and text prompts. Our final goal is to build a robust system that can generalize to novel viewpoints, scenes, and vocabularies.

1.2 Thesis Overview

In this thesis, with the goal of developing systems which can generalize to to novel viewpoints, scenes and vocabularies, we explore different representation learning methods based on Siamese learning, masked visual modeling, and generatively pre-training. A detailed overview of each part is as follows:

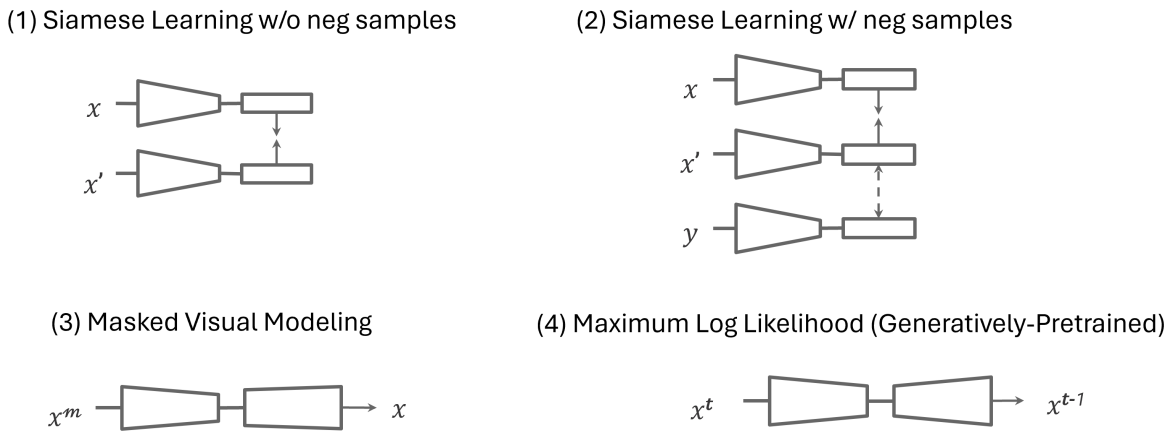

Figure 1.1: Overview of our representation learning methods. We explore four typical representation learning methods: (1) Siamese learning without negative samples; it forces the model to make consistent predictions between the input x and its positive view x'; (2) Siamese learning with both positive and negative samples; it forces the model to learn representations that will minimize the feature distance between x and its positive view x', and maximize the distance between x and the negative sample y; (3) masked visual modeling; it incorporates masked views x' of the input x into the model training. We explore two training objectives: (3a) mask and reconstruction; and (3b) mask and consistency prediction. (4) maximum log likelihood; it is used for text-to-image generative pre-training to learn unified representations for vision and language.

Part I Siamese Learning for Robust Semantic Instance Segmentation In this part, we aim at developing robust systems based on Siamese learning to understand "things" in videos and 3D volumetric images. We propose a model named *MSNet* to perform viewpoint-invariant instance segmentation in aerial videos (chapter 2). For 3D volumetric images, we force the model to learn spatially-consistent representation which are robust to variant noise conditions (chapter 3).

Part II Masked Visual Modeling for Generalized Human Action Analysis In this part, we focus on analyzing human behavior from temporal cues based masked visual modeling. We explore different modalities for human action analysis, including videos, 3D skeletons, point clouds, and meshes. We first propose to leverage adversarially masked consistency for scene-invariant action recognition (chapter 4). We then propose a masked vertex modeling technique for 3D mesh-based action recognition (chapter 5). Finally, we conduct generalized human action

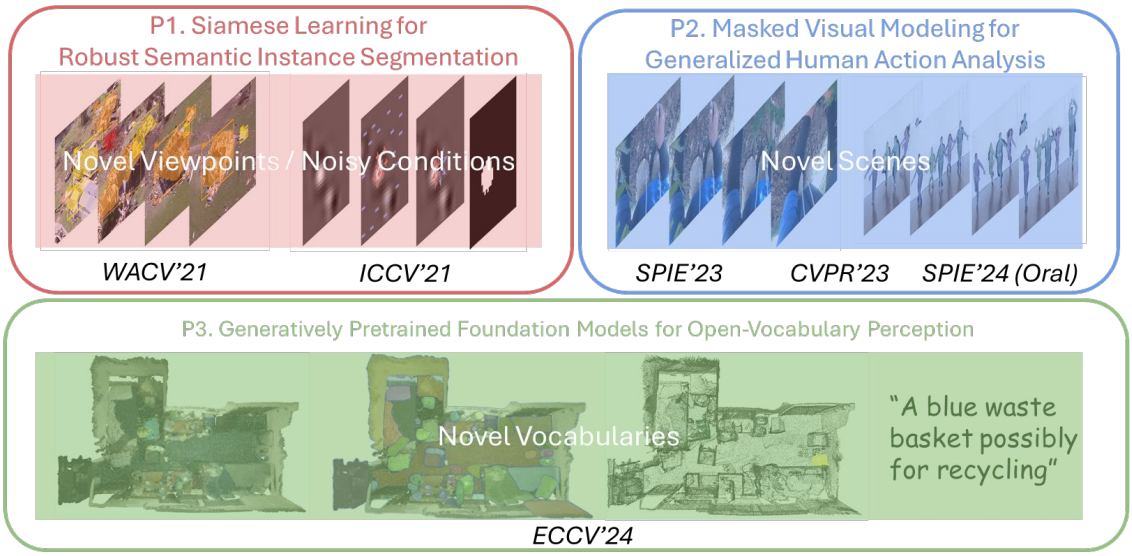

Figure 1.2: Thesis roadmap: Generalizable visual representation learning for novel viewpoints, scenes and vocabularies.

recognition by jointly modeling videos and 3D meshes (chapter 6).

Part III Generatively Pretrained Foundation Models for Open-Vocabulary Perception
In the last part, we leverage generatively pre-trained vision-language models and develop systems that can handle novel vocabularies and text prompts. We evaluate the model on the open-vocabulary 3D scene understanding tasks including 3D semantic segmentation and visual grounding (chapter 7).

1.3 Thesis Contributions

The study in this thesis demonstrates the effectiveness of representation learning techniques for increasing generalization of deep learning models. The specific findings are as follows:

Siamese learning leads to more generalized representations. Forcing the model to make consistent predictions across the temporal [349] (chapter 2) and spatial [348] (chapter 3) domains leads to more generalized and robust representations.

Masked visual modeling learns scene-invariant representations. By incorporate masked visual modeling into the model design, the model is able to learn scene-invariant representations in both "mask and consistent learning" [352] (chapter 4) and "mask and reconstruct" [351]

(chapter 5).

The 2D and 3D representations are complementary to each other, even when the 3D representations are noisy estimations. We prove that noisy 3D body pose estimations are helpful for domain-invariant representations learning in videos [350, 353] (chapter 6).

Generatively pretrained multi-modal representations are beneficial for visual perception. We aim at leveraging text-to-image diffusion models for open-vocabulary perception task [354] (chapter 7). This demonstrates the effectiveness of generatively-pretrained models are also beneficial for the perception task.

Part I

Siamese Learning for Robust Semantic Instance Segmentation

In this part, we aim at developing robust systems based on Siamese learning to understand "things" in videos and 3D volumetric images. We propose a model named *MSNet* to perform viewpoint-invariant instance segmentation in aerial videos. For 3D volumetric images, we force the model to learn spatially-consistent representation which are robust to variant noise conditions (chapter 3).

Chapter 2

Temporal consistency learning for video instance segmentation

In this chapter, we explore the benefit of temporal consistency for viewpoint-invaraint feature representation learning, and the model could perform well on instance segmentation task with sharp viewpoint changes.

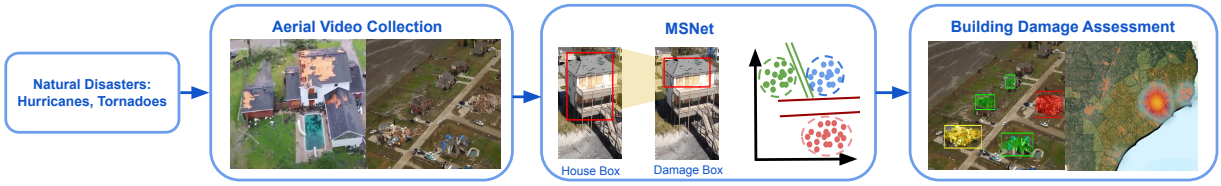

Figure 2.1: Illustration of the natural disaster damage assessment pipeline. Aftermaths of natural disasters are recorded by drones. Our model is able to detect damage masks and damage scales in different locations. The damage detections along with drones' GPS trajectory could generate a damage assessment location heatmap to aid timely disaster relief efforts.

2.1 Overview

In recent years, natural disasters have impacted many vulnerable areas around the world. In 2019, there have been ten natural disaster events with damages of more than 1 billion dollars each across the United States [81]. Timely response to natural disasters plays a crucial role

in disaster relief. However, current damage assessments are mostly based on manual damage detection and documentation, which is slow, expensive and labor-intensive work [217].

With the increasing availability of consumer-grade drones, a large number of aerial videos are recorded and shared across social media [166]. After a natural disaster, like a hurricane or a flood, people frequently share drone footage of the district, or the authorities could dispatch drones themselves to assess the damage of the area. These videos could serve as valuable resources for automatic damage assessment. Compared with satellite imagery used in previous damage assessment task works [39, 95, 240], drone videos have the advantage of capturing detailed observations of each building from different angles other than just from a top-down perspective. Valuable structural information of the buildings could be extracted from drone videos for further damage evaluation, *i.e.*, whether the buildings are going to collapse.

Consider the example in Figure 7.2, there are three challenges for automatic building damage assessment. The first is the diversity of buildings, the level of damages and the location of damages. Buildings could include homes, schools, coastal buildings, factories, and other facilities. Some might be slightly damaged, and others might be completely damaged. Some might only have severe damage on the roof. The second challenge is the detection of small objects and debris. The drone videos are usually recorded from a high altitude where many of the damaged parts are only represented by a few dozen pixels (See Section 2.3). The third challenge is the changes of viewpoints as the drone flies over the area. The damage of a building might only be visible from a certain viewpoint. This leads to problems like missed detection and inconsistent detections by a single image-based detector.

To overcome the aforementioned challenges, we have collected the first dataset with aerial videos for natural disaster damage assessment. Our dataset, namely ISBDA (Instance Segmentation in Building Damage Assessment), consists of fine-grained building damage bounding box and mask annotations of different damage levels. This provides the first quantitative benchmark for evaluating building damage assessment models. Our second contribution is to propose a new neural network model, *MSNet*, to address the difficulties of accurately detecting damages in buildings with aerial videos. Our model makes use of the hierarchical relationship between building and damage, and inter-frame spatial consistency of multiple viewpoints to train more robust representations. To summarize, our contribution is fourfold:

- We present the first natural disaster building damage assessment dataset, namely ISBDA, using aerial drone videos. It is annotated with fine-grained instance-level building and damage bounding boxes and masks. It provides the first quantitative benchmark for assessing damage assessment in aerial videos.

- We propose a novel neural model termed Hierarchical Region Proposal Network (HRPN), which explores the hierarchical spatial relationship among different objects, and thus significantly improving the model performance.
- We propose an unsupervised score refinement model named Score Refinement Network (SRN) based on inter-frame consistency to tackle the challenges of detections using drone videos.
- We empirically validate our model on the proposed ISBDA dataset for damage assessment, in which our model achieves the best results compared to state-of-the-art object detection models.

2.2 Related Work

Natural Disaster Damage Assessment Datasets. Existing damage assessment dataset can be roughly categorized into two types: ground-level images and satellite imagery. The ground-level images were mostly collected from social media [215]. Those datasets only have image-level labels available, because the scene captured by a single ground-level image is highly limited. Besides, due to the lack of geo-tags in social media, ground-level images may not be suitable for large-scale damage assessment. Another disaster data source is satellite imagery based on remote sensing [39, 95, 128, 240, 246]. However, the main limitation of satellite imagery is that it could not provide detailed damage information due to the long distance to the captured buildings and its limited vertical viewpoint. We are the first to propose a dataset from drone video viewpoints (typically about forty-five degrees) for damage assessment tasks with instance-level damage annotations.

Damage Detection Approaches. Current damage detection approaches can be put into three categories. The first category is using supervised machine learning methods which include pixel-based relevant change detection [27] and object-based local descriptors [291]. The second category includes unsupervised methods [91, 206, 216] that generally refer to outlier detection in scene changes. The third category, a recent trend on damage assessment is using semi-supervised approaches [92] aimed at using less human-labeled data and maintaining higher accuracy. Other literature also proposed deep learning frameworks such as Convolutional Neural Networks (CNN) [8, 215] to predict the damage level of each image. However, existing models only worked on building bounding box prediction tasks, which lack specific

locations of damaged parts.

Anchor-based Region Proposal Networks. Existing literature on anchor-based region proposal networks mostly adopted dense anchoring scheme, where anchors are sampled densely over the spatial feature space with predefined scales and aspect ratios. The most representative work is Region Proposal Network (RPN) introduced in Faster R-CNN [238], which designed a light fully convolutional network to map sliding windows to a low-dimensional feature space. This framework has been widely adopted in later research [54, 101]. Some research [324] focused on using meta-learning to dynamically generate anchors from the arbitrary customized prior boxes. Other research works [26, 36, 335] adopted cascade architecture to regress bounding boxes iteratively for progressive anchor refinement. Some researchers [301] tried to remove the iteration process by predicting the center of objects of interest. However, there is still a lack of region proposal networks that could utilize spatial hierarchical relationships among objects which could potentially improve detection accuracy.

Figure 2.2: Visualization of our ISBDA dataset. The green, yellow and red polygons denote damages in Slight, Severe and Debris levels, respectively. The rectangles composed of solid lines represent damaged building bounding boxes. The polygons with dotted lines represent segmentation masks of damaged parts.

Detection Score Refinement. Current research in detection score refinement can be categorized into two streams, bounding box score refinement and mask score refinement. In bounding box score correction, most works focused on making modifications on the basis of Non-maximum Suppression (NMS) algorithm, such as Fitness NMS [289] and SoftNMS [21]. Jiang *et al.* [124] proposed IoU-Net that directly predicted box IoU, and the predicted IoU was used for the bounding boxes refinement. In terms of score refinement in mask level, Mask Scoring

R-CNN [118] was proposed by adding a MaskIoU head to regress the IoU between the predicted mask and its ground truth mask. One limitation of this approach is that it can only refine the mask scores, which nearly has no impact on the bounding box branch. Our proposed score refinement algorithm based on inter-frame consistency is able to achieve consistent improvement in both bounding box and mask branches.

2.3 The ISBDA Dataset

2.3.1 Data Collection

In order to fully assess building damages in different scenarios and locations, we have collected ten videos from social media platforms, which recorded severe hurricane and tornado disaster aftermaths in recent years. Specifically, the aerial videos were recorded after Hurricane Harvey in 2017, Hurricane Micheal and Hurricane Florence in 2018 and other three tornadoes (EF-2 or EF-3) in 2017, 2018 and 2019, respectively. The affected areas recorded in the videos include Florida, Missouri, Illinois, Texas, Alabama and North Carolina in the United States. The total length of the collected videos is about 84 minutes.

To get individual frames, we first obtain video clips from the ten videos that: (1) do not have apparent camera rotations; and (2) fly with moderate and stable speed. To further improve the annotation efficiency and cover different scenarios, we extract one frame out of every ten frames from these video clips. Overall, we have collected 1,030 frames for instance-level building and damage annotation.

One important problem is to define damage scale and corresponding standards which can cover various types of damages in different scenes. Following the damage assessment practice, Joint Damage Scale [95], we divide building damages into three levels: Slight, Severe and Debris. Slight refers to visible cracks or appearance damages. Severe refers to partial wall or roof collapse, which are apparent structural damages. Debris refers to completely collapsed buildings.

2.3.2 Hierarchical Instance-level Annotation

To provide fine-grained localization information of individual damages, we formulate the damage assessment task as an instance segmentation problem. We annotate both the polygons of damaged buildings and the specific damaged parts of the buildings. In order to explore the hier-

archical relationships between building and damaged part instances (*i.e.*, specific damaged parts are within corresponding damaged building boxes), we also include the mappings between each damaged part ID and its corresponding damaged building ID. The dataset is annotated by three experienced annotators, and one pass of verification is performed for each annotation to ensure accuracy.

2.3.3 Dataset Statistics

Overall, 1,030 images sampled from 10 videos are annotated with instance-level building masks and damaged part masks. The dataset has 2,961 damaged part instances which are divided into three levels: Slight, Severe, and Debris. Following Microsoft COCO's [172] size definition, we calculate the number of damaged part instances in different sizes for each damage scale, shown in Table 2.1.

Damage Scale	Small	Medium	Large	Total
Slight	204	1169	746	2119
Severe	-	120	440	560
Debris	-	54	228	282

Table 2.1: Distribution of annotation sizes. Small: area less than 32×32; Medium: area greater than 32×32 and less than 96×96; Large: area greater than 96×96. Area is measured as the number of pixels in the segmentation mask.

We also analyze the distribution of the area of damage segmentation in the ISBDA dataset, shown in Figure 2.3. We observe that the majority of the damage segmentation are relatively small. Visualization of the ISBDA dataset and annotations is shown in Figure 2.2.

2.4 Method

2.4.1 Overview

To provide fine-grained localization information, similar to some of the existing works [95], we formulate the damage assessment task as an instance segmentation problem. Moreover, our model will predict damage-level instance masks instead of building-level, which is a more challenging task due to the high damage variance and small damaged area. We propose a new

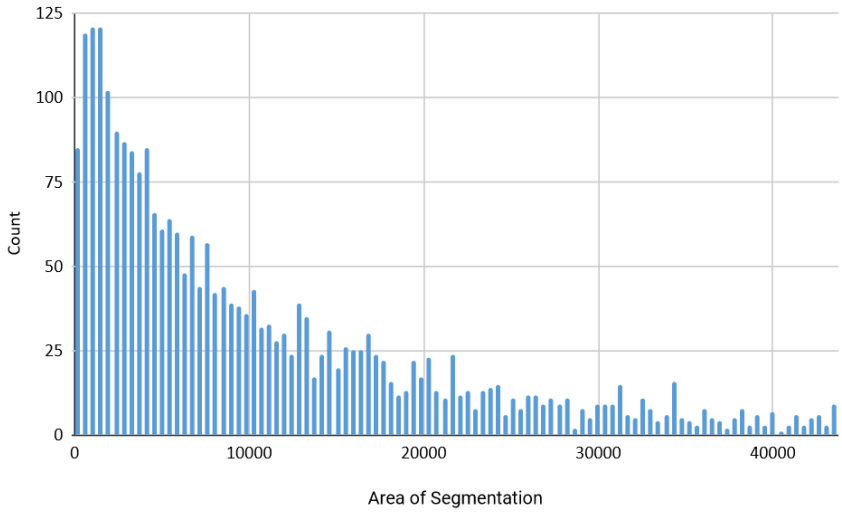

Figure 2.3: The distribution of the area of damage segmentation in our ISBDA dataset. We only show the distribution of areas below 90th percentile of the whole dataset for better visualization purpose. Area is measured as the number of pixels in the segmentation mask.

model named *MSNet* in order to learn more robust representations in different scenarios with different viewpoints. It includes two types of supervision: supervision of building bounding boxes for low-level damage anchor sampling and mask segmentation; and supervision of temporal and spatial relationships between adjacent video frames. In summary, it has the following key components:

Pyramid Backbone Network uses ResNet-50 based Feature Pyramid Network (FPN) [173] to extract spatial features of input images.

Hierarchical Region Proposal Network first generates high-level building proposals and then uses them to supervise low-level anchor sampling and damage proposals generation.

Score Refinement Network is proposed to calibrate the confidence scores of instances in adjacent frames which share common appearance features but have confidence score variances.

Mask R-CNN Head includes the R-CNN head for bounding box and class prediction, and the Mask head for mask prediction [101].

In the rest of this section, we will introduce the above components and the learning objectives in details.

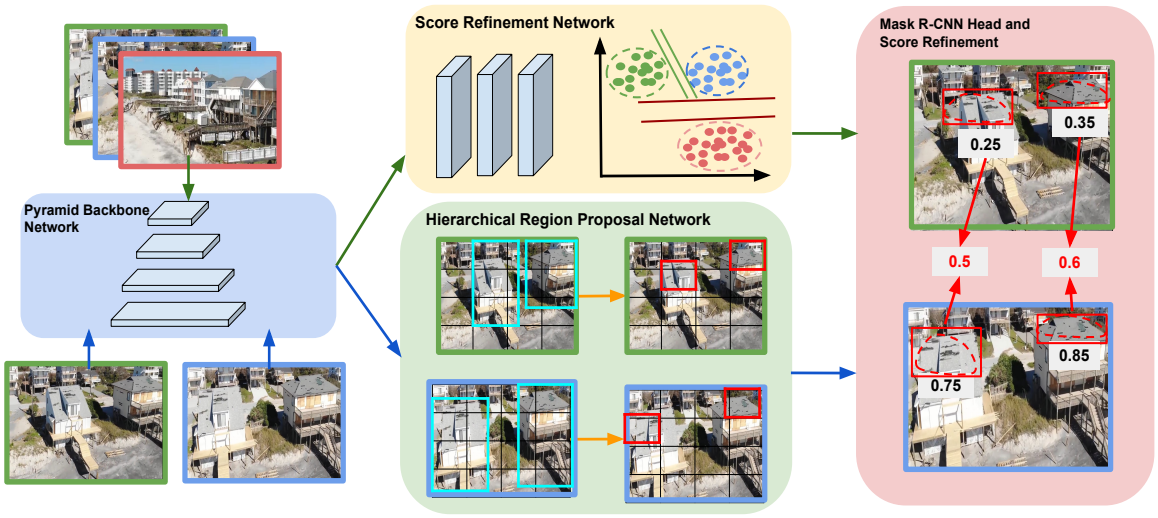

Figure 2.4: Network architecture of *MSNet*. The left part contains a pyramid backbone network to extract features in multi-scale levels. The backbone network is shared in the two neural network's training. The first neural network (Bottom) is for generating instance segmentation results. Specifically, for each image, Hierarchical Region Proposal Network takes the encoded features to generate proposals for damaged buildings. The building proposals are used to give supervision on damage proposals generation (Yellow Arrow). The second branch (Top) is for the training of Score Refinement Network. The adjacent frames (images with green and blue edges) along with one negative sample (image with red edges) are firstly fed into the Pyramid Backbone Network, then Score Refinement Network is trained with the proposed Multi-scale Consistency Loss to learn feature similarity. These two branches are joined at the end, where Mask R-CNN Head generates bounding box and mask predictions. Finally, the score refinement algorithm is performed to calibrate the confidence scores.

2.4.2 Hierarchical Region Proposal Network

Traditional Region Proposal Network (RPN) treats all objects in the same spatial level, and uniformly generates dense anchors over the feature space. If we adopt a conventional RPN scheme and train the RPN with building and damage proposals simultaneously, the hierarchical relationship between buildings and damaged parts will not be utilized. Therefore, we propose a new model, termed Hierarchical Region Proposal Network (HRPN), to address the aforementioned problems.

In HRPN, there are two RPNs sharing the same backbone network: a high-level RPN and a low-level RPN. The high-level RPN is trained with damaged building boxes with binary la-

bels indicating whether the proposal is a damaged building or not. The low-level RPN utilizes building proposal outputs from the high-level RPN for anchor sampling. We sample anchors based on one of the two metrics: Intersection over Union (IoU) and Inner Intersection (II) between high-level region proposals and low-level anchors. For each low-level low-level (damage) anchor $A_{\widetilde{a}}$, we define its sampling score as:

$$\mathrm{S}_{IoU}(A_{\widetilde{a}}, A_p) = \max_{A_p \in P} \frac{A_{\widetilde{a}} \bigcap A_p}{A_{\widetilde{a}} \bigcup A_p} \tag{2.1}$$

$$\mathrm{S}_{II}(A_{\widetilde{a}}, A_p) = \max_{A_p \in P} \frac{A_{\widetilde{a}} \bigcap A_p}{A_{\widetilde{a}}} \tag{2.2}$$

where P is a set of high-level (building) region proposals. For each anchor, we compute its sampling score and only keep anchors with scores larger than a certain threshold S. Then the sampled anchors are used for damage proposals generation.

2.4.3 Score Refinement Network

In previous works [22], the confidence scores are determined by single-frame detection, while correspondence between two adjacent frames is not utilized. We propose a score refinement model based on inter-frame temporal and spatial correspondence termed Score Refinement Network (SRN). The input of the model is randomly generated triplets and each triplet is composed of one frame and its adjacent frame as a positive frame and another random frame as a negative frame. By incorporating multi-scale features from the FPN backbone, we design a multi-scale consistency loss to force SRN to learn feature representations such that one sample's distance to its positive sample is closer than its distance to the negative one. We aim to refine the scores of instances in adjacent frames which share common appearance features but have confidence score variances.

Inspired by [307], we use patch mining to build triplets and each is composed of one sample P_i, its relative adjacent frame P_i^+ and its random sample P_i^-. The triplets are sampled based on the fact that the average drone speed is 50 mph and thus the frame variances within half seconds are small. Therefore, given a frame x_t at time t and the video frame rate r, the positive sample is defined as the frame in range $[x_t - 0.5r, x_t + 0.5r]$. The negative sample is defined as the frame in range $[0, x_t - 10r] \bigcup [x_t + 10r, T]$. T is the maximum frame number of the video.

Multi-scale features usually demonstrate significant performance improvement in object detection tasks [101, 173]. Therefore, we propose Multi-scale Consistency Loss (MCL) which makes use of multi-scale feature maps. For two image patches X_i, X_j, we firstly obtain the

feature maps of each image from the last four layers of the FPN backbone, namely P_{ik}, P_{jk}, where $k \in [1, 2, 3, 4]$. These feature maps are used as input to SRN. For an input feature P, we can obtain its feature from the last SRN layer as $f(P)$, where f is a feature encoder which is composed of three fully connected layers. Then, we propose a spatial-wise similarity metric of two feature maps P_{ik}, P_{jk} in FPN level k using:

$$Sim(P_{ik}, P_{jk}) = \sum_{w=0}^{W} \sum_{h=0}^{H} \frac{f(P_{ik}^{wh}) \cdot f(P_{jk}^{wh})}{\|f(P_{ik}^{wh})\| \|f(P_{jk}^{wh})\|} \tag{2.3}$$

$$D(P_{ik}, P_{jk}) = 1 - Sim(P_{ik}, P_{jk}) \tag{2.4}$$

Given a set of triplets and each triplet is denoted as (X, X^+, X^-) , we aim to train SRN which can learn feature representations such that $D(X, X^-) > D(X, X^+)$ using the Multi-scale Consistency Loss (MCL):

$$\mathcal{L}_{mcl}(X, X^+, X^-) = \sum_{i=1}^{L} \max\{0, D(X_i, X_i^+) - D(X_i, X_i^-) + m\} \tag{2.5}$$

where m is a margin constraint parameter, and L is the number of multi-scale layers.

2.4.4 Training

In this section, we provide detailed descriptions of the training procedure. The first part of the loss function is the HRPN loss, which is defined as:

$$\mathcal{L}_{hrpn} = \mathcal{L}_{rpn}^{h} + \mathcal{L}_{rpn}^{l}. \tag{2.6}$$

Here, $\mathcal{L}_{rpn}^{h}$ and $\mathcal{L}_{rpn}^{l}$ represent the loss of high-level RPN and low-level RPN, respectively. The low-level RPN conducts anchor sampling and proposal generation under the supervision of high-level RPN. As described in Section 2.4.2, the losses of damage proposals which are filtered out under the supervision of high-level building proposals are not computed in the HRPN loss. The definition of RPN loss follows [238]. $\mathcal{L}_{cls}$, $\mathcal{L}_{box}$, and $\mathcal{L}_{mask}$ follow the definitions in [101]. $\mathcal{L}_{mcl}$ is computed using Equation 2.4.3.

The final multi-task loss of our proposed approach is calculated using:

$$\mathcal{L} = \mathcal{L}_{hrpn} + \mathcal{L}_{cls} + \mathcal{L}_{box} + \mathcal{L}_{mask} + \mathcal{L}_{mcl}. \tag{2.7}$$

Method	AP	AP_{25}	AP_{50}	AP^{bb}	AP^{bb}_{25}	AP^{bb}_{50}
PolarMask+D	22.3	29.1	15.4	24.4	29.6	18.2
Mask R-CNN+D	34.4	40.6	26.9	35.9	40.9	29.4
Mask R-CNN+B+D	32.2	39.5	23.3	34.0	40.3	25.7
Ours	**37.2**	**44.2**	**28.8**	**38.7**	**44.4**	**31.5**

Table 2.2: Cross scene evaluation results. We report detection and instance segmentation results. AP denotes instance segmentation results and AP^{bb} denotes bounding box detection results. In the results area, rows 1 and row 2 use the PolarMask and Mask R-CNN frameworks with only damage masks (D) as input; row 3 uses Mask R-CNN co-trained with damaged buildings (B) and damages (D) as the baseline model. The results show that our proposed method gains significant improvements compared to state-of-the-art models.

The HRPN and Mask R-CNN Head can be trained end-to-end together with SRN. However, in that case, the model training and inference would be heavy due to the multi-scale feature similarity calculation. Therefore, we only calibrate confidence scores of the model which has the best instance segmentation performance.

2.4.5 Inference

In test time, we use HRPN to generate building region proposals. Then the building proposals are used as supervision for damage anchor sampling and proposal generation, as described in Section 2.4.2. In the second stage, the model extracts features using RoIAlign for each damage proposal and performs proposal classification, bounding box regression and mask prediction.

During the inference of SRN, given two adjacent frames P and Q, we firstly extract the last four layers from the Pyramid Backbone Network for each frame. The four layers are used as input for SRN described in Section 2.4.3 to extract similarity feature maps. Then we use RoIAlign to align the extracted features with each bounding box. For each prediction (including bounding box and mask) in frame P, we calculate its similarity score with each prediction in frame Q, using equation 2.3 with the aligned feature maps as input. Then we can obtain the prediction in frame Q that has the highest similarity score with it. The average of these two confidence scores is used as their final scores. Note that we only refine confidence scores that fall within the range of $[C_0, C_1]$.

2.5 Experiments

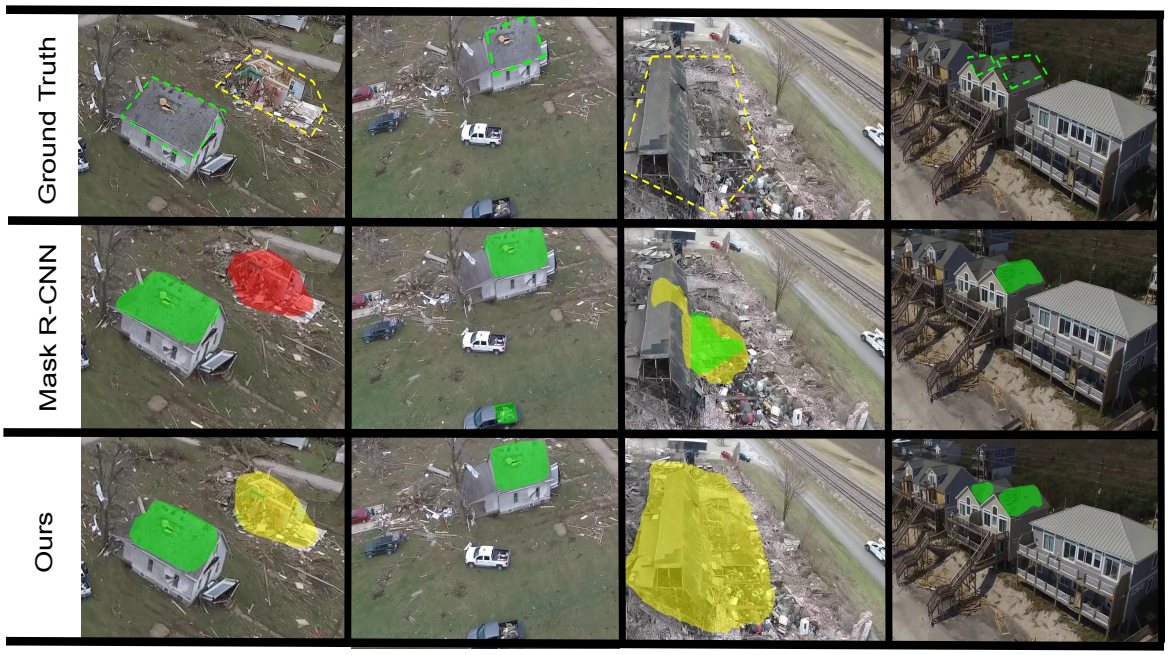

Figure 2.5: Visualization of the predicted damage segmentation. This figure demonstrates that our proposed model can alleviate the following errors: (1) label misclassification (first column, left to right); (2) false positive segmentation in the complex scenario with cars and buildings (second column); (3) incompleted masks in noisy video scenario (third column); and (4) missed masks (fourth column).

In this section, we compare our *MSNet* model with state-of-the-art baselines on the proposed ISBDA dataset. We randomly split the dataset into subsets with no overlapping scenes. We train our model using 80% of the dataset, and test on the rest 20% dataset. We repeat the split and experiments 3 times and report the results in Table 2.2. The final reported results are the average over the evaluation results of all splits.

We report the standard COCO instance segmentation metric [172] including AP (averaged over all IoU thresholds), AP@0.25, AP@0.5, and AP_S, AP_M, AP_L (AP at different scales). Unless noted, AP is evaluating using mask IoU.

Model	AP	AP_{25}	AP_{50}	AP^{bb}	AP^{bb}_{25}	AP^{bb}_{50}
Baseline	35.0	41.9	27.8	36.8	42.9	29.9
Baseline + HRPN	39.3	46.6	31.0	41.4	47.1	33.7
	(+4.3)	**(+4.7)**	**(+3.2)**	**(+4.6)**	**(+4.2)**	**(+3.8)**
Baseline + HRPN + SRN	40.0	47.7	31.3	42.1	48.1	33.9
	(+5.0)	**(+5.8)**	**(+3.5)**	**(+5.3)**	**(+5.2)**	**(+4.0)**

Table 2.3: Effect of HRPN and SRN. We use Mask R-CNN co-trained with building and damage instances as the baseline model. The results show that HRPN component gains significant improvement by 4.3% AP compared with the baseline model. Combined with HRPN, the SRN component also gets consistent improvement in both bounding box and mask branches.

2.5.1 Implementation Details

We compare our model with two recent state-of-the-art instance segmentation models, PolarMask [316] and Mask R-CNN [101]. All models use ResNet-50 based FPN as a backbone network. We train all the networks for 100 epochs, with a starting learning rate of 0.003 then we decrease it to 0.001 after 10 epochs. Mini-batch SGD is used as the optimizer with batch size equals 8. We initialize all the backbone networks with the weights pre-trained on COCO [172]. The input images are resized to have the shorter side being 800 and the longer side less or equal to 1333. For testing, an NMS with threshold 0.5 is used and top 100 detections are retained for each image.

For the score refinement procedure, SRN is trained using hard negative mining. We firstly generate 1,000 (X, X^+) pairs from different videos, and randomly extract 5 negative samples for each (X, X^+) pair as described in Section 2.4.3. We calculate the loss of 5 negative samples, and choose the top K ones with the highest losses as in [307] to optimize. For the experiments, we use $K = 1$. Adam optimizer [135] is used for network training with learning rate 0.001, and each batch is composed of one (X, X^+) pair and 5 negative samples. For testing, we choose C_0 = 0.2, and C_1 = 0.7 for the range described in Section 2.4.5.

2.5.2 Comparison to state-of-the-art

Baseline methods. We compare our method with state-of-the-art models and their variants customized for the damage instance segmentation problem. PolarMask [316] is a single shot

instance segmentation model with damage masks as input only. Mask R-CNN [101] is one of the state-of-the-art instance segmentation models. Two variants of Mask R-CNN are used as baselines: (1) Mask R-CNN with damage bounding boxes and masks as input; and (2) Mask R-CNN co-trained with damaged buildings and damages. Damaged building bounding boxes are used for RPN and R-CNN head training, and damage masks are used for the training of Mask head.

Quantitative results. Table 2.2 lists the damage instance segmentation results. Compared with PolarMask, our model is able to obtain significant improvement, *e.g.*, an absolute increment of 14.9% mask AP. For the Mask R-CNN baselines, we observe that Mask R-CNN trained with damage masks could be confused by the high variance of damage masks in different locations and scenarios. When the Mask R-CNN model is trained with building boxes and damage masks, the errors in building detection will impact the damage detection in the second stage. Also, the model could not precisely predict the damage masks from large building bounding boxes. Our proposed model utilizes the hierarchical nature of the damaged buildings and damaged parts, and outperforms the baseline with 5.0% AP in the segmentation branch and 4.7% AP in the bounding box branch.

Qualitative analysis. We qualitatively demonstrate the advantages of our model in Figure 3.4, showing that our proposed model can alleviate the following errors: (1) label misclassification (first column); (2) false positive segmentation in the complex scenario with cars and buildings (second column); (3) incompleted masks in noisy video scenario (third column); and (4) missed masks (fourth column). Thanks to the HRPN module and the inter-frame supervision, our model is able to generate accurate and robust detections even in very noisy scenarios like the third column of Figure 3.4.

2.5.3 Ablation Study

We evaluate our method on the ISBDA dataset. We use ResNet-50 FPN as a backbone network for ablation study. All experiments in this section are performed on one split.

Different IoU and II thresholds. In Figure 2.6, we compare the effects of different thresholds for IoU and II on the model performance using equations in Section 2.4.2. We train our model with IoU and II from 0.0 to 0.5 in steps of 0.1. For the model with IoU as metrics, the

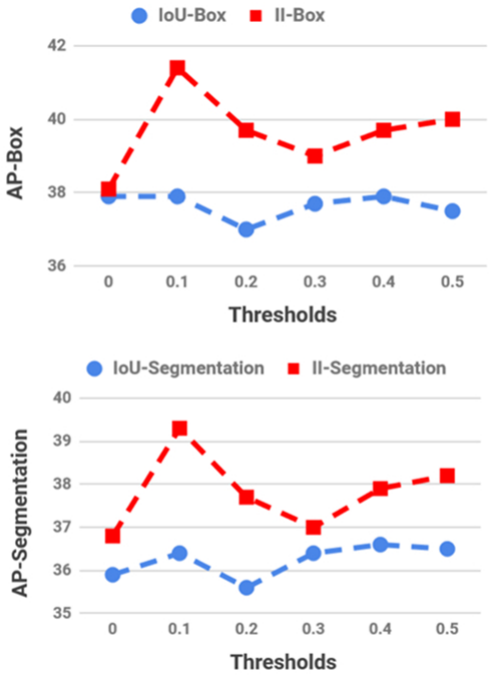

Figure 2.6: mAP of bounding box and segmentation using different IoU and II thresholds. The blue and red lines denote IoU and II metrics, respectively.

model gets the best performance when IoU equals 0.4. For the model with II as metrics, the model achieves the best performance when it equals 0.1.

Choices of IoU and II metrics. In Table 2.4, we report the best performance model among different IoU and II thresholds, respectively, where IoU equals 0.4 and II equals 0.1. We observe that II metric gains 2.7% AP improvement compared with IoU metric. By analyzing the AP in different sizes, we find that the small objects get the most significant improvement for 7.1% absolute value. This is probably because in IoU calculation, small damage anchors only occupy a small portion of its union with a large building bounding box. Therefore, small damage instances may not be well detected. On the other hand, II could properly handle such cases as it performs anchor sampling by calculating the intersection within the damage anchors.

Effect of HRPN and SRN. In Table 2.3, we experiment with the effect of HRPN and SRN. We observe that the HRPN component gains significant improvement by 4.3% AP compared with the baseline model. The SRN component further improves the model performance in both bounding box and mask branches.

M	AP	AP_{25}	AP_{50}	AP_S	AP_M	AP_L
IoU	36.6	42.5	30.1	47.4	41.1	38.6
II	39.3	46.6	31.0	54.5	38.0	42.0

Table 2.4: Results of different anchor sampling metrics.

2.6 Conclusion

In this paper, we investigate the problem of conducting damage assessment using user-generated aerial video data. We provide the first benchmark, namely ISBDA, for quantitative evaluation for models to assess building damage in aerial videos. Also, our proposed *MSNet* is able to explore the hierarchical spatial relationship among different objects and calibrate confidence scores to improve the model performance in both bounding box and mask branches. We empirically validate our model on the proposed ISBDA dataset, in which our model achieves the best results compared to state-of-the-art object detection models. We believe our dataset, together with our models, will facilitate future research in remote sensing and damage assessment for better and faster natural disaster relief.

Chapter 3

Spatial consistency learning for 3D semantic segmentation

In this chapter, we explore the benefit of spatial consistency for robust feature representation learning, and the model could generalize well towards variant noise conditions for semantic segmentation in 3D volumetric images.

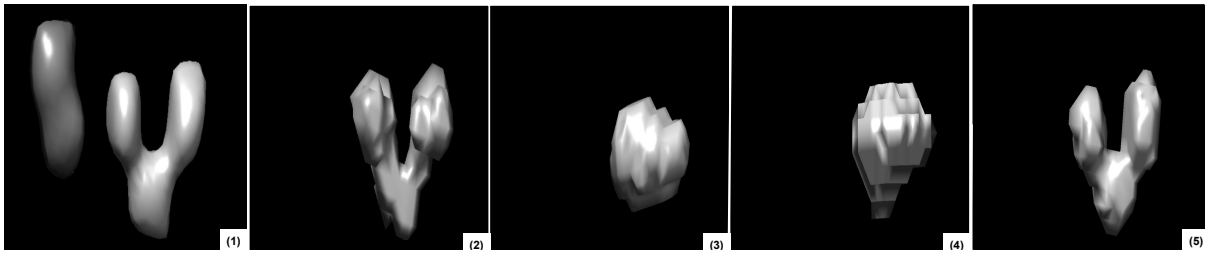

Figure 3.1: Illustration of 3D semantic segmentation using image-level class labels as supervision. This figure shows: (1) input 3D cryo-ET image; (2) ground truth segmentation; (3) semantic segmentation generated by Grad-CAM baseline. It only covers the most discriminative area; (4) Grad-CAM results augmented by our cross-image co-occurrence learning module. It is able to cover more integral areas; (5) segmentation generated by our *CIVA-Net*, which utilizes inter-voxel affinity relations to predict segmentation with accurate class boundaries. We do not visualize noise for better visualization purposes.

3.1 Overview

Recently, there has been an increasing interest in semantic segmentation for 3D images [33, 281]. 3D semantic segmentation methods that rely on point-wise annotations have been suc-

cessfully developed and achieved promising performance [33, 233, 281]. However, the full segmentation methods are generally data-hungry. To alleviate the time and labor-intensive data annotation process, weakly-supervised methods have been widely developed for two popular 3D data representations: point clouds [202, 220, 311, 319] and meshes [15, 267]. As the dominant 3D representation for biomedical images, voxel grids have not figured prominently in these developments, especially in the area that uses image-level class labels as supervision for full semantic segmentation. Existing weakly-supervised volumetric segmentation approaches still highly rely on the supervision of 2D slices [25], bounding boxes [320, 342] or sparse point annotations [235].

In this paper, we introduce a weakly supervised learning approach using image-level labels for 3D volumetric segmentation, with the focus on cryo-electron tomography (cryo-ET). In recent years, cryo-ET emerges as a revolutionary in situ 3D structural biology imaging technique for studying macromolecular complexes and virus structures in single cells [31]. Cryo-ET captures the 3D native structure and spatial distribution of all macromolecular complexes and other subcellular components without disrupting the cell [143]. During the COVID-19 pandemic, cryo-ET serves as a powerful imaging technique to study the structures of individual viruses and their interaction with host cells [132, 175]. Nevertheless, cryo-ET data is heavily affected by a low signal-to-noise ratio (SNR) due to the complex cytoplasm environment and missing wedge effects. Moreover, the cryo-ET based COVID-19 analysis is greatly impeded by the lack of ground truth data for model training. The ground truth masks of cryo-ET tomograms are generally obtained by template matching or human annotation. Template matching takes about 81 days to obtain the ground truth masks of one structure on one tomogram using one CPU core. If we use human annotation, annotating all structures on one tomogram takes about a month by a structural biology expert. To help the timely understanding of the virus infection, accurate semantic segmentation for 3D structures needs to be performed with fewer annotation efforts required.

Therefore, we propose a weakly-supervised 3D volumetric segmentation method based on image-level class labels. In our setting, image-level labels only indicate the classes that appeared in our input samples. Consider the example in Figure 7.2, there are three main challenges regarding semantic segmentation on cryo-ET images with image-level supervision. First, the cryo-ET images suffer from severe imaging limits such as noise and missing wedge effects (See Figure 3.3). Such limits greatly impede robust and accurate 3D semantic segmentation. Second, most of the advanced weakly supervised semantic segmentation (WSSS) methods on 2D images are based on class activation maps (CAM). However, the CAMs can only cover the most

discriminative area of the object and sometimes can incorrectly activate background regions, which can be summarized as under-activation and over-activation problems. The model thus cannot predict segmentation with accurate class boundaries. Third, the volumetric segmentation problem would be more challenging in 3D images due to the complex spatial structures, where semantic segmentation requires accurate boundary prediction.

To overcome the aforementioned challenges, we present a novel framework that utilizes both cross-image consensus and inter-voxel affinity relations. To address the under-activation and over-activation issues brought by CAM, we utilize the cross-image consensus among the same image group (i.e. images with the same class labels) to generate more consistent and integral object regions. This design provides high-quality supervision for the segmentation network. To detect accurate segmentation boundaries of complex 3D structures with only image-level labels available, we utilize the fine-grained inter-voxel affinity relations for the training of the segmentation network. Our framework can yield robust segmentation as it utilizes both cross-image and inter-voxel relations. To the best of our knowledge, we are the first to propose a 3D volumetric semantic segmentation model based on image-level supervision. To summarize, the contributions of this paper are three-fold:

- We propose a cross-image co-occurrence learning module to tackle the challenges brought by CAM and imaging limits.
- We propose an inter-voxel affinity learning module to predict segmentation with accurate boundaries of complex 3D structures with only image-level class labels available.
- Our experiments show that our method, namely *CIVA-Net*, achieves comparable performance to state-of-the-art models trained with stronger supervision.

3.2 Related Work

Weakly Supervised Semantic Segmentation on 2D Images. Recent studies [19, 133, 171, 269] presented promising results in 2D semantic segmentation with weak labels. Different kinds of supervision have been studied to reduce the labor cost for dense annotations, such as bounding box [133, 269], scribble [171], and point annotation [19]. Among those types of supervision, the image label is more popular as it requires the cheapest labor cost. The general framework for image-level tasks was firstly generating pixel-level seeds by using CAM-based methods [344] and then using these seeds as pseudo-supervision to train a full segmentation network. However, as CAM often failed to find the integral object region, several works [9,

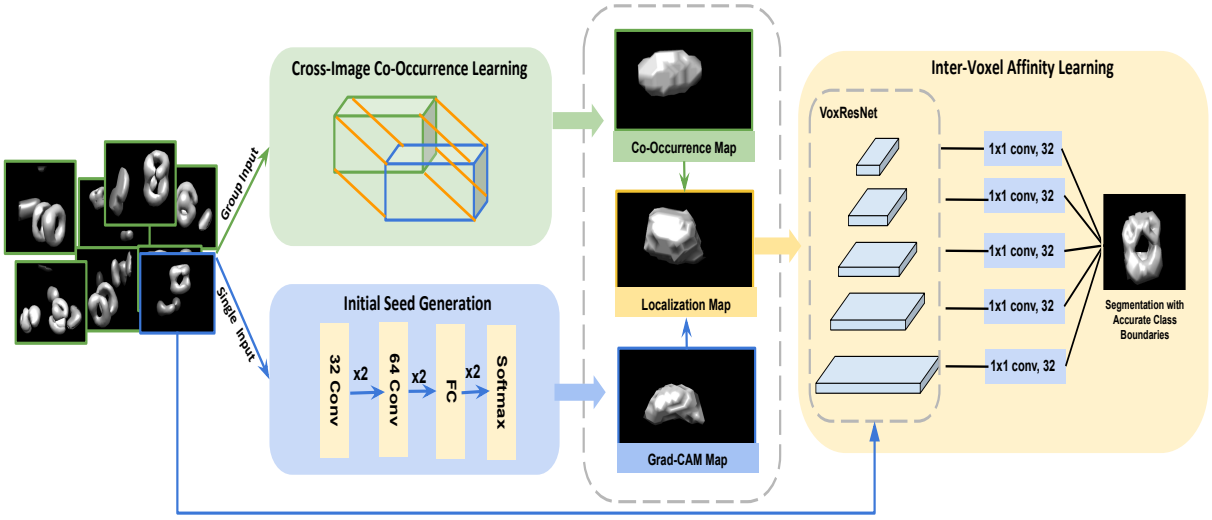

Figure 3.2: Network architecture of *CIVA-Net*. The left part includes the initial seeds generation module, which is combined with a cross-image co-occurrence learning module to generate more integral seed areas. For the initial seed generation, Grad-CAM is used to take single images as input to train a classification network. For the cross-image co-occurrence learning module, it takes group image as input to generate the co-occurrence map by utilizing group consensus embedding. Those two branches are combined at the end to produce the final localization map. The right part contains the inter-voxel affinity learning module. It utilizes the voxel affinity pairs sampled from the localization map to train a full segmentation network. During inference time, the inter-voxel affinity learning module will take raw 3D images as input to predict semantic segmentation results.

10, 137] were proposed to improve the accuracy of pseudo-labels. Compared to 2D weakly-supervised methods, 3D volumetric segmentation is more challenging as it involves imaging limits and more complex 3D spatial structures.

Object Co-Segmentation on 2D Images. Object co-segmentation aims to predict the segmentation of common objects for an image group [77, 78, 109]. Many 2D co-segmentation approaches were trained with strong pixel-level masks [34, 153]. Some weakly supervised methods used co-segmentation for initial seeds generation or incorporated the co-segmentation module to an end-to-end framework [77, 262]. However, 3D object co-segmentation has not been fully explored. We propose a novel cross-image co-occurrence learning module to generate consistent and integral object areas.

Semantic Segmentation on 3D Images. Current 3D semantic segmentation approaches can be put into three categories: supervised, semi-supervised, and unsupervised learning. Supervised learning approaches have gained popularity in recent years [51]. Cicek et al. proposed 3D U-Net [51] which extended previous U-Net architecture by replacing all 2D operations with their 3D counterparts. Chen et.al proposed VoxResNet [33] which was inspired by deep residual learning in 2D image recognition tasks. To reduce the need for large-size densely-labeled training data, some researchers proposed semi-supervised approaches for biomedical image segmentation [235]. For example, 2D slices were proposed as supervision to predict full object segmentation [51]. Point annotations were also adapted to reduce human annotation costs [235]. Other research proposed a network that was optimized by the weighted combination of a common supervised loss for labeled inputs and used a regularization loss for both labeled and unlabeled data [157]. Several unsupervised learning methods were based on learning anatomical prior [55] or training adversarial networks [127]. However, there is still a lack of volumetric segmentation methods based on image-level class labels, which can greatly reduce the annotation time and cost. Therefore, we propose a novel framework in order to predict accurate semantic segmentation with only image-level supervision.

3.3 Method

3.3.1 Overview

In this section, we describe our model for 3D semantic segmentation using image-level class labels as supervision, which we call *CIVA-Net*. The input of our model includes a single image and its class label c; and an image group that shares the same class label c. Our model contains two novel designs: (1) a cross-image co-occurrence learning module for integral region generation; (2) an inter-voxel affinity learning module that explores voxel affinity relations for precise semantic segmentation. In summary, it has the following four key components:

Initial Seed Generation takes a single image as input to train a classification network and generates pseudo voxel-level label.

Cross-Image Co-Occurrence Learning (CO) first obtains group consensus embedding from the image group. Then, it turns back to segment the common areas for the single image through co-occurrence learning. The co-occurrence map is combined with the initial seeds to produce the final localization map.

Inter-Voxel Affinity Learning (IVA) is proposed to explore the fine-grained inter-voxel re-

lations from the localization map for voxel affinity pairs generation.

Semantic Segmentation under Affinity Supervision is to predict the full image segmentation under the supervision of voxel affinity pairs.

See Figure 3.2 for a high-level summary of the model, and the sections below for more details.

3.3.2 Initial Seed Generation

Following previous weakly-supervised methods [77, 117, 311], we choose the CAM-based method to generate initial localization clues on 3D volumetric data. We use the Grad-CAM [254] with a 3D convolutional neural network as the model backbone. Grad-CAM plays three essential roles in our model. First, the localization map produced by Grad-CAM is used to define seed areas of objects. Second, the 3D CNN backbone of Grad-CAM is used as a feature encoder to produce group consensus, as described in Section 3.3.3. Third, Grad-CAM is used to produce image-level class labels during model inference.

Grad-CAM first trains a classification network using the image-level labels, and then obtains the pseudo segmentation label for certain classes. Specifically, given an image, in order to obtain the class-discriminative localization map $G^c \in \mathbb{R}^{T \times U \times V}$ of depth T, width U, and height V for class c, we first compute the gradient of the score for class c, y^c (layer before the softmax), with respect to feature map activations A^m of a convolutional layer, *i.e.* $\frac{y^c}{A^m}$. These gradients flowing back are global-average-pooled over the width, height and depth dimensions (indexed by i, j and k respectively) to obtain the neuron importance weights α_m^c:

$$\alpha_m^c = \frac{1}{Z} \sum_i \sum_j \sum_k \frac{\partial y^c}{\partial A_{ijk}^m}. \tag{3.1}$$

We perform a weighted combination of forward activation maps followed by a ReLU to obtain the localization map:

$$G_s^c = \mathrm{ReLU}\left(\sum_m \alpha_m^c A^m \right). \tag{3.2}$$

Then we perform spline interpolation [96] to resize the $T \times U \times V$ localization map to the original image size $D \times H \times W$, where D, H, and W denote the image depth, height, and width, respectively.

3.3.3 Cross-Image Co-Occurrence Learning

Unlike most of the existing weakly-supervised methods which learned from independent images [9, 10, 311], we propose a model to utilize cross-image relations to generate a more integral and consistent object area. The model aims to tackle the over-activation and under-activation challenges brought by Grad-CAM. The model first receives a group of images as input for the generation of a consensus representation [337] in a high-dimensional space with a learned feature encoder. The representation describes the common patterns of the image group that shares the same class label. Then, it turns back to segment the common areas for each sample by computing a co-occurrence map.

Specifically, given a group of images $\mathcal{I} = \{I_n\}_{n=1}^{N}$ with the same class label c, we first obtain its group consensus embedding. We employ the 3D convolutional network of Grad-CAM by removing the last fully connected layers as the 3D feature encoder $\mathcal{F}$. Our proposed method first extracts latent features $e_n = \mathcal{F}(I_n)$ of each single image I_n. The group consensus representation $\hat{e}$ of image group $\mathcal{I}$ can be calculated by:

$$\hat{e} = \text{Softmax}\left(\sum_{n=1}^{N} e_n\right). \tag{3.3}$$

$\hat{e}$ describes the common attributes of this image group. We aim to obtain the co-occurrence matrix between individual image feature $e_n \in \mathbb{R}^{C\times D\times H\times W}$ and the consensus embedding $\hat{e} \in \mathbb{R}^{C\times D\times H\times W}$, where C, D, H, W represent channel size, image depth, height and width. We first reshape e_n and $\hat{e}$ to $\mathbb{R}^{C\times N}$, and then perform a matrix multiplication between e_n transpose and $\hat{e}$. The result is an $N \times N$ matrix. Then we apply the max pooling operation to the second dimension of the matrix and get an $N \times 1$ matrix. Finally, we shape the $N \times 1$ matrix back to the input image shape, which is $D \times H \times W$. This matrix represents the co-occurrence relations between the individual image and group consensus embedding in voxel-level. The final co-occurrence map for class c is denoted as P^c.

To generate a consistent and integral segmentation for each individual image, we combine the co-occurrence map P^c and class-discriminative localization map G^c obtained in Section 3.3.2 by:

$$M^c_{ijk} = w_1 G^c_{ijk} + w_2 P^c_{ijk}, \tag{3.4}$$

where M^c_{ijk} is the voxel-level element in the merged localization map M^c. Note that we apply rank normalization [276] to G^c and P^c before the combination.

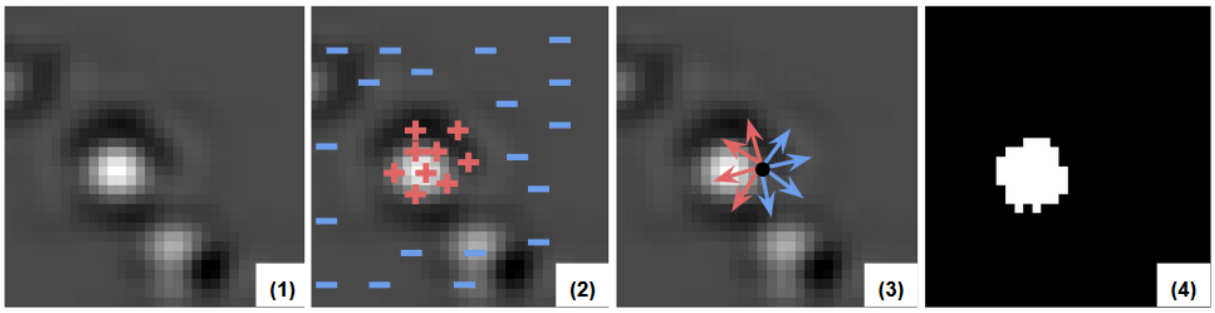

Figure 3.3: Semantic segmentation under the supervision of voxel affinity pairs. The figure shows: (1) the input biomedical image with heavy noise; (2) pseudo labels generated by localization map M^c; (3) voxel affinity pairs (S^-) sampled from M^c; (4) semantic segmentation generated by *CIVA-Net*. We only show one of the 2D slices for better visualization purposes.

3.3.4 Inter-Voxel Affinity Learning

Most of the existing weakly-supervised learning work directly trained a full segmentation network using the augmented voxel-wise pseudo labels [311, 313]. However, as the pseudo labels are not accurate, especially at the object boundaries, the model may not be able to learn from those inaccurate labels in a ordinary full segmentation manner. Therefore, we aim to utilize inter-voxel relations to force the model to predict object segmentation with precise class boundaries. We will first sample the voxel affinity pairs from the coarse localization map obtained in Section 3.3.3. Then, the model will train a segmentation network using the affinity pairs as supervision.

Inter-Voxel Affinity Mining. Because semantic segmentation requires precise object boundary prediction, inspired by [10], we propose a method to explore fine-grained inter-voxel relations of the localization map. Therefore, we carefully examine the merged localization map M^c to sample voxel affinity pairs. We first convert each voxel to a foreground or background class based on a threshold of $\hat{S}$. For foreground voxels, we further construct a class-map from M^c by choosing the class with the best score for each voxel. We obtain the pseudo class-map $\hat{M}$ where each voxel denotes the most probable class including a background class. Finally, we sample pairs of neighboring voxels from the pseudo class-map $\hat{M}$, and categorize them into two sets $\mathcal{S}^-$ and $\mathcal{S}^+_{bg}$ according to their class equivalence by:

$$\mathcal{S} = \{(p, q) \mid \|\mathbf{x}_p - \mathbf{x}_q\| < \gamma, \forall p \neq q\}, \tag{3.5}$$

$$\mathcal{S}^{-} = \{(p,q) \mid \hat{M}(\mathbf{x}_p) \neq \hat{M}(\mathbf{x}_q), (p,q) \in \mathcal{S}\}, \tag{3.6}$$

$$\mathcal{S}^{+}_{bg} = \{(p,q) \mid \hat{M}(\mathbf{x}_p) = \hat{M}(\mathbf{x}_q) = \mathbf{0}, (p,q) \in \mathcal{S}\}, \tag{3.7}$$

where (p,q) is the index of voxel affinity pair, and both x_p and x_q are of the form (i,j,k). γ is a radius limiting the maximum distance of a pair. $\mathbf{0}$ in Eqn 3.7 represents the background class. $\mathcal{S}$ represents the voxel pairs in which the distance of each pair is less than the radius γ. $\mathcal{S}^{-}$ represents a set of voxel pairs in which p and q have different class labels. $\mathcal{S}^{+}_{bg}$ represents a set of voxel pairs in which p and q have the same background class labels.

Semantic Segmentation with Voxel Affinity Supervision. We propose an inter-voxel affinity network (IVA) which predicts semantic segmentation with precise class boundaries. The input of the network is the 3D volumetric image and its voxel affinity pairs which are used as supervision for the network training. The network structure is shown in Figure 3.2. It uses VoxResNet [33] as the backbone network. Similar to the network structure used in [10], we first apply 1×1 convolution to each input feature map, and then the results are resized, concatenated, and fed into the last 1×1 convolution layer. The network output is object segmentation denoted by $\mathcal{O} \in [0,1]^{D\times H\times W}$. Because no ground-truth segmentation is available for training, we utilize the voxel affinity pairs to generate precise segmentation boundaries. The key assumption is that a class boundary exists somewhere between a pair of voxels with different class labels. Specifically, any path between negative pairs in Eqn. 3.6 must contain at least one foreground voxel (denotes as 1); any path between positive pairs in Eqn. 3.7 should only contain background voxels (denote as 0). The pair distance is limited by radius γ. As the 3D object could have visible holes, we do not sample foreground voxel pairs to supervise the model training. We propose the following 3D affinity matrix. For each pair of voxels $\mathbf{x}_p$ and $\mathbf{x}_q$, we define their semantic affinity a_{pq} as:

$$a_{pq} = 1 - \max_{k\in\Pi_{pq}} \mathcal{O}(\mathbf{x}_k), \tag{3.8}$$

where Π_{pq} is a set of voxels on the path between $\mathbf{x}_p$ and $\mathbf{x}_q$.

We utilize class equivalence relations between voxels as supervision for learning a_{ij}. The affinity is learned by minimizing cross-entropy between the one-hot vector of the binary affinity label and the predicted affinity in Eqn. 3.8:

$$\mathcal{L}_{\mathcal{O}} = -\sum_{(p,q)\in\mathcal{S}^{-}} \frac{\log(1-a_{pq})}{2|\mathcal{S}^{-}|} - \sum_{(p,q)\in\mathcal{S}^{+}_{\text{bg}}} \frac{\log a_{pq}}{2|\mathcal{S}^{+}_{\text{bg}}|}.$$

3.3.5 Training

During the training of Grad-CAM backbone, we use cross-entropy loss for class label prediction:

$$\mathcal{L}_{\mathcal{B}} = \sum_{i=1}^{N} \mathrm{CE}(cls^{i}, cls^{*i}), \tag{3.9}$$

where cls^{i} is the predicted label and cls^{*i} is the ground truth label. After obtaining the class-discriminative map generated by Grad-CAM, the 3D convolutional neural network is used as a feature encoder for image groups in co-occurrence learning. We get the merged localization map M^{c} by combining the Grad-CAM map and co-occurrence map. We then sample voxel affinity pairs by exploring affinity relations in M^{c}. These pairs are used as supervision for the training of the inter-voxel affinity network using loss $\mathcal{L}_{\mathcal{O}}$ described in Section 3.3.4. The final loss of our proposed approach is calculated using:

$$\mathcal{L} = \mathcal{L}_{\mathcal{B}} + \mathcal{L}_{\mathcal{O}}. \tag{3.10}$$

3.3.6 Inference

To predict the semantic segmentation for each image, we first use Grad-CAM to predict its class label c. Then we obtain the Grad-CAM map of class c and convert it to binary map $\bar{G}^{c}$. The 3D biomedical image is used as input for the inter-voxel affinity network to predict object segmentation. Because a single image could contain multiple target objects, we first retrieve the segmentation boundary proposals $\mathcal{O}_{b}^{1}, \mathcal{O}_{b}^{2}, ..., \mathcal{O}_{b}^{n}$ and choose the proposal that has the highest mIoU with $\bar{G}^{c}$ as the final segmentation. To further leverage the low-level contextual information, we implement 3D-CRF which replaces the original CRF [141] with 3D counterparts to refine the segmentation results.

3.4 Experiment

In this section, we compare our *CIVA-Net* with the state-of-the-art baselines on both simulated and real datasets of cryo-ET at different signal-to-noise ratios (SNR). We randomly split each dataset into training, test, and validation set, with ratios 70%, 15%, and 15%, respectively. We train our model on the training set, choose hyper-parameters of *CIVA-Net* based on the validation set, and report our results on the test set.

	SNR003					SNR005				
Method	mIoU	1bxn	1f1b	1yg6	covid	mIoU	1bxn	1f1b	1yg6	covid
Respond-CAM	15.2	27.0	12.4	2.31	19.1	9.9	6.6	11.9	1.9	19.0
Grad-CAM	14.8	20.3	15.3	1.44	22.3	9.7	11.1	0.2	24.6	24.0
CIVA-Net	20.6	29.2	12.4	6.89	34.1	24.4	16.9	11.7	38.6	30.5
CIVA-Net (3D-CRF)	39.9	48.2	28.7	52.6	30.0	38.8	46.4	24.3	55.8	28.7

Table 3.1: Comparison of *CIVA-Net* and the image-level baselines on two realistically simulated datasets.

3.4.1 Dataset

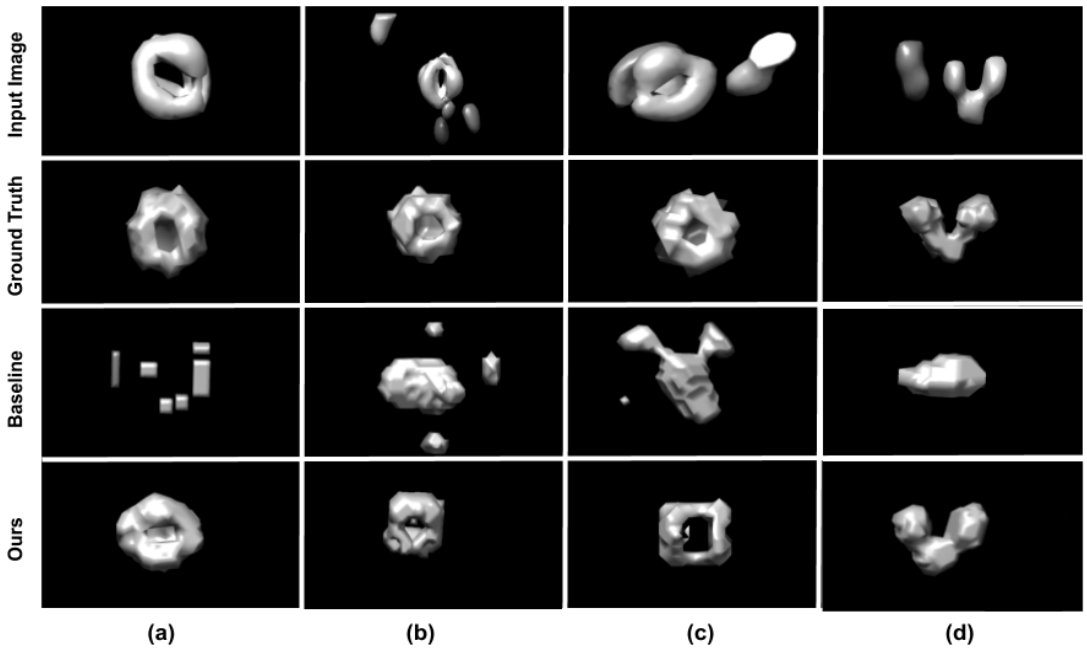

Figure 3.4: Visualization of the semantic segmentation results. We use Grad-CAM as the visualization baseline.

We follow common practice in cryo-ET analysis to evaluate our method on subtomograms [330, 331]. A subtomogram from a tomogram is a small cubic volume generally containing one macromolecule structure. To test the robustness and generalization of *CIVA-Net*, we process both simulated and real datasets to obtain submotograms containing one major structure and its neighbor structures.

Simulated Datasets. The subtomogram dataset simulation utilizes a standard procedure in work [71], which takes into account the tomographic reconstruction process with missing wedges and contrast transfer function. Besides the COVID-19 structural class, we also choose three representative macromolecule complexes in our simulated datasets (1bxn, 1f1b, and 1yg6). We simulate two datasets close to experimental conditions for all four classes, with SNR 0.03 and 0.05. Each dataset consists of 1,000 samples for each structure. Following prior work [170, 330], we resize each subtomogram to 32^3 due to GPU memory constrains. The simulated dataset contains 8,000 samples in total.

Real Dataset. To validate our model in experimental conditions, we use the publicly available Poly-GA dataset as our real dataset [94]. This dataset contains 756 subtomograms with unbalanced classes. It consists of 617 *Ribosome* subtomograms, 58 *26S* subtomograms, and 81 *TRiC* subtomograms. Such unbalanced class distribution is common in biomedical image processing. Each subtomogram is rescaled to size 32^3.

3.4.2 Evaluation Metrics

Following prior work [7, 309], we use the standard metrics of the mean intersection of union (mIoU) in these experiments. We also compute the class-specific mIoU to measure the model performance for each class.

3.4.3 Baseline Methods

Image-level Baselines. Following existing work [311], CAM-based methods are chosen as image-level baselines when there are no existing literature on 3D segmentation using image-level supervision. We choose Grad-CAM [254] and Respond-CAM [339] with the 3D CNN backbone as our baselines. We use the open-source implementation from [1].

State-of-the-art Baselines with Stronger Supervision. We also compare *CIVA-Net* with two of the state-of-the-art 3D segmentation models, 3D U-Net and VoxResNet using the open source code from [2] and [3]. For 3D segmentation trained with 2D slice supervision, 3D U-Net is one of the state-of-the-art models. We train 3D U-Net with the ground truth segmentation of one 2D slice, which covers 6.8% ground truth voxels. VoxResNet is trained with 3D full segmentation. Specifically, 2D slice supervision means the network learns from one 2D slice annotation and predicts a dense 3D segmentation. Full segmentation supervision is used when the full 3D masks are available, and the network densely segments new volumetric images.

Method	mIoU	ribo	26S	TRiC
Respond-CAM	12.8	14.5	6.1	2.7
Grad-CAM	19.0	22.8	0.4	0.0
CIVA-Net	36.1	37.1	25.4	36.3
CIVA-Net (3D-CRF)	67.8	74.2	32.3	39.9

Table 3.2: Comparison of *CIVA-Net* and the image-level baselines on the real dataset.

3.4.4 Implementation Details

Grad-CAM and Respond-CAM use the same network structure in [339] and share the same hyper-parameter settings and training configurations. The models are trained with a learning rate of 0.001. Adam [135] is used as the optimizer with batch size 32. The networks converge at about 20 epochs. 3D U-Net is trained with a learning rate of 0.0002. Adam [135] is used as the optimizer with batch size 1. We use the same hyper-parameter settings and training configurations for the experiments with 2D slices and full segmentation supervision. The network converges after about 50 epochs. For the training of VoxResNet, the learning rate is initially set to 0.001 and decreases at every iteration with exponential decay [255]. Adam [135] is used as the optimizer with batch size 16. The model converges after about 200 epochs.

For our *CIVA-Net*, it directly uses the Grad-CAM baseline as a part of its backbone network. The inter-voxel affinity network is trained from scratch and uses Stochastic Gradient Descent for network optimization with batch size 1. The learning rate is initially set to 0.0001 and decreases at every iteration with polynomial decay [184]. The model converges after about 3 epochs. The radius γ used in affinity pairs sampling is set to 2, and other hyper-parameters are determined by the validation set for each dataset. The model trained on 4,000 subtomograms takes 8 hours to converge with a single GTX 1080 Ti machine.

3.4.5 Quantitative Results

Comparison to Image-Level Baselines. Table 4.6 lists the evaluation results on two simulated datasets. As we can see, our model outperforms two image-level baselines in all classes and performs significantly better in the average mIoU metric. We report the mIoU evaluation results on the real dataset in Table 3.2. Our model also achieves superior performance on both average mIoU and class mIoU. For some classes with significantly fewer samples (26S and TRiC), our model can also generalize to these unbalanced classes and predict precise segmentation.

Method	$\text{mIoU}^{003}_{\text{snr}}$	$\text{mIoU}^{005}_{\text{snr}}$	$\text{mIoU}_{\text{real}}$
Supervision: Voxel-level			
3D U-Net$_s$	30.3	34.2	52.7
VoxResNet$_f$	77.0	78.5	89.8
Supervision: Image-level			
Ours	39.9	38.8	62.7

Table 3.3: Comparison of *CIVA-Net* and the state-of-the-art semantic segmentation models on both simulated and real datasets. 3D U-Net$_s$ is 3D U-Net trained with 2D slices. VoxResNet$_f$ is VoxResNet model trained with full segmentation supervision.

Grad-CAM	CO	IVA	3D-CRF	$\text{mIoU}^{003}_{\text{snr}}$	$\text{mIoU}^{005}_{\text{snr}}$	$\text{mIoU}_{\text{real}}$
✓				14.8	9.7	19.0
✓	✓			17.9	18.3	20.5
✓	✓	✓		20.6	24.4	36.1
✓	✓	✓	✓	39.9	38.8	62.7

Table 3.4: Performance of ablated versions of our model.

With the post-processing of our 3D-CRF module, the model can achieve better performance by leveraging low-level contextual information.

Comparison to State-of-the-Art Segmentation. In Table 3.3, we compare our final result with the existing state-of-the-art segmentation models that rely on stronger supervision. Similar to other state-of-the-art weakly supervised methods using image-level labels [311], there is still a performance gap between our proposed model and the state-of-the-art fully segmentation methods, but our weakly supervised approach achieves better performance to 3D U-Net models trained with stronger supervision.

3.4.6 Qualitative Analysis

We qualitatively demonstrate the advantages of our model in Figure 3.4. The first row is the input image. We can see it contains a major macromolecule and neighbor structures. The second row is the ground truth segmentation. The third and fourth rows are the semantic segmentation predicted by the baseline method (Grad-CAM) and our *CIVA-Net*. Compared with the baseline method, our *CIVA-Net* can alleviate the following errors: (a) wrong segmentation; (b) incomplete segmentation brought by heavy noise; (c) false-positive segmentation in complex

scenarios with neighbor structures; (d) segmentation with wrong class boundaries. Due to the cross-image co-occurrence and inter-voxel affinity learning designs, our model can generate accurate and robust segmentation in different scenarios.

3.4.7 Ablation Study

Ablation Study of *CIVA-Net*. We test various ablations of our model on both simulated and real datasets to substantiate our design decisions. The mIoU evaluation results are shown in Table 3.4. We observe that each component of our model gains consistent improvements on all datasets.

Ablation Study of Co-Occurrence Learning Module. We test the effects of different weights used in combining Grad-CAM map and co-occurrence map. The experiments are performed on the dataset with SNR 0.03. We report the mIoU evaluation results in Figure 3.5 with Grad-CAM map weight from 0 to 1 in steps of 0.1. Assuming the weight of Grad-CAM is w_1, then the weight of the co-occurrence map is $1 - w_1$. We observe that the model gets significant performance improvements from combining Grad-CAM map with the co-occurrence map. The model gets the best performance when the Grad-CAM weight equals 0.3 and the co-occurrence weight equals 0.7.

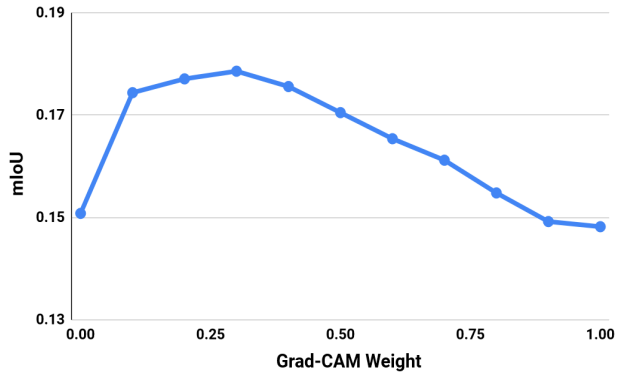

Figure 3.5: Ablation Study of Co-Occurrence Learning.

Ablation Study of Inter-Voxel Affinity Learning Module. To demonstrate the effectiveness of our inter-voxel affinity learning module, we compare our module with ordinary VoxResNet that directly takes the pseudo segmentation label as ground truth to train a full segmentation network using cross-entropy loss [33]. The results are reported in Table 3.5. The first row shows the mIoU of the pseudo segmentation labels. The second row shows the performance of

Setting	$\text{mIoU}^{003}_{\text{snr}}$	$\text{mIoU}^{005}_{\text{snr}}$	$\text{mIoU}_{\text{real}}$
Pseudo Label	17.9	18.3	20.5
Seg. w/o IVA	18.1	20.3	22.7
Seg. w/ IVA (Ours)	20.6	24.4	36.1

Table 3.5: Ablation Study of Inter-Voxel Affinity Learning.

VoxResNet trained with cross-entropy loss. The third row shows the model trained with voxel affinity pairs. We can see that the model can achieve better performance with our inter-voxel affinity learning module.

3.5 Conclusion

In this paper, we propose a novel weakly supervised approach for 3D semantic segmentation on cryo-ET images. Unlike most existing methods that require voxel-wise densely labeled training data, our weakly-supervised *CIVA-Net* is the first 3D model that only needs image-level class labels as guidance to learn accurate volumetric segmentation. Our model utilizes cross-image co-occurrence for integral and consistent region generation, and explores inter-voxel affinity relations to predict segmentation with accurate boundaries. Our experiments show that *CIVA-Net* can achieve comparable performance to the models trained with stronger supervision. Our model can be easily generalized to other 3D biomedical images. Moreover, our work fundamentally relates to COVID-19 research. We experiment on two simulated datasets containing the COVID-19 class and achieve superior performance. As a result, our model will assist the analysis of the 3D native structure of COVID-19 under the cryo-electron microscope, to benefit the design of effective therapeutics against COVID-19.

Part II

Masked Visual Modeling for Generalized Human Action Analysis

In this part, we focus on analyzing human behavior from temporal cues based masked visual modeling. We explore different modalities for human action analysis, including videos, 3D skeletons, point clouds, and meshes. We first propose to leverage adversarially masked consistency for scene-invariant action recognition (chapter 4). We then propose a masked vertex modeling technique for 3D mesh-based action recognition. Finally, we conduct generalized human action recognition by jointly modeling videos and 3D meshes (chapter 5).

Chapter 4

Adversarially masked consistency for video-based action recognition

In this chapter, we explore the benefit of adversarially learned masks for the unsupervised video domain adaptation task. The model is able to perform generalized human action analysis towards novel scenes across different geo-locations.

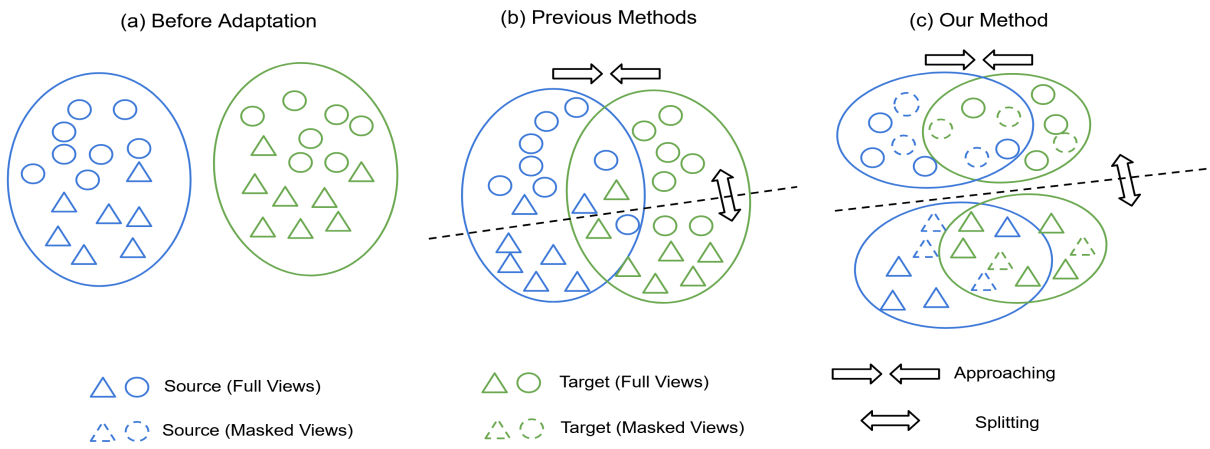

Figure 4.1: Visualization of the feature space for unsupervised domain adaptation methods. Existing state-of-the-art video domain adaptation models [38, 84, 290] used full-view input data to perform domain alignment as shown in (b). In this work, we propose a model that learns from adversarially masked samples, which can lead to the learning of effective domain-invariant and class-discriminative representations.

4.1 Overview

Egocentric vision [90, 126, 150, 164, 182, 186, 212, 214, 273, 275, 333] has attracted increasing attention in the computer vision community. It serves as key elements for various research fields, such as human-object interaction [186, 333], action recognition [212, 273], action anticipation [182, 214], social interaction analysis [150], and augmented reality [126, 164, 275]. Besides, egocentric vision has been popular in many real-world applications, among which egocentric action recognition is an important and challenging task compared to third-person action recognition, due to the presence of ego-motion caused by the action performer. Such camera motion introduces heavy noises that complicate the extraction of visual representation from the video frames [212]. Moreover, egocentric action recognition task usually requires high-fidelity modeling of human behaviors [333], such as *cut a vegetable* instead of coarse-grained actions such as *cook*. This requires the model to effectively recognize small objects and their mutual interaction. To train a discriminate model that is robust to sharp domain gaps, supervised approaches rely on collecting and annotating a large number of videos, which is expensive and may not be feasible in practice.

To address the lack of fine-grained data annotations, Unsupervised Domain Adaptation (UDA) setting is commonly used to transfer a model learned on a labeled source domain to an unlabelled target domain. However, existing unsupervised domain adaptation benchmarks for egocentric action recognition are limited to a single environment [56, 57] (*i.e.* kitchens), with small domain variances (*i.e.* different kitchens are treated as different domains). As a first step in this direction, we propose a new unsupervised domain adaptation benchmark, named U-Ego4D. We leverage the massive-scale Ego4D dataset [90]. It records daily-life activity videos spanning hundreds of scenarios (household, outdoor, workplace, leisure, etc.). The proposed U-Ego4D treats actions in different regions as different domains, which is more challenging. Moreover, the same action can happen in the same or different scenarios. For example, the action *cut*, can happen in a kitchen such as *cut dough* and *cut a vegetable*, or happen outdoors such as *cut grass*.

There are several challenges in training a model that is robust to various scenarios with only labeled source data available. First, there are multiple factors that could lead to domain gaps, including different backgrounds, lighting conditions, viewpoints, interacted objects, and motion variances. To bridge the domain gap, most of the existing state-of-the-art methods use adversarial learning to align two domains based on the full-view data. However, it might lead to trivial solutions (*i.e.* the model might be over-fitted to differentiate the source and target domain

only based on lighting differences, while other factors are neglected). Second, the decision boundary learned on labeled source videos may generalize poorly to the target domain. The model may overfit the source data well but is less discriminative for the target.

To tackle the aforementioned challenges, we propose a transformer-based model that utilizes masked data to avoid trivial solutions and learns more generalizable representations. This is different from existing state-of-the-art methods, which only take full-view data as inputs, as illustrated in Figure 4.1. Our model consists of two novel designs: Generative Adversarial Domain Alignment Network (GADAN) and a Masked Consistency Learning (MCL) module. GADAN simultaneously learns a masking generator and a domain-invariant encoder in an adversarial way. The domain-invariant encoder is trained to minimize the feature distance between the source and target domain. The mask generator, conversely, aims at producing challenging masks by maximizing the domain distance. To increase the model's class-discriminative ability, MCL enforces the prediction consistency between the masked target videos and their full forms, and enhances the understanding of spatial-temporal context. We show the efficacy of our model on the Epic-Kitchen and the proposed U-Ego4D benchmarks. Our contributions are three-fold:

- We propose the U-Ego4D benchmark, to enable the evaluation of video domain adaptation models in a more challenging and practical scenario.
- We introduce a new transformer-based model, which contains the Generative Adversarial Domain Alignment Network and the Masked Consistency Learning module to learn effective domain-invariant and class-discriminative representations.
- Our method outperforms existing state-of-the-art models on Epic-Kitchen and U-Ego4D benchmarks.

4.2 Related Work

Egocentric Vision. Egocentric vision is more complicated compared to third-person perception. It brings various challenges, such as sharp viewpoint movement, object occlusions, and environmental bias [134, 160, 211, 227, 228, 249, 271]. To help the model focus on the regions of interest and better recognize different actions, Sudhakaran [273] proposed to use both long-term and short-term attention mechanisms to recognize fine-grained actions. Lu [191] introduced a two-stream deep neural network which consists of an appearance-based stream and a motion-based stream for action recognition. Another stream focuses on leveraging multi-modal

information, such as RGB, depth, audio, and event camera [131, 156, 228]. [131] introduced a novel architecture for multi-modal temporal-binding. It is able to combine multiple modalities within a range of temporal offsets. The proposed framework combined three modalities (*i.e.* RGB, Flow and Audio) for egocentric action recognition. [156] proposed a transformer-based method which includes inter-frame attention encoder and mutual-attentional fusion block. The model consumes both RGB and depth images as inputs. [228] proposed to use event-camera data to distill optical flow information. Leveraging event-camera data is demonstrated to be effective and can improve performance of up to 4% with respect to RGB only information. However, those methods need extra sensors and increase the computational cost, as multiple backbone networks are needed to encode different modalities. To tackle the challenges brought by egocentric videos, we propose a masked consistency learning module to help the model learn the spatial-temporal context.

Video Domain Adaptation. Video domain adaptation has been studied to bridge domain gaps from different perspectives. One of the important tasks is cross-viewpoint domain adaptation [138, 151, 181, 236, 268, 350]. These works focused on learning geometric transformations of a camera but neglected other domain shifts such as environment differences. To learn viewpoint-invariant representations, 3D representations such as skeletons and human meshes are used as model inputs [181, 268]. The other stream for video domain adaptation focuses on environmental changes. Some of the recent works applied adversarial training for domain alignment [37, 120, 219]. Specifically, Gradiant Reverse Layer [151] was adapted to C3D [286], TRN [345] or both [219] architectures. Chen *et al.* [37] proposed an attention-based model to attend to the temporal dynamics of videos. Pan *et al.* [219] introduced a cross-domain attention model to learn relevant information. In contrast to previous literature which used complete videos as model inputs, we propose a model which leverages adversarially generated masks to better align the source and target domain.

Masked Visual Modeling. Masked visual modeling has gained attention in both Natural Language Processing and Computer Vision. It learns effective representations by masking and reconstruction. Some early works[298] treated the masks as a noise type [297] or missing regions and used inpainting objectives [221]. More recently, transformer backbones became more and more popular due to their flexibility to mask different patches [16, 66, 102, 310, 317, 346, 351]. BEiT [16] followed the success of BERT [63] and proposed methods to learn visual representations by predicting the discrete tokens [237]. He [102] proposed an encoder-decoder architec-

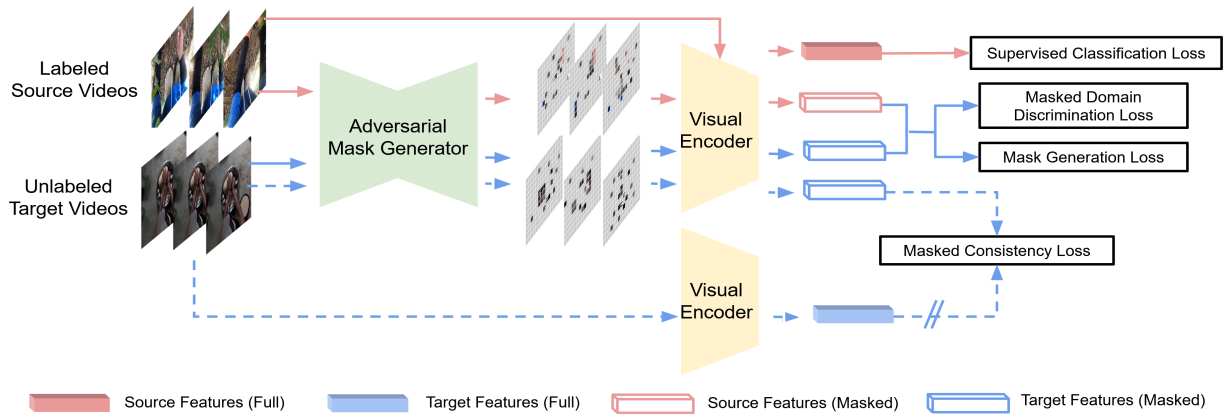

Figure 4.2: Overview of the proposed framework. There are two training stages to learn domain-invariant and class-discriminative representations. The goal of stage one (denoted by solid lines) is to align the source and target domains. As directly aligning the two domains using full views may lead to trivial solutions, we propose an adversarial mask generator to produce masked samples. This module is trained with the domain-invariant encoder in an adversarial way. For the training of stage two (denoted by dashed lines), we propose a Masked Consistent Learning module to enhance the model's understanding of the spatial-temporal context, and thus increase the class-discrimination ability. We first initialize the class-discriminative visual encoder using weights learned in stage one. Then we force the visual encoder to have consistent predictions on the full and masked views of the same target video. Our two-stage training framework learns effective domain-invariant and class-discriminative representations, with robustness to large domain gaps.

ture for masked image modeling. [265] proposed to adversarially learn the masks for masked visual modeling and performed evaluation on the image classification task. VideoMAE [283] was inspired by MAE [102] and adapted mask and reconstruction strategies to the video domain. One of the recent works applied the masked image modeling strategy for unsupervised domain adaptation on the image classification task [108]. In contrast to previous methods which use random masking strategies, we propose to generate challenging masks for the domain adaptation task in an adversarial way. Moreover, we do not use any reconstruction objectives and simply force the masked and unmasked views to have consistent predictions.

4.3 Method

4.3.1 Overview

In this section, we introduce our model for unsupervised video domain adaptation. Contrast to previous methods [38, 84, 212, 249, 290] which take full data views as inputs to train a domain-alignment or class-discriminative loss, we aim at developing a model that can benefit from the masked forms for better domain alignment and context understanding. Following previous works for unsupervised domain adaptation [211, 249, 290], we adopt a multi-stage training schema. Our model consists of two stages. For stage one, We train the Generative Adversarial Domain Alignment Network. Specifically, the adversarial mask generator and domain-invariant visual encoder are trained in an adversarial way. The adversarial mask generator aims at producing challenging samples to maximize the distance between the source and target domain. Conversely, the domain-invariant visual encoder is trained with those masked samples to minimize the domain distance. For stage two, we further fine-tune the domain-invariant encoder by forcing the label predictions between the masked and full unlabeled videos to be consistent. As there are no ground truth labels for the target samples, we generate the pseudo-labels using complete videos. In summary, our proposed method has the following key components:

- **Adversarial Mask Generator** is trained to generate challenging masks that will maximize the domain gap between masked source and target samples.
- **Domain-Invariant Visual Encoder** is trained to minimize the domain gap between the masked source and target samples. It is trained with Adversarial Mask Generator in an adversarial way.
- **Masked Consistency Loss** enforces the masked target videos and their full forms to have consistent label predictions.
- **Class-Discriminative Visual Encoder** is initialized with the weights of Domain-Invariant Visual Encoder trained in stage one, and is fine-tuned using the masked consistency loss.

See Figure 7.3 for a high-level summary of the model, and the sections below for more details.

4.3.2 Unsupervised Domain Adaptation

Given a set of labeled source videos $\mathcal{D}_S = \{(\mathbf{V}^{i\{s\}}, y^i)\}_{i=1}^{N_S}$ and unlabelled target videos $\mathcal{D}_T = \{\mathbf{V}^{i\{t\}}\}_{i=1}^{N_T}$, the goal of UDA task is to learn a model $\mathcal{H}$ which minimizes the task risk $\epsilon_{\mathcal{D}_T}(\mathcal{H})$

in the target domain, *i.e.* $\epsilon_{\mathcal{D}_T}(\mathcal{H}) = \mathbb{P}_{\mathcal{D}_T}[\mathcal{H}(x) \neq \mathcal{H}^*(x))]$, where $\mathcal{D}_T$ is the unlabeled target samples, and $\mathcal{H}^*$ is the ideal model in all model space. The model $\mathcal{H}$ consists of a feature extractor $\mathcal{F}$ and a classification head $\mathcal{G}$, *i.e.* $\mathcal{H}(x) = \mathcal{G}(\mathcal{F}(x))$.

4.3.3 Generative Adversarial Domain Alignment Network

To align the source and target domains, adversarial adaptive learning methods [84, 212, 290] are used to regularize the source and target representations, so as to minimize the distance between the empirical source and target mapping distributions: $\mathcal{H}(x_s)$ and $\mathcal{H}(x_t)$. Adversarial domain alignment is one of the commonly used strategies. A domain discriminator is trained along with the gradient reverse layer (GRL) [84] to minimize the domain distance. The gradient reversal mechanism ensures that the distributions over the two domains are forced to be similar (as indistinguishable as possible for the domain classifier), thus resulting in domain-invariant representations. Given an ideal model that is domain-invariant, the source classification model can be directly applied to classify the samples from the target domain. Given a binary domain label, d, indicating if an example $x \in \mathbf{S}$ or $x \in \mathbf{T}$, the domain discriminator is defined as,

$$\mathcal{L}_d = \sum_{x \in \{\mathbf{S},\mathbf{T}\}} -d \log(\mathcal{D}(\mathcal{F}(x))) - (1-d) \log(1 - \mathcal{D}(\mathcal{F}(x))) \tag{4.1}$$

Where $\mathcal{F}$ is the visual encoder and $\mathcal{D}$ is the domain classification head. However, there are multiple factors that could lead to domain gaps, including different backgrounds, lighting conditions, viewpoints, interacted objects, and motion variances. Existing state-of-the-art methods [38, 84, 249, 290] directly use Eqn. 4.1 to align two domains based on the full input data. However, it might lead to trivial solutions. To tackle this problem, we propose to minimize the distance between the empirical source and target mapping distributions learned from the **masked forms**: $h(m_s \odot x_s)$ and $h(m_t \odot x_t)$, where m_s and m_t are element-wise masks and $\odot$ is the Hadamard product. The model can learn from samples that are adversarially masked based on domain classification loss. The masked domain discrimination loss $\mathcal{L}_d^m$ is defined as:

$$\mathcal{L}_d^m = \sum_{x \in \{\mathbf{S},\mathbf{T}\}} -d \log(\mathcal{D}(\mathcal{F}(x \odot m))) - (1-d) \log(1 - \mathcal{D}(\mathcal{F}(x \odot m))) \tag{4.2}$$

There are multiple video masking options, such as pixel-wise masking, tube masking, and frame-wise masking. Another important aspect is the masking ratio. He *et. al* [102] demonstrated that large mask ratios are essential for effective self-supervised learning. In contrast to previous methods, we propose to learn the mask in an adversarial way. The learned masks

save the efforts to adjust masking hyper-parameters, and can produce challenging samples for domain-invariant learning. Given an RGB video $x \in \mathbb{R}^{t\times c\times w\times h}$, the adversarial mask generation model $\mathcal{M}$ produces an element-wise mask $m = \mathcal{M}(x)$, which is in the shape of $\mathbb{R}^{t\times c\times w\times h}$ with values in $[0, 1]$. The generation model is trained with the objective of maximizing the distribution shifts between the two domains. On the other hand, the domain-invariant visual encoder $\mathcal{F}$ takes the masked videos from the source and target domains as inputs, and tries to minimize the mapping distributions. The two models are jointly learned using the following function:

$$\mathcal{M}^\star, \mathcal{F}^\star = \arg\min_{\mathcal{F}} \max_{\mathcal{M}} \mathcal{D}^m(x_s, x_t; \mathcal{F}, \mathcal{M}) \,. \tag{4.3}$$

Where $\mathcal{D}^m$ denotes the masked feature distance. To stabilize the training process, inspired by Generative Adversarial Network (GAN) [88], we first freeze the mask generator and train the visual encoder using masked domain discrimination loss with GRL. In this way, the visual encoder learns domain-invariant representations that are as indistinguishable as possible for the domain classifier. Then we freeze the visual encoder and train the mask generator with masked domain discrimination loss only (without GRL). The mask generator learns to generate challenging masked videos which will maximize the domain distance (*i.e*, the masked views of the source and target videos are as distinguishable as possible for the domain classifier). The adversarial mask generator $\mathcal{M}$ consists of a U-Net architecture [243] and a pixel-wise softmax layer σ to ensure that the sum of the generated mask equals one.

4.3.4 Masked Consistency Learning

The goal of the Generative Adversarial Domain Alignment Network is to learn domain-invariant representations using masked views. However, the models learned on labeled source videos may overfit the source domain but are less discriminative for the target [129]. To help the model learn effective class-discriminative features, we enforce the model to make consistent predictions on the full and masked videos.

Specifically, we use an Adversarial Mask Generator trained in stage one to generate masked samples for unlabeled target videos. Moreover, we take the full video forms as inputs to generate pseudo-labels, and force the model to have consistent predictions on the masked and full videos. The proposed consistency learning module has two purposes: (1) Using the masked samples as a type of strong data augmentation. Based on [338], the unlabeled target samples can be divided into two groups: source-like samples and target-specific samples. The source-like samples can

already be easily classified after the domain alignment; the target-specific samples, however, are more likely to confuse the video classification model. We aim to apply the adversarial mask generation model in Sec 4.3.3 to generate more target-like samples. (2) Enforcing the model to have consistent predictions for better context learning. To recognize a human action, a model can utilize clues from different parts of the video. This can be local spatial information, which originates from the same region as the corresponding cell in the feature map, or context information, which comes from nearby patches in the spatial-temporal domain that can belong to different parts of the object or its background [107, 108]. The proposed Masked Consistency Learning (MCL) can help the model learn context relations on the unlabeled target domain, which will further improve the class-discriminative ability.

Specifically, MCL first generates adversarial masks using the mask generator trained in stage one, and then applies an element-wise multiplication of mask and video. In this way, the masked target prediction $\hat{y}^M$ can only rely on the limited information of the remaining video pixels:

$$\hat{y}^M = \mathcal{G}(\mathcal{F}(x \odot m)) \tag{4.4}$$

This makes the prediction more difficult. The masked consistency loss $\mathcal{L}_c^m$ can be represented as

$$\mathcal{L}_c^m = \sum_{x \in \{\mathbf{T}\}} -p^T \log \hat{y}^M \tag{4.5}$$

where p^T denotes a pseudo-label. The proposed model uses pseudo-labels as there is no ground truth available for the target domain. The pseudo-label is the prediction of the visual encoder and classification head of the complete target video x.

$$p^T = \operatorname{argmax} \mathcal{G}(\mathcal{F}(x)) \,. \tag{4.6}$$

4.3.5 Training

The model training includes two stages. For the first stage, we train the domain-invariant visual encoder using the masked domain discrimination loss $\mathcal{L}_d^m$ and the supervised classification loss $\mathcal{L}_\mathcal{S}$. Training with supervised classification loss expects the presence of labels and thus can only be applied to the labeled source input. The supervised classification loss can be represented as: $\mathcal{L}_\mathcal{S} = \sum_{i=1}^{N} \mathrm{CE}(cls^i, cls^{*i})$, where cls^i is the predicted label and cls^{*i} is the ground truth label, and N is the number of labeled source videos. To stabilize the training, the mask generator and visual encoder training proceed in alternating periods. In stage two, we freeze the mask generator and train the visual backbone using the masked consistency loss $\mathcal{L}_c^m$.

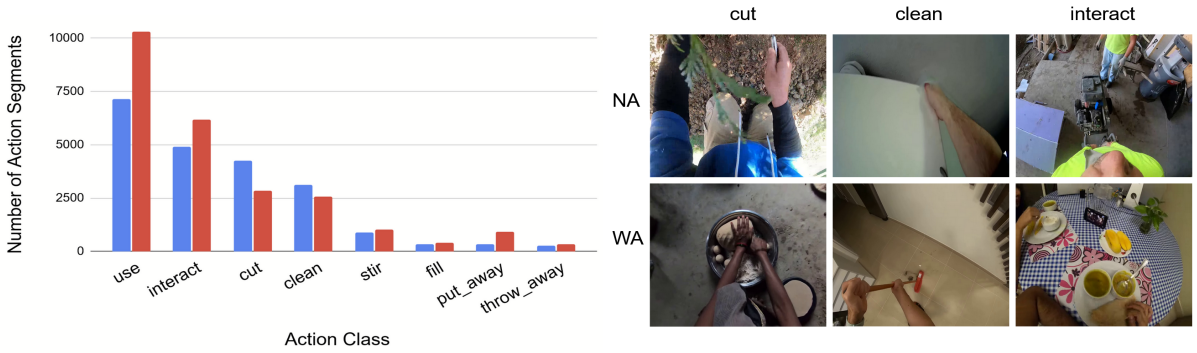

Figure 4.3: **Left:** Class distributions per domain for the U-Ego4D benchmark. **Right:** Videos collected from different regions are treated as different domains. Different from the Epic-Kitchen dataset which is limited to the kitchen scenario, the same action in the U-Ego4D benchmark can happen in totally different environments.

4.4 Experiment

4.4.1 Datasets

Epic-Kitchen. We evaluate our model on the commonly-used Epic-Kitchen dataset [56]. Epic-Kitchen contains egocentric videos for fine-grained activities in the kitchen environment. It has three domains (D1, D2, and D3), and each domain is a different kitchen. The task is to adapt between each pair of kitchens, which have different visual appearances. This benchmark contains 8 verb action classes, which occur in combination with different nouns. We use the standard training and test splits provided by the previous works [56, 249] to conduct our experiments.

U-Ego4D. We construct a new unsupervised domain adaptation benchmark called U-Ego4D. It builds on the massive-scale Ego4D dataset [90], which records daily-life activity videos spanning hundreds of scenarios (*e.g.* household, outdoor, workplace, leisure). We select two regions with the largest number of videos, North America and West Asia as two different domains (*i.e.* domain NA and domain WA). In this way, we can analyze the domain gaps between different regions. Moreover, in U-Ego4D, the same action can happen in the same or different scenarios (*e.g.* indoors and outdoors), which increases the action diversities. For example, in Figure 4.3, the action "interact" can happen indoors (during a meal) or happen outdoors. For the action categories, we select the 8 largest classes: (use, interact, clean, put_away, cut, throw_away, stir,

and fill). The class distributions are shown in Figure 4.3. The in-balanced class distributions bring additional challenges for the domain adaptation task. Besides, U-Ego4D is **3x** larger than Epic-Kitchen in terms of video length and clip number. Specifically, U-Ego4D has 35.67 hours of video with 35,937 video clips in total, while Epic-Kitchen only has 7.98 hours with 10,094 clips. Please refer to supplementary materials for data pre-processing details.

4.4.2 Implementation Details

We use the transformer-based architecture, ViT-Small[283], as our visual encoder. Following previous works [38, 84, 249, 290], the model is initialized with Kinetics-400 [130] pre-trained weights. Our model uses the same visual encoder and initialization strategies as the baselines with ViT backbones for a fair comparison. Following previous works for unsupervised domain adaptation [211, 249, 290], we adopt a multi-stage training schema. The adversarial mask generator and domain-invariant encoder are trained in stage one. The class-discriminative visual encoder is trained using the masked consistency loss in stage two. We use a clip length of 16 with a spatial size of 224 $\times$ 224 to train all the models. AdamW [190] is used with β equals (0.9, 0.999) as the model optimizer with a learning rate 1e-4. We use batch size 8 for all the experiments. For the implementation of the mask generator, we use a U-Net structure with a depth of 4. For inference, we use 16 randomly sampled frames per video and use the visual encoder along with the classification head to recognize the action. For more details, please see the supplementary material.

4.4.3 Baselines

We compare our model with state-of-the-art unsupervised domain adaptation models. ADDA [290] is a general framework which combined discriminative modeling, untied weight sharing, and a GAN loss for unsupervised domain adaptation. DANN [84] proposed gradient reverse layer (GRL) for domain-invariant representation learning. TA^{3}N [38] adapted attention mechanisms to explicitly attend to the temporal dynamics using domain discrepancy for effective domain alignment. CoMix is a contrastive learning framework which leveraged background mixing to produce augmented samples [249]. TransVAE [312] combines seven different loss functions [38, 84, 345] for spatial-temporal disentanglement and domain gap minimization. All of those models were originally implemented with the I3D backbone. We also run experiments for DANN [84] and CoMix [249] by replacing I3D with the same ViT backbone as our proposed model for a fair comparison. Note that it is not feasible to replace the backbone of TransVAE

Method	Backbone	Epic-Kitchens						Average
		D2→D1	D3→D1	D1→D2	D3→D2	D1→D3	D2→D3	
Supervised Source	I3D	35.4	34.6	32.8	35.8	34.1	39.1	35.3
ADDA [290]	I3D	36.3	36.1	35.4	41.4	34.9	40.8	37.4
DANN [84]	I3D	38.3	38.8	37.7	42.1	36.6	41.9	39.2
TA^{3}N [38]	I3D	40.9	39.9	34.2	44.2	37.4	42.8	39.9
CoMix [249]	I3D	38.6	42.3	42.9	49.2	40.9	45.2	43.2
TranSVAE [312]	I3D	50.3	48.0	50.5	58.0	50.3	**58.6**	52.6
Supervised Target	I3D	57.0	57.0	64.0	64.0	63.7	63.7	61.5
Supervised Source	ViT	44.7	45.6	53.3	55.6	46.7	47.8	49.0
DANN [84]	ViT	49.8	47.5	58.9	57.3	53.8	52.4	53.3
CoMix [249]	ViT	46.3	47.3	56.7	59.3	51.4	52.3	52.2
Ours	ViT	**50.7**	**48.2**	**64.6**	**60.8**	**55.5**	56.6	**56.1**
Supervised Target	ViT	57.6	57.6	66.5	66.5	67.2	67.2	63.8

Table 4.1: Experimental Results on Epic-Kitchens Dataset. Our model achieves the best average performance among all state-of-the-art methods.

with ViT, as some of the loss functions in TransVAE are specifically designed for its architecture. All of these models use single modality features as our proposed method. There are several recent works conducting video-based unsupervised domain adaptation using multi-modal data which combines RGB and Flow. Although our method solely uses RGB information, we still take this set of methods into account following the previous work [312]. Specifically, we consider MM-SADA [211], STCDA [271], CMCD [249], CleanAdapt [59], MixDANN [327] and CIA [323].

4.4.4 Main Results

Results on Epic-Kitchen. We compare our model with state-of-the-art unsupervised video domain adaptation models on the Epic-Kitchen dataset and report the results in Table 4.1. Our model achieves the best performance on the 5 of 6 splits as well as the best average performance of 56.1%. Besides, we observe that the performance gap between our method and the supervised target model has been reduced to 7.7%, which demonstrates the potential to close the domain gap using masked video modeling methods.

Compare to Multi-Modal Methods. We further compare our model with recent video-based unsupervised domain adaptation methods that use multi-modalities, *i.e.*, RGB features

Method	Epic-Kitchens						Average
	D2→D1	D3→D1	D1→D2	D3→D2	D1→D3	D2→D3	
Supervised Source	43.0	43.0	43.2	55.5	42.5	48.0	45.9
MM-SADA [211]	48.2	50.9	49.5	56.1	44.1	52.7	50.3
STCDA [271]	49.0	**52.6**	52.0	55.6	45.5	52.5	51.2
CMCD [249]	49.5	48.7	50.3	56.3	46.3	52.0	51.0
CleanAdapt [59]	46.2	47.8	52.7	54.4	47.0	52.7	50.3
MixDANN [327]	50.3	51.0	56.0	54.7	47.3	52.4	52.0
CIA [323]	49.8	52.2	52.5	57.6	47.8	53.2	52.2
Ours	**50.7**	48.2	**64.6**	**60.8**	**55.5**	**56.6**	**56.1**

Table 4.2: Experimental Results on Epic-Kitchen with comparisons to approaches using multi-modality data as the input. Our model, which only uses RGB videos, achieves the best average performance among all state-of-the-art methods.

and optical flows, although our model only uses RGB features. The results are reported in Table 4.2. We observe that our method achieves the best average result among all of the six multi-modal methods.

Results on U-Ego4D. We compare our model with state-of-the-art unsupervised video domain adaptation models on the proposed U-Ego4D benchmark and report the results in Table 4.3. Our model achieves the best performance including the best average performance of 53.9%. Although our model achieves promising performance, there is still a large performance gap between our method and the supervised target model, which is 15.5%. This shows the great potential for further improvement to bridge the large domain gap caused by different regions. Besides, we observe that on Epic-Kitchen, all the methods can help increase the performance from Supervised-Source by at least an absolute 3.2% (CoMix) to 7.1% (ours), while on U-Ego4D we are only seeing an increase of 2.5% (CoMix) to 3.6% (ours). Low-bound (Supervised-Source) methods on both datasets have similar performance (49.0% v.s. 50.3%), while U-Ego4D has a higher upper bound (supervised-target) 69.4% compared to Epic-Kitchen which is 63.8%. This suggests that there is still a larger domain gap after applying the state-of-the-art domain adaptation methods and thus the proposed U-Ego4D dataset is more challenging.

Method	Backbone	U-Ego4D		Average
		NA→WA	WA→NA	
Supervised Source	ViT	52.5	48.1	50.3
DANN [84]	ViT	56.9	48.6	52.8
CoMix [249]	ViT	56.3	48.9	52.6
Ours	ViT	**58.4**	**49.4**	**53.9**
Supervised Target	ViT	67.5	71.3	69.4

Table 4.3: Experimental Results on the proposed U-Ego4D Dataset. Different from Epic-Kitchen which specifically focuses on domain transfer among different kitchens, U-Ego4D focuses on a more practical setting: domain adaptation between different regions. Our model achieves the best performance compared to state-of-the-art models.

Method	Backbone	Epic-Kitchen						Average
		D2→D1	D3→D1	D1→D2	D3→D2	D1→D3	D2→D3	
S.O.	ViT	44.7	45.6	53.3	55.59	46.7	47.8	49.0
S.O.+GADAN	ViT	50.2	47.9	61.0	56.8	54.3	53.3	53.9
S.O.+GADAN+MCL	ViT	**50.7**	**48.2**	**64.6**	**60.8**	**55.5**	**56.6**	**56.1**

Table 4.4: Ablation Studies on Epic-Kitchen. Each of our proposed modules brings stable performance improvement in all tasks. S.O. stands for the source-only model, GADAN stands for Generative Adversarial Domain Alignment Network, and MCL stands for Masked Consistency Learning.

4.4.5 Ablation Studies and Analysis

Effectiveness of Different Components. We test various ablations of our model on the Epic-Kitchen dataset to substantiate our design decisions. The results are shown in Table 4.4. We observe that each component of our model brings consistent improvements in all six splits. Overall, compared to the source-only baseline, Generative Adversarial Domain Alignment Network improves the average accuracy from 49.0% to 53.9%, and Masked Consistency Learning can further improve the accuracy to 56.1%.

Comparison of Different Loss Functions and Masking Strategies for GADAN. We test our Generative Adversarial Domain Alignment Network with different loss functions and masking strategies. The results are shown in Table 4.5. We perform all the experiments with the same

Method	Top-1 (%)
S.O.	53.3
S.O.+DL	58.9
S.O.+MDL + random tube (r=0.5)	59.8
S.O.+MDL + random tube (r=0.75)	59.6
S.O.+MDL + random tube (r=0.9)	59.3
S.O. + AMG (GADAN)	**61.0**

Table 4.5: Comparison of Different Loss Functions and Masking Strategies. DL stands for the domain discrimination loss (with full-view inputs). MDL stands for the masked domain discrimination loss. AMG stands for our adversarial mask generator.

Method	Top-1 (%)
GADAN	61.0
GADAN + naive pseudo labeling	62.2
GADAN + $\mathrm{MCL}_{\mathrm{CE}}$ + random tube	63.4
GADAN + $\mathrm{MCL}_{\mathrm{MSE}}$ + AMG	62.8
GADAN + $\mathrm{MCL}_{\mathrm{CE}}$ + AMG	**64.6**

Table 4.6: Comparison of Different Loss Functions and Masking Strategies. DL stands for the domain discrimination loss (with full-view inputs). MDL stands for the masked domain discrimination loss. AMG stands for our adversarial mask generator.

hyper-parameters on the D1→D2 split of Epic-Kitchen for a fair comparison. Row 1 shows the performance of the source-only baseline. Row 2 shows the performance of the source-only baseline with the domain discrimination loss (DL) and gradient reverse layer. For rows 3-5, we replace the masks produced by the Adversarial Mask Generator in GADAN with random tube masks. We test three mask ratios: 0.5, 0.75 and 0.9. The last row is the result of our full GADAN model. We observe that the proposed adversarial mask generator can bring 1.2% performance improvement compared to the best model that is trained with random masks. Moreover, as our masks are directly learned using the mask generation objective, it saves the efforts to adjust the hyper-parameters, such as mask types and mask ratios.

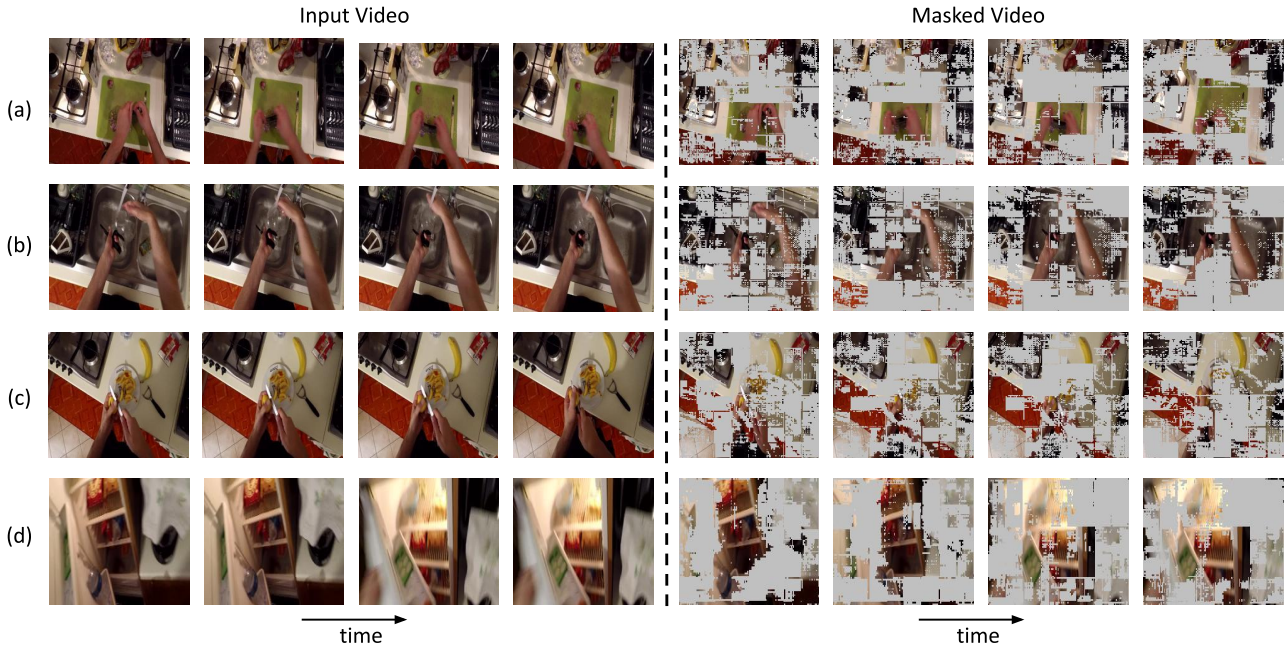

Figure 4.4: Visualizations of the Adversarially-Learned Masks.

Comparison of Different Loss Functions and Masking Strategies for MCL. We test MCL with different consistency losses and masking strategies. We perform all the experiments with the same hyper-parameters on Epic-Kitchen's D1→D2 split. The results are shown in Table 4.6. Row 1 shows the performance of GADAN. Row 2 shows the performance of GADAN with the naive pseudo-labeling method. Row 3 shows the performance of GADAN with MCL. Here we replace masks produced by AMG with random tubes. Comparing row 3 and row 5 (our full model), we find that adversarially generated masks lead to better performance for masked consistency learning. We also test different consistency losses including the cross-entropy loss (MCL_{CE}) and mean squared error loss (MCL_{MSE}). For the cross-entropy loss, we force the hard labels predicted from masked and full views to be consistent. For the mean squared error loss, we force the soft logits of masked and full views to be consistent. Comparing row 4 and row 5, we find that cross-entropy loss leads to better performance compared to mean squared error loss.

Visualizations of the Adversarially-Learned Masks. We visualize the adversarially-learned masks in Figure 4.4. We observe that after applying the learned masks, only the key instances are kept for each frame. Specifically, in (a) and (b), the regions that describe human-object interaction (person's hands, green board, sink) are retained. In (c) and (d), some of the key objects are preserved, such as bananas and refrigerators. In this way, the model is able to use the gen-

erated hard samples for domain-invariant and class-discriminative feature learning, and thus further improve performance.

4.4.6 Discussion

Although our model is trained and evaluated on eight different splits, potential dataset biases can still cause negative societal impact in a real-world deployment. For example, due to the small size of the action datasets, they may not properly represent actions performed by minority groups. Therefore, the models trained on these datasets (whether domain-adapted or not) might still under-represent some groups of people in the real world applications.

4.5 Conclusion

We have presented a novel transformer-based model for unsupervised domain adaptation in egocentric videos. We are the first to show that masked video modeling can benefit both domain-invariant and class-discriminative feature learning. Our method also establishes new state-of-the-art performance on Epic-Kitchen and U-Ego4D. We believe our dataset, together with our models, will facilitate future research in the domain adaptation and generalization field.

Chapter 5

Masked vertex modeling for 3D mesh-based action recognition

In this chapter, we propose the first work that leverages 3D mesh representations for generalized human action recognition. We propose two pre-training objectives, namely masked vertex modeling and future frame prediction, to help the model learn better spatial-temporal context. The model is able to perform generalized human action analysis towards novel viewpoints and scenes, thanks to the robust 3D mesh representations.

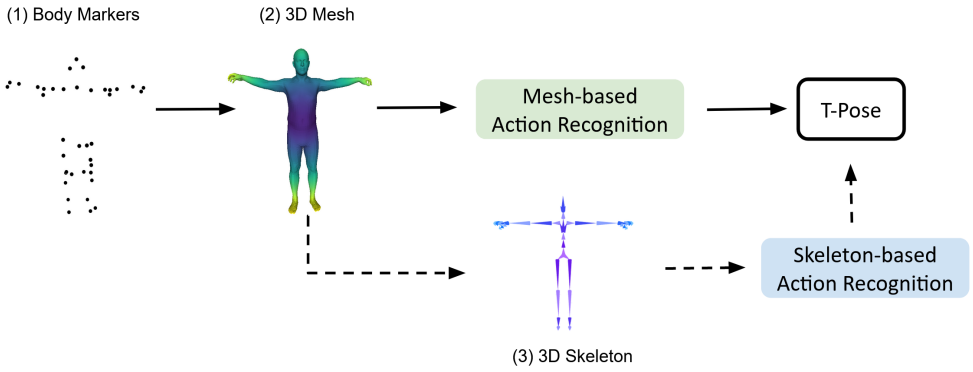

Figure 5.1: Current state-of-the-art MoCap-based action recognition methods first convert body markers into a human body mesh, which is used to predict a standardized 3D skeleton. The 3D skeleton is used as input for action recognition models (dashed line). We propose a method that directly models the dynamics of raw mesh sequences (solid line). Our method saves the manual effort to derive skeleton representation, and achieves superior recognition performance by leveraging surface motion and body shape knowledge from meshes.

5.1 Overview

Motion Capture (MoCap) is the process of digitally recording the human movement, which enables the fine-grained capture and analysis of human motions in 3D space [194, 231]. MoCap-based human perception serves as key elements for various research fields, such as action recognition [73, 210, 213, 218, 231, 256], tracking [213], pose estimation [5, 136], imitation learning [341], and motion synthesis [213]. Besides, MoCap is one of the fundamental technologies to enhance human-robot interactions in various practical scenarios including hospitals and manufacturing environment [100, 142, 195, 201, 205, 347]. For example, Hayes [100] classified automotive assembly activities using MoCap data of humans and objects. Understanding human behaviors from MoCap data is fundamentally important for robotics perception, planning, and control.

Skeleton representations are commonly used to model MoCap sequences. Some early works [18, 154] directly used body markers and their connectivity relations to form a skeleton graph. However, the marker positions depend on each subject (person), which brings sample variances within each dataset. Moreover, different MoCap datasets usually have different numbers of body markers. For example, ACCAD [218], BioMotion[287], Eyes Japan [73], and KIT [197] have 82, 41, 37, and 50 body markers respectively. This prevents the model to be trained and tested on a unified framework. To use standard skeleton representations such as NTU RGB+D [259], Punnakkal *et al.* [231] first used Mosh++ to fit body markers into SMPL-H meshes, and then predicted a 25-joint skeleton [179] from the mesh vertices [242]. Finally, a skeleton-based model [263] was used to perform action recognition. Although those methods achieved advanced performance, they have the following disadvantages. First, they require several manual steps to map the vertices from mesh to skeleton. Second, skeleton representations lose the information provided by original MoCap data (*i.e.*, surface motion and body shape knowledge). To overcome those disadvantages, we propose a mesh-based action recognition method to directly model dynamic changes in raw mesh sequences, as illustrated in Figure 7.1.

Though mesh representations provide fine-grained body information, it is challenging to classify high-dimensional mesh sequences into different actions. First, unlike structured 3D skeletons which have joint correspondence across frames, there is no vertex-level correspondence in meshes (*i.e.*, the vertices are unordered). Therefore, the local connectivity of every single mesh can not be directly aggregated in the temporal dimension. Second, mesh representations encode local connectivity information, while action recognition requires global understanding in the whole spatial-temporal domain.

To overcome the aforementioned challenges, we propose a novel Spatial-Temporal Mesh Transformer (*STMT*). *STMT* leverages mesh connectivity information to build patches at the frame level, and uses a hierarchical transformer which can freely attend to any intra- and inter-frame patches to learn spatial-temporal associations. The hierarchical attention mechanism allows the model to learn patch correlation across the entire sequence, and alleviate the requirement of explicit vertex correspondence. We further define two self-supervised learning tasks, namely masked vertex modeling and future frame prediction, to enhance the global interactions among vertex patches. To reconstruct masked vertices of different body parts, the model needs to learn prior knowledge about the human body in the spatial dimension. To predict future frames, the model needs to understand meaningful surface movement in the temporal dimension. To this end, our hierarchical transformer pre-trained with those two objectives can further learn spatial-temporal context across entire frames, which is beneficial for the downstream action recognition task.

We evaluate our model on common MoCap benchmark datasets. Our method achieves state-of-the-art performance compared to skeleton-based and point-cloud-based models. The contributions of this paper are three-fold:

- We introduce a new hierarchical transformer architecture, which jointly encodes intrinsic and extrinsic representations, along with intra- and inter-frame attention, for spatial-temporal mesh modeling.
- We design effective and efficient pretext tasks, namely masked vertex modeling and future frame prediction, to enable the model to learn from the spatial-temporal global context.
- Our model achieves superior performance compared to state-of-the-art point-cloud and skeleton models on common MoCap benchmarks.

5.2 Related Work

Action Recognition from Depth and Point Cloud. 3D action recognition models have achieved promising performance with depth [180, 250, 251, 304, 315] and point clouds [75, 185, 233, 308]. Depth provides reliable 3D structural and geometric information which characterizes informative human actions. In MVDI [315], dynamic images [20] were extracted through multi-view projections from depth videos for 3D action recognition. 3D-FCNN [250] directly exploited a 3D-CNN to model depth videos. Another popular category of 3D human action recognition is based on 3D point clouds. PointNet [232] and PointNet++ [233] are the pioneer-

ing works contributing towards permutation invariance of 3D point sets for representing 3D geometric structures. Along this avenue, MeteorNet [185] stacked multi-frame point clouds and aggregates local features for action recognition. 3DV [308] transferred point cloud sequences into regular voxel sets to characterize 3D motion compactly via temporal rank pooling. PST-Net [75] disentangled space and time to alleviate point-wise spatial variance across time. Action recognition has shown promising results with 3D skeletons and point clouds. Meshes, which are commonly used in representing human bodies and creating action sequences, have not been explored for the action recognition task. In this work, we propose the first mesh-based action recognition model.

MoCap-Based Action Recognition. Motion-capture (MoCap) datasets [73, 203, 210, 213, 218, 231, 256] serve as key elements for various research fields, such as action recognition [73, 203, 210, 213, 218, 231, 256], tracking [213], pose estimation [5, 136], imitation learning [341], and motion synthesis [213]. MoCap-based action recognition was formulated as a skeleton-based action recognition problem [231]. Various architectures have been investigated to incorporate skeleton sequences. In [68, 178, 336], skeleton sequences were treated as time-series inputs to RNNs. [106, 303] respectively transformed skeleton sequences into spectral images and trajectory maps and then adopted CNNs for feature learning. In [322], Yan *et al.* leveraged GCN to model joint dependencies that can be naturally represented with a graph. In this paper, we propose a novel method to directly model the dynamics of raw mesh sequences which can benefit from surface motion and body shape knowledge.

Masked Autoencoder. Masked autoencoder has gained attention in Natural Language Processing and Computer Vision to learn effective representations using auto-encoding. Stacked denoising autoencoders [298] treated masks as a noise type and used denoising autoencoders to denoise corrupted inputs. ViT [67] proposed a self-supervised pre-training task to reconstruct masked tokens. More recently, BEiT [16] proposed to learn visual representations by reconstructing the discrete tokens [237]. MAE [102] proposed a simple yet effective asymmetric framework for masked image modeling. In 3D point cloud analysis, Wang *et al.* [299] chose to first generate partial point clouds by calculating occlusion from random camera viewpoints, and then completed occluded point clouds using autoencoding. Point-BERT [328] followed the success of BERT [63] to predict the masked tokens learned from points. However, applying self-supervised learning to temporal 3D sequences (*i.e.* point cloud, 3D skeleton) has not been fully explored. One probable reason is that self-supervised learning on high-dimensional 3D

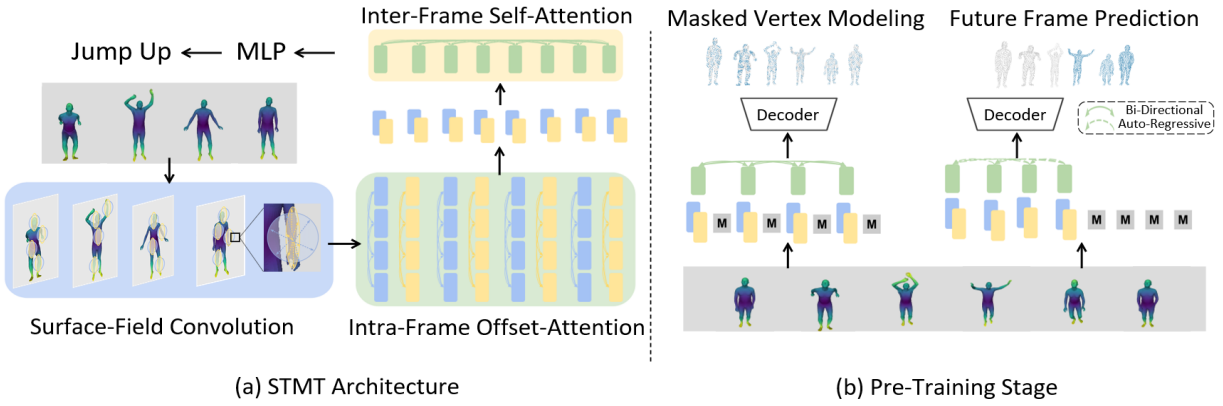

Figure 5.2: Overview of the proposed framework. **(a) Overview of *STMT*.** Given a mesh sequence, we first develop vertex patches by extracting both intrinsic (geodesic) and extrinsic (euclidean) features using surface field convolution. The intrinsic and extrinsic features are denoted by yellow and blue blocks respectively. Those patches are used as input to the intra-frame offset-attention network to learn appearance features. Then we concatenate intrinsic patches and extrinsic patches of the same position. The concatenated vertex patches (green blocks) are fed into the inter-frame self-attention network to learn spatial-temporal correlations. Finally, the local and global features are mapped into action predictions by MLP layers. **(b) Overview of Pre-Training Stage.** We design two pretext tasks: masked vertex modeling and future frame prediction for global context learning. Bidirectional attention is used for the reconstruction of masked vertices. Auto-regressive attention is used for the future frame prediction task.

temporal sequences is computationally-expensive. In this work, we propose an effective and efficient self-supervised learning method based on masked vertex modeling and future frame prediction.

5.3 Method

5.3.1 Overview

In this section, we describe our model for mesh-based action recognition, which we call *STMT*. The inputs of our model are temporal mesh sequences: $\mathbf{M} = ((\mathbf{P}_1, \mathbf{A}_1), (\mathbf{P}_2, \mathbf{A}_2), \cdots, (\mathbf{P}_t, \mathbf{A}_t))$, where t is the frame number. $\mathbf{P}_i \in \mathbb{R}^{N\times 3}$ represents the vertex positions in Cartesian coordinates, where N is the number of vertices. $\mathbf{A}_i \in \mathbb{R}^{N\times N}$ represents the adjacency matrix of the mesh. Element $\mathbf{A}_i^{mn} \in \mathbf{A}_i$ is one when there is an edge from vertex V_m to vertex V_n, and zero

when there is no edge. The mesh representation with vertices and their adjacent matrix is a unified format for various body models such as SMPL [189], SMPL-H [242], and SMPL-X [222]. In this work, we use SMPL-H body models from AMASS [194] to obtain the mesh sequences, but our method can be easily adapted to other body models.

Mesh's local connectivity provides fine-grained information. Previous methods [98, 261] proved that explicitly using surface (*e.g.*, mesh) connectivity information can achieve higher accuracy in shape classification and segmentation tasks. However, classifying temporal mesh sequences is a more challenging problem, as there is no vertex-level correspondence across frames. This prevents graph-based models from directly aggregating vertices in the temporal dimension. Therefore, we propose to first leverage mesh connectivity information to build patches at the frame level, then use a hierarchical transformer which can freely attend to any intra- and inter-frame patches to learn spatial-temporal associations. In summary, it has the following key components:

- **Surface Field Convolution** to form local vertex patches by considering both intrinsic and extrinsic mesh representations.
- **Hierarchical Spatial-Temporal Transformer** to learn spatial-temporal correlations of vertex patches.
- **Self-Supervised Pre-Training** to learn the global context in terms of appearance and motion.

See Figure 5.2 for a high-level summary of the model, and the sections below for more details.

5.3.2 Surface Field Convolution

Because displacements in grid data are regular, traditional convolutions can directly learn a kernel for elements within a region. However, mesh vertices are unordered and irregular. Considering the special mesh representations, we represent each vertex by encoding features from its neighbor vertices inspired by [232, 233]. To fully utilize meshes' local connectivity information, we consider the mesh properties of extrinsic curvature of submanifolds and intrinsic curvature of the manifold itself. Extrinsic curvature between two vertices is approximated using Euclidean distance. Intrinsic curvature is approximated using Geodesic distance, which is defined as the shortest path between two vertices on mesh surfaces. We propose a light-weighted surface field convolution to build local patches, which can be denoted as:

$$\boldsymbol{F}_{VG}^{\prime(x,y,z)} = \sum_{(\delta_x,\delta_y,\delta_z)\in G(x,y,z)} \boldsymbol{W}^{(\delta_x,\delta_y,\delta_z)} \cdot \boldsymbol{F}^{(x+\delta_x,y+\delta_y,z+\delta_z)} \tag{5.1}$$

$$\boldsymbol{F}_{VE}^{\prime(x,y,z)} = \sum_{(\zeta_x,\zeta_y,\zeta_z)\in E(x,y,z)} \boldsymbol{W}^{(\zeta_x,\zeta_y,\zeta_z)} \cdot \boldsymbol{F}^{(x+\zeta_x,y+\zeta_y,z+\zeta_z)} \tag{5.2}$$

G and E is the local region around vertex (x, y, z). In this paper, we use k-nearest-neighbor to sample local vertices. $(\delta_x, \delta_y, \delta_z)$ and $(\zeta_x, \zeta_y, \zeta_z)$ represent the spatial displacement in geodesic and euclidean space, respectively. $\boldsymbol{F}^{(x,y,z)}$ denotes the feature of the vertex at position (x, y, z).

5.3.3 Hierarchical Spatial-Temporal Transformer

We propose a hierarchical transformer that consists of intra-frame and inter-frame attention. The basic idea behind our transformer is three-fold: (1) Intra-frame attention can encode connectivity information from the adjacency matrix, while such information can not be directly aggregated in the temporal domain because vertices are unordered. (2) Frame-level offset-attention can be used to mimic the Laplacian operator to learn effective spatial representations. (3) Inter-frame self-attention can learn feature correlations in the spatial-temporal domain.

Intra-Frame Offset-Attention

Graph convolution networks [24] show the benefits of using a Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{E}$ to replace the adjacency matrix $\mathbf{E}$, where $\mathbf{D}$ is the diagonal degree matrix. Inspired by this, offset-attention has been proposed and achieved superior performance in point-cloud classification and segmentation tasks [93]. We adapt offset-attention to attend to vertex patches. Specifically, the offset-attention layer calculates the offset (difference) between the self-attention (SA) features and the input features by element-wise subtraction. Offset-attention is denoted as:

$$\boldsymbol{F}_{out} = OA(\boldsymbol{F}_{in}) = \phi(\boldsymbol{F}_{in} - \boldsymbol{F}_{sa}) + \boldsymbol{F}_{in}. \tag{5.3}$$

where ϕ denotes a non-linear operator. $\boldsymbol{F}_{in} - \boldsymbol{F}_{sa}$ is proved to be analogous to discrete Laplacian operator [93], *i.e.* $\boldsymbol{F}_{in} - \boldsymbol{F}_{sa} \approx \boldsymbol{L}\boldsymbol{F}_{in}$. As Laplacian operators in geodesic and euclidean space are expected to be different, we propose to use separate transformers to model intrinsic patches and extrinsic patches. Specifically, the aggregated feature for vertex V is denoted as:

$$\boldsymbol{F}_{V}^{\prime(x,y,z)} = OA_G(\boldsymbol{F}_{VG}^{\prime(x,y,z)}) \oplus OA_E(\boldsymbol{F}_{VE}^{\prime(x,y,z)}) \tag{5.4}$$

Here $F_{VG}^{\prime(x,y,z)} \in \mathbb{R}^{N\times d_g}$ and $F_{VE}^{\prime(x,y,z)} \in \mathbb{R}^{N\times d_e}$ are local patches learned using Equ. 5.1 and Equ. 5.2. $F_{V}^{\prime(x,y,z)} \in \mathbb{R}^{N\times d}$ denotes the local patch for position (x, y, z), where $d = d_g + d_e$. The weights of OA_G and OA_E are not shared.

Inter-Frame Self-Attention

Given $F_V^{\prime}$ which encodes local connectivity information, we use self-attention (SA) [295] to learn semantic affinities between different vertex patches across frames. Specifically, let $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ be the *query, key* and *value*, which are generated by applying linear transformations to the input features $F_V^{\prime} \in \mathbb{R}^{N\times d}$ as follows:

$$\begin{aligned}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) &= F_V^{\prime} \cdot (\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v) \\ \boldsymbol{Q}, \boldsymbol{K} &\in \mathbb{R}^{N\times d_a}, \quad \boldsymbol{V} \in \mathbb{R}^{N\times d} \\ \boldsymbol{W}_q, \boldsymbol{W}_k &\in \mathbb{R}^{d\times d_a}, \quad \boldsymbol{W}_v \in \mathbb{R}^{d\times d}\end{aligned} \tag{5.5}$$

where $\boldsymbol{W}_q$, $\boldsymbol{W}_k$ and $\boldsymbol{W}_v$ are the shared learnable linear transformation, and d_a is the dimension of the query and key vectors. Then we can use the query and key matrices to calculate the attention weights via the matrix dot-product:

$$\boldsymbol{A} = (\tilde{\alpha})_{i,j} = \mathrm{softmax}(\frac{\boldsymbol{Q} \cdot \boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_a}}). \tag{5.6}$$

$$\boldsymbol{F}_{sa} = \boldsymbol{A} \cdot \boldsymbol{V} \tag{5.7}$$

The self-attention output features $\boldsymbol{F}_{sa}$ are the weighted sums of the value vector using the corresponding attention weights. Specifically, for a vertex patch in position (x, y, z), its aggregated feature after inter-frame self-attention can be computed as: $\boldsymbol{F}_{sa}^{(x,y,z)} = \sum \boldsymbol{A}^{(x,y,z),(x',y',z')} \times \boldsymbol{V}^{(x',y',z')}$, where (x', y', z') belongs to the Cartesian coordinates of $\boldsymbol{F}_V^{\prime}$.

5.3.4 Self-Supervised Pre-Training

Self-supervised learning has achieved remarkable results on large-scale image datasets [102]. However, self-supervised learning for temporal 3D sequences (*i.e.* point cloud, 3D skeleton) remains to be challenging and has not been fully explored. There are two possible reasons: (1) self-supervised learning methods rely on large-scale datasets to learn meaningful patterns [52]. However, existing MoCap benchmarks are relatively small compared to 2D datasets like ImageNet [62]. (2) Self-supervised learning for 3D data sequences is computationally expensive in terms of memory and speed. In this work, we first propose a simple and effective method to

augment existing MoCap sequences, and then define two effective and efficient self-supervised learning tasks, namely masked vertex modeling and future frame prediction, which enable the model to learn global context. The work that is close to us is OcCO [299], which proposed to use occluded point cloud reconstruction as the pretext task. OcCO has a computationally-expensive process to generate occlusions, including point cloud projection, occluded point calculation, and a mapping step to convert camera frames back to world frames. Different from OcCO, we randomly mask vertex patches or future frames on the fly, which saves the pre-processing step. Moreover, our pre-training method is designed for temporal mesh sequences and considers both bi-directional and auto-regressive attention.

Data Augmentation through Joint Shuffle

Considering the flexibility of SMPL-H representations, we propose a simple yet effective approach to augment SMPL-H sequences by shuffling body pose parameters. Specifically, we split SMPL-H pose parameters into five body parts: bone, left/right arm, and left/right leg. We use $I_{bone}, I_{leg}^{left}, I_{leg}^{right}, I_{arm}^{left}, I_{arm}^{right}$ to denote the SMPL-H pose indexes of the five body parts. Then we synthesize new sequences by randomly selecting body parts from five different sequences. We keep the temporal order for each part such that the merged action sequences have meaningful motion trajectories. The input to Joint Shuffle are SMPL-H pose parameters $\theta \in \mathbb{R}^{b\times t\times n\times 3}$, where b is the sequence number, t is the frame number, and n is the joint number. We randomly select the shape β and dynamic parameters ϕ from one of the five SMPL-H sequences to compose a new SMPL-H body model. Given b SMPL-H sequences, we can synthesize ${}^{b}C_5 = \frac{b!}{5!(b-5)!}$ number of new sequences. We prove that the model can benefit from large-scale pre-training in Section 5.4.6.

Masked Vertex Modeling with Bi-Directional Attention

To fully activate the inter-frame bi-directional attention in the transformer, we design a self-supervised pretext task named Masked Vertex Modeling (MVM). The model can learn human prior information in the spatial dimension by reconstructing masked vertices of different body parts. We randomly mask r percentages of the input vertex patches, and force the model to reconstruct the full sequences. Moreover, we use bi-directional attention to learn correlations among all remaining local patches. Each patch will attend to all patches in the entire sequence. It models the joint distribution of vertex patches over the whole temporal sequences x as the following product of conditional distributions, where x_i is a single vertex patch:

Algorithm 1 Pseudocode of STMT Joint Shuffle

function STMT_JOINT_SHUFFLE($\theta \in \mathbb{R}^{b\times t\times n\times 3}, I_{bone}, I_{leg}^{left}, I_{leg}^{right}, I_{arm}^{left}, I_{arm}^{right}$)
  $\theta_s \leftarrow random\_sample(\theta, 5)$ ▷ $\theta_s \in \mathbb{R}^{5\times t\times n\times 3}$, randomly sample five SMPL-H sequences
  $t_{max} \leftarrow get\_max\_length(\theta_s)$ ▷ compute the maximum sequence length in θ_s
  $\theta_{new} \leftarrow Initialize(t_{max}, n, 3)$
  $P \leftarrow \{I_{bone}, I_{leg}^{left}, I_{leg}^{right}, I_{arm}^{left}, I_{arm}^{right}\}$
  for i in $0, 1, 2, 3, 4$ **do**
    $\theta_s \leftarrow repeat(\theta_s[i], (t_{max}, n, 3))$ ▷ pad each sequence to the max length using repeating
    $\theta_{new}[P[i]] \leftarrow \theta_s[i][P[i]]$ ▷ assign the body-part sequence
  return θ_{new}

$$p(x) = \prod_{i=1}^{N} p(x_i|x_1, .., x_i, ..., x_N). \tag{5.8}$$

Where N is the number of patches in the entire sequence x after masking. Every patch will attend to all patches in the entire sequence. In this way, bi-directional attention is fully-activated to learn spatial-temporal features that can accurately reconstruct completed mesh sequences.

Future Frame Prediction with Auto-Regressive Attention

The masked vertex modeling task is to reconstruct masked vertices in different body parts. The model can reconstruct completed mesh sequences if it captures the human body prior or can make a movement inference from nearby frames. As action recognition requires the model to understand the global context, we propose the future frame prediction (FFP) task. Specifically, we mask out all the future frames and force the transformer to predict the masked frames. Moreover, we propose to use auto-regressive attention for the future frame prediction task, inspired by language generation models like GPT-3 [23]. However, directly using RNN-based models [49] in GPT-3 to predict future frames one by one is inefficient, as 3D mesh sequences are denser compared to language sequences. Therefore, we propose to reconstruct all future frames in a single forward pass. For auto-regressive attention, we model the joint distribution of vertex patches over a mesh sequence x as the following product of conditional distributions, where x_i is a single patch at frame t_i:

$$p(x) = \prod_{i=1}^{N} p(x_i | x_1, x_2, ..., x_M). \tag{5.9}$$

Where N is the number of patches in the entire sequence x after masking. $M = (t_i - 1) \times n$, where n is the number of patches in a single frame. Each vertex patch depends on all patches that are temporally before it. The auto-regressive attention enables the model to predict movement patterns and trajectories, which is beneficial for the downstream action recognition task.

5.3.5 Training

In the pre-training stage, we use PCN [329] as the decoder to reconstruct masked vertices and predict future frames. The decoder is shared for the two pretext tasks. Since mesh vertices are unordered, the reconstruction loss and future prediction loss should be permutation-invariant. Therefore, we use Chamfer Distance (CD) as the loss function to measure the difference between the model predictions and ground truth mesh sequences.

$$\begin{aligned} CD(M_{pred}, M_{gt}) = \frac{1}{|M_{pred}|} \sum_{x \in M_{pred}} \min_{y \in M_{gt}} ||x - y||_2 + \\ \frac{1}{|M_{gt}|} \sum_{y \in M_{gt}} \min_{x \in M_{pred}} ||y - x||_2 \end{aligned} \tag{5.10}$$

CD (5.10) calculates the average closest euclidean distance between the predicted mesh sequences M_{pred} and the ground truth sequences M_{gt}. The overall loss is a weighted sum of masked vertex reconstruction loss and future frame prediction loss:

$$L = \lambda_1 CD(M_{pred}^{MVM}, M_{gt}) + \lambda_2 CD(M_{pred}^{FFP}, M_{gt}) \tag{5.11}$$

In the fine-tuning stage, we replace the PCN decoder with an MLP head. Cross-entropy loss is used for model training.

5.4 Experiment

5.4.1 Datasets

Following previous MoCap-based action recognition methods [231, 274], we evaluate our model on the most widely used benchmarks: KIT[197] and BABEL [231]. **KIT** is one of the largest

Method	Input	KIT	
		Top-1 (%)	Top-5 (%)
2s-AGCN-FL [264] (CVPR'19)	3D Skeleton	42.44	75.60
2s-AGCN-CE [264] (CVPR'19)	3D Skeleton	57.46	81.54
CTR-GCN [42] (ICCV'21)	3D Skeleton	64.65	87.90
MS-G3D [188] (CVPR'20)	3D Skeleton	65.38	87.90
PSTNet[75] (ICLR'21)	Point Cloud	56.93	88.21
SequentialPointNet[158] (arXiv'21)	Point Cloud	59.75	88.01
P4Transformer[74] (CVPR'21)	Point Cloud	62.15	88.01
STMT(Ours)	Mesh	**65.59**	**90.09**

Table 5.1: Experimental Results on KIT and BABEL Dataset.

MoCap datasets. It has 56 classes with 6,570 sequences in total. (2) **BABEL** is the largest 3D MoCap dataset that unifies 15 different datasets. BABEL has 43 hours of MoCap data performed by over 346 subjects. We use the 60-class subset from BABEL, which contains 21,653 sequences with single-class labels. We randomly split each dataset into training, test, and validation set, with ratios of 70%, 15%, and 15%, respectively. Note that existing action recognition datasets with skeletons only are not suitable for our experiments, as they do not provide full 3D surfaces or SMPL parameters to obtain the mesh representation.

Motion Representation. Both KIT and BABEL's MoCap sequences are obtained from AMASS dataset in SMPL-H format. A MoCap sequence is an array of pose parameters over time, along with the shape and dynamic parameters. For skeleton-based action recognition, we follow previous work [231] which predicted the 25-joint skeleton from the vertices of the SMPL-H mesh. The movement sequence is represented as $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_L)$, where $\mathbf{x}_i \in \mathbb{R}^{J\times 3}$ represents the position of the J joints in the skeleton in Cartesian coordinates. For point-cloud-based action recognition, we directly use the vertices of SMPL-H model as the model input. The point-cloud sequence is represented as $\mathbf{P} = (\mathbf{p}_1, \cdots, \mathbf{p}_L)$, where $\mathbf{p}_i \in \mathbb{R}^{V\times 3}$, and V is the number of vertices. For mesh-based action recognition, we represent the motion as a series of mesh vertices and their adjacent matrix over time, as introduced in Section 5.3.1. See Sup. Mat. for more details about datasets and pre-processing.

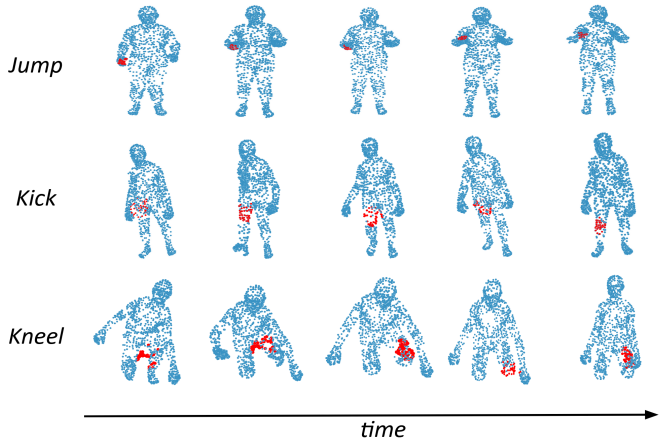

Figure 5.3: Visualization of inter-frame attention. Red denotes the highest attention.

5.4.2 Baseline Methods

We compare our model with state-of-the-art 3D skeleton-based and point cloud-based action recognition models, as there is no existing literature on mesh-based action recognition. 2s-AGCN [264], CTR-GCN [42], and MS-G3D [188] are used as skeleton-based baselines. Among those methods, 2s-AGCN trained with focal loss and cross-entropy loss are used as benchmark methods in the BABEL dataset [231]. For the comparison with point-cloud baselines, we choose PSTNet [75], SequentialPointNet[158], and P4Transformer [74]. Those methods achieved top performance on common point-cloud-based action recognition benchmarks.

5.4.3 Implementation Details

For skeleton-based baselines, we use the official implementations of 2s-ACGN, CTR-GCN, and MS-G3D. For point-cloud-based baselines, we use the official implementations of PSTNet, SequentialPointNet, P4Transformer. We pre-train *STMT* for 200 epochs with a batch size of 32. The model is fine-tuned for 50 epochs with a batch size of 64. Adam optimizer [135] is used with a learning rate of 0.0001 for both pre-training and fine-tuning. See Sup. Mat. for more implementation details.

5.4.4 Main Results

Comparison with State-of-the-Art Methods. As indicated in Table 5.1, *STMT* outperforms all other state-of-the-art models. Our model can outperform point-cloud-based models by 3.44%

Intrinsic	Extrinsic	MVM	FFP	Top-1 (%)
✓				63.40
✓	✓			64.03
✓	✓	✓		64.96
✓	✓		✓	64.13
✓	✓	✓	✓	**65.59**

Table 5.2: Performance of ablated versions. Intrinsic and Extrinsic stand for the intrinsic (geodesic) and extrinsic (euclidean) features in surface field convolution. MVM stands for Masked Vertex Modelling. FFP stands for Future Frame Prediction.

and 4.11% on KIT and BABEL datasets in terms of top-1 accuracy. Moreover, compared to skeleton-based methods which involve manual efforts to convert mesh vertices to skeleton representations, our model achieves better performance by directly modeling the dynamics of raw mesh sequences.

We visualize the inter-frame attention weights of our hierarchical transformer in Figure 5.3. We observe that the model can pay attention to key regions across frames. This supports the intuition that our hierarchical transformer can take the place of explicit vertex tracking by learning spatial-temporal correlations.

5.4.5 Ablation Study

Ablation Study of *STMT*. We test various ablations of our model on the KIT dataset to substantiate our design decisions. We report the results in Table 6.3. Note that Joint Shuffle is used in all of the self-supervised learning experiments (last three rows). We observe that each component of our model gains consistent improvements. The comparison of the first two rows proves the effectiveness of encoding both intrinsic and extrinsic features in vertex patches. Comparing the last three rows with the second row, we observe a consistent improvement using self-supervised pre-training. Moreover, the downstream task can achieve better performance with MVM compared to FFP. One probable reason is that the single task for future frame prediction is more challenging than masked vertex modeling, as the model can only see the person movement in the past. The model can achieve the best performance with both MVM and FFP, which demonstrates that the two self-supervised tasks are supplementary to each other.

Method	Top-1 (%)
w/o pre-training	64.03
pre-training w/o JS	64.13
pre-training w/ JS	**65.59**

Table 5.3: Comparison of Different Pre-Training Strategies. JS stands for Joint Shuffle.

r	Pre-Train Loss ($\times 10^4$)	Fine-Tune Accuracy (%)
0.1	0.39	64.44
0.3	0.41	64.55
0.5	0.40	**65.59**
0.7	0.43	64.19
0.9	0.48	65.07
Rand	0.43	64.75

Table 5.4: Effect of Different Masking Ratios.

5.4.6 Analysis

Different Pre-Training Strategies. We pre-train our model with different datasets and summarize the results in Table 5.3. The first row shows the case without pre-training. The second shows the result for the model pre-trained on the KIT dataset (without Joint Shuffle augmentation). The third shows the result for the model pre-trained on KIT dataset (with Joint Shuffle). We observe our model can achieve better performance with Joint Shuffle, as it can synthesize large-scale mesh sequences.

Different Masking Ratios. We investigate the impact of different masking ratios. We report the converged pre-training loss and the fine-tuning top-1 classification accuracy on the test set in Table 5.4. We also experiment with the random masking ratio in the last row. For each forward pass, we randomly select one masking ratio from 0.1 to 0.9 with step 0.1 to mimic flexible masked token length. The model with a random masking ratio does not outperform the best model that is pre-trained using a single ratio (*i.e.* 0.5). We observe that as the masking ratio increases, the pre-training loss mostly increases as the task becomes more challenging. However, a challenging self-supervised learning task does not necessarily lead to better performance. The model with a masking ratio of 0.7 and 0.9 have a high pre-train loss, while the fine-tune accuracy is not higher than the model with a 0.5 masking ratio. The conclusion is similar to

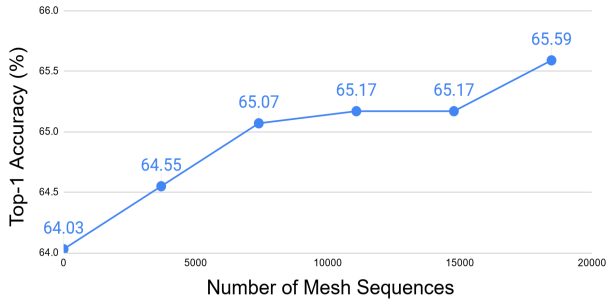

Figure 5.4: Effect of Different Number of Mesh Sequences.

Method	Input	Top-1 (%)
2s-AGCN-FL [264]	3D Skeleton	58.67
2s-AGCN-CE [264]	3D Skeleton	57.49
CTR-GCN [42]	3D Skeleton	62.25
MS-G3D [188]	3D Skeleton	60.01
PSTNet[75]	Point Cloud	51.48
SequentialPointNet[158]	Point Cloud	60.60
P4Transformer[74]	Point Cloud	57.84
STMT(Ours)	Mesh	**64.04**

Table 5.5: Experimental results on body poses estimated by VIBE [136] on NTU RGB+D dataset. The skeleton, point cloud, and mesh representations are derived from the same noisy body estimations.

the comparison of MVM and FFP training objectives, where a more challenging self-supervised learning task may not be optimal.

Different Number of Mesh Sequences for Pre-Training. We test the effect of different numbers of mesh sequences used in pre-training. We report the fine-tuning top-1 classification accuracy in Figure 5.4. We observe that a large number of pre-training data can bring substantial performance improvement. The proposed Joint Shuffle method can greatly enlarge the dataset size without any manual cost, and has the potential to further improve model performance.

Experimental Results on Noisy Body Pose Estimations. Body pose estimation has been a popular research field [125, 136, 161], but how to leverage the estimated 3D meshes for downstream perception tasks has not been fully explored. We apply the state-of-the-art body pose estimation model VIBE [136] on videos of NTU RGB+D dataset to obtain 3D mesh sequences. Skeleton and point cloud representations are derived from the estimated meshes to train the

baseline models (see Sup. Mat.). We report the results in Table 5.5. We observe that *STMT* can outperform the best skeleton-based and point cloud-based action recognition model by 1.79% and 3.44% respectively. This shows that *STMT* with meshes as input, is more robust to input noise compared to other state-of-the-art methods with 3D skeletons or point clouds as input.

5.5 Conclusion

In this work, we propose a novel approach for MoCap-based action recognition. Unlike existing methods that rely on skeleton representation, our proposed model directly models the raw mesh sequences. Our method encodes both intrinsic and extrinsic features in vertex patches, and uses a hierarchical transformer to freely attend to any two vertex patches in the spatial and temporal domain. Moreover, two self-supervised learning tasks, namely Masked Vertex Modeling and Future Frame Prediction are proposed to enforce the model to learn global context. Our experiments show that *STMT* can outperform state-of-the-art skeleton-based and point-cloud-based models.

Chapter 6

Generalized human action recognition by jointly modeling videos and 3D meshes

In this chapter, we propose a generalized human action recognition framework by jointly considering RGB videos and estimated 3D meshes. We demonstrate that 2D and 3D representations are complementary to each other, even when the 3D representations are noisy estimations. The model is able to perform generalized human action analysis and bridge challenging domain gaps, such as sim-to-real transfer.

6.1 Overview

Recent advancements in action recognition using deep learning have shown promising results [29, 99, 285]. However, the efficacy of these methods heavily depends on the availability of extensive labeled data specific to the target domain, which is often unfeasible in real-world scenarios. For instance, identifying novel actions can be costly due to the necessity of hiring actors to perform predefined actions. Furthermore, in order to achieve reliable performance, it is necessary to ensure the consistency between the training and test data in terms of the environment, subjects, and camera perspectives. This adds further complexity [211, 284] to the data collection efforts. These limitations make the data collection process time-consuming, labor-intensive, and sometimes unattainable, particularly in cases involving harmful or violent activities.

Incorporating synthetic data brings significant advantages, including reducing data collection efforts, producing well-aligned training data across multiple modalities, and facilitating the recognition of rare or violent actions. Despite the practical benefits associated with synthetic data, there are several challenges in training a model that can effectively handle diverse scenarios when only labeled synthetic data are available. Primarily, domain gaps may arise due to various factors such as different backgrounds, lighting conditions, viewpoints, interacted objects, and motion variances.

To address these challenges, research in video domain adaptation has explored various strategies to mitigate domain gaps from multiple perspectives. One of the main streams is cross-viewpoint domain adaptation [138, 151, 181, 236, 268], and the research efforts have concentrated on learning geometric transformations of camera viewpoints. However, these work often overlooked other domain shifts, such as environmental differences. In pursuit of viewpoint-invariant representations, some studies utilized 3D representations such as skeletons and human meshes as model inputs [181, 268, 351]. Another line of research in video domain adaptation focuses on addressing environmental changes. Recent works in this domain have employed adversarial training methods for domain alignment [37, 120, 219, 352]. Notably, techniques such as the Gradient Reverse Layer [151] have been adapted to architectures like C3D [286], TRN [345], or both [219]. Chen *et al.* [37] proposed an attention-based model to capture temporal dynamics in videos, while Pan *et al.* [219] introduced a cross-domain attention mechanism to discern relevant information. In contrast to previous methodologies that relied solely on complete videos as model inputs, we propose a novel approach that integrates both 2D RGB videos and 3D meshes to learn viewpoint-invariant representations.

To this end, we propose a multi-modal action recognition model which jointly takes RGB videos and 3D meshes as inputs to learn domain-invariant representations. There are two parallel action recognition branches based on RGB and mesh respectively. For RGB-based action recognition, we use a light-weighted student model (*i.e.* I3D) for its real-time inference speed. For mesh-based action recognition, we leverage a spatial-temporal transformer named *STMT*, which is proposed in our previous work [351]. It consumes meshes as inputs to learn domain-invariant representations and serves as a strong teacher model. The student model distills the pseudo-labels produced by the teacher model, and thus achieves better performance while maintaining the real-time inference speed. We evaluate our proposed method on the Mixamo→Kinetics dataset dataset, which consists of a variety of real and synthetic videos of human actions. We show that our model achieves superior performance compared to other state-of-the-art baselines. In short, our contributions are three-fold:

- We propose a novel unsupervised video domain adaptation model which consumes RGB videos and estimated 3D meshes to learn generalized feature representations.
- We propose a multi-modal distillation framework where the light-weighted student model can learn from the strong teacher model to achieve reliable performance.
- Our model, achieves superior performance compared to other domain adaptation baselines while maintaining real-time inference speed. This enables our model to be deployed in practice.

6.2 Related Work

Data-Efficient Action Recognition. The effectiveness of human action recognition significantly relies on the availability of sufficient training data, which is not always feasible in real-world scenarios. Considering the real-world data limitations, various data-efficient learning methodologies [4, 28, 85, 196, 208, 349] have been proposed which achieved promising performance. Zero-shot action recognition approaches [85, 196, 208] were proposed to learn semantic representations from word vectors or annotated attributes to identify unseen action categories. However, a notable performance gap exists between zero-shot models and those trained with real videos, due to the inherent ambiguity present in texts. Few-shot action recognition models, on the other hand, [4, 28] have employed episodic training to acquire a metric [65, 225, 326] or an optimizer [80, 247] from a set of base tasks, enabling learning of novel categories with only a few real videos available. Recently, weakly- and webly-supervised approaches have raised considerable research interest [35, 278, 296, 314]. Common practices in these methodologies include outlier removal [314] and label correction [278, 296]. Chen et al. [35] proposed a method to denoise unannotated web training data by transferring learned similarities from a clean set of base categories. While traditional few-shot learning approaches necessitate annotated videos in base categories and genuine videos in novel categories, we diverge by leveraging synthetic videos, which can be generated indefinitely with minimal annotation costs.

Video Domain Adaptation. Video domain adaptation has been studied to bridge domain gaps from different perspectives. One of the important tasks is cross-viewpoint domain adaptation [138, 151, 181, 236, 268]. These works focused on learning geometric transformations of a camera but neglected other domain shifts such as environment differences. To learn viewpoint-invariant representations, 3D representations such as skeletons and human meshes are used as

model inputs [181, 268]. The other stream for video domain adaptation focuses on environmental changes. Some of the recent works applied adversarial training for domain alignment [37, 120, 219]. Specifically, Gradiant Reverse Layer [151] was adapted to C3D [286], TRN [345] or both [219] architectures. Chen *et al.* [37] proposed an attention-based model to attend to the temporal dynamics of videos. Pan *et al.* [219] introduced a cross-domain attention model to learn relevant information. In contrast to previous literature which used complete videos as model inputs, we propose a model which leverages both 2D RGB videos and 3D meshes as inputs to learn viewpoint-invariant representations.

Action Recognition using 3D Representations. To learn generalized feature representations, various models for 3D action recognition have been proposed, leveraging multiple 3D modalities. These can be broadly categorized into depth-based approaches [180, 250, 251, 304, 315], skeleton-based methods [68, 106, 178, 263, 303, 322, 336], and point-cloud-based techniques [75, 185, 233, 308]. Depth-based models utilize depth representations to capture reliable 3D geometric cues that remain robust across different viewpoints. For instance, MVDI [315] extracted multi-view images from depth videos for 3D action recognition. Meanwhile, skeleton sequences, capturing spatial-temporal information, have demonstrated resilience against scene and viewpoint variations. Various architectures have been explored for incorporating skeleton representations. Some methods treated skeleton sequences as time-series inputs to Recurrent Neural Networks (RNNs) [68, 178, 336], while others transformed them into spectral images and trajectory maps for feature learning with Convolutional Neural Networks (CNNs) [106, 303]. Additionally, Yan et al. [322] utilized Graph Convolutional Networks (GCNs) to model joint dependencies naturally encoded within graphs. Another popular approach in 3D human action recognition involves point clouds. Pioneering works like PointNet [232] and PointNet++ [233] encoded 3D point sets in a permutation-invariant manner. In this study, we propose leveraging our previous work on mesh-based action recognition [351] to learn domain-invariant representations. 3D mesh representations offer viewpoint-invariant characteristics compared to 2D RGB representations, enhancing the generalization ability of our model, which jointly learns from 2D and 3D synthetic motion sequences.

Synthetic Humans. Synthetic humans have been widely used across various computer vision tasks, including body pose estimation [41, 86, 177, 226, 266, 293], depth estimation [292], pedestrian detection [198, 226], trajectory forecasting [168, 169], person re-identification [234], and face recognition [140, 199]. Despite these applications, the utilization of synthetic data for action recognition remains relatively unexplored. Previous methodologies predominantly re-

lied on pre-made 3D models sourced from platforms like the Unity Asset Store, encompassing human models, object models, and texture models, to generate synthetic data [60, 83, 244]. De Souza et al. [60] proposed a model capable of jointly predicting real and synthetic action classes within a multi-task framework, while [230] utilized the Unity3D game engine to programmatically define synthetic activities. More recently, Varol et al. [294] extracted motion sequences from genuine data and augmented the estimated sequences with novel viewpoints to address the challenge of cross-view action recognition.

6.3 Method

In this section, we describe our domain-adaptive action recognition model. We illustrate the difference between our proposed method and previous action recognition models in Figure 7.3. (a) shows one of the most popular action recognition model (*i.e.* I3D [29]). I3D is a light-weighted model which jointly takes RGB and estimated flows as model inputs. The model can achieve real-time inference speed. However, it suffers from domain generalization problems caused by novel viewpoints and scenes. (b) is a multimodal domain generalization model [350]. It has two branches. The first is an I3D model. The second is a Spatial-Temporal Mesh Transformer, namely *STMT*[351]. It takes the 3D meshes estimated from RGB videos as model inputs. Those three branches are jointly trained. During inference, the model takes RGB, flows, and estimated 3D meshes as inputs. The model can generalize well to novel viewpoints and scenarios, thanks to the robust 3D representations. (c) The proposed method. We aim at developing a model that can generalize to novel scenes and achieve real-time inference speed. There are two parallel action recognition branches based on RGB and mesh respectively. For RGB-based action recognition, we use a light-weighted student model (*i.e.* I3D) for its real-time inference speed. For mesh-based action recognition, we leverage a spatial-temporal transformer named *STMT*, which is proposed in our previous work [351]. It consumes meshes as inputs to learn domain-invariant representations and serves as a strong teacher model. The student model distills the pseudo-labels produced by the teacher model. During inference, only the student model (I3D) is used and thus achieves better performance while maintaining the real-time inference speed.

See Figure 1 for a high-level summary of the model, and the sections below for more details.

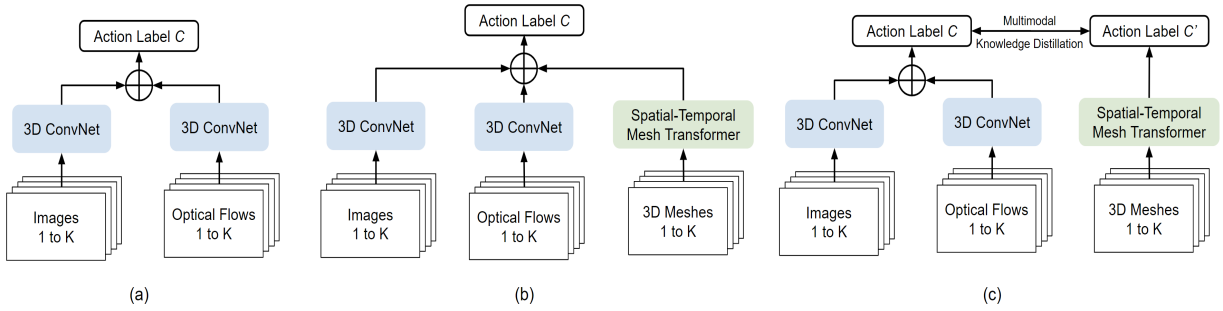

Figure 6.1: Overview of previous action recognition methods [29, 350] and the proposed method. (a) RGB-based action recognition model (*i.e.* I3D [29]). I3D is a light-weighted model which jointly takes RGB and estimated flows as model inputs. The model can achieve real-time inference speed. However, it suffers from domain generalization problems caused by novel viewpoints and scenes. (b) is a domain generalization model [350]. It has two branches. The first is an I3D model. The second is a Spatial-Temporal Mesh Transformer, namely *STMT* [351]. It takes the 3D meshes estimated from RGB videos as model inputs. Those three branches are jointly trained. During inference, the model takes RGB, flows, and estimated 3D meshes as inputs. The model can generalize well to novel viewpoints and scenarios, thanks to the robust 3D representations. (c) The proposed method. We aim at developing a model that can generalize to novel scenes and achieve real-time inference speed. There are two parallel action recognition branches based on RGB and mesh respectively. For RGB-based action recognition, we use a light-weighted student model (*i.e.* I3D) for its real-time inference speed. For mesh-based action recognition, we leverage a spatial-temporal transformer named *STMT*, which is proposed in our previous work [351]. It consumes meshes as inputs to learn domain-invariant representations and serves as a strong teacher model. The student model distills the pseudo-labels produced by the teacher model. During inference, only the student model (I3D) is used and thus achieves better performance while maintaining the real-time inference speed.

6.3.1 Action Recognition from RGB Videos

We use I3D [29] as our encoder of RGB videos, considering its real-time inference speed. The architecture of I3D is shown in Figure 7.3 (a). It leverges two parallel branches to encoder RGB frames and optical flow respectively. The final predictions are fused at the end to get the final action label. While I3D achieves real-time inference speed, it is not able to generalize well to

novel domains.

6.3.2 Action Recognition from 3D Meshes

Considering the limitation of RGB-based action recognition, we propose to leverage an extra modality: 3D meshes. 3D meshes can either be exported from game simulators without costs, or be obtained from body pose estimation models when the MoCap sensors are not availble. Specifically, we represent the 3D meshes as: $\mathbf{M} = ((\mathbf{P}_1, \mathbf{A}_1), (\mathbf{P}_2, \mathbf{A}_2), \cdots, (\mathbf{P}_t, \mathbf{A}_t))$, following [350, 351]. Where t is the frame number. $\mathbf{P}_i \in \mathbb{R}^{N\times 3}$ represents the vertex positions, where N is the number of vertices. $\mathbf{A}_i \in \mathbb{R}^{N\times N}$ represents the adjacency matrix of the mesh. This is a unified format for various body models such as SMPL [189], SMPL-H [242], and SMPL-X [222]. In this paper, we use the body pose estimation model named VIBE [136] to obtain the 3D meshes, but it can be easily adapted to other body pose estimation models. We leveraged a spatial-temporal transformer for mesh-based action recognition, which was inspired by our previous work STMT[351]. It consumes mesh sequences as inputs, and predicts the action labels. we first develop vertex patches by extracting both intrinsic (geodesic) and extrinsic (euclidean) features using surface field convolution [351]. Those patches are used as input to the intra-frame attention network to learn appearance features. Then we concatenate intrinsic patches and extrinsic patches of the same position. The concatenated vertex patches are fed into the inter-frame self-attention network to learn spatial-temporal correlations. Finally, the local and global features are mapped into action predictions by MLP layers.

6.3.3 Multimodal Knowledge Distillation for RGB Videos and 3D Meshes

Different from previous work [350], which requires 3D meshes for both training and inference stages, we propose a multimodal knowledge distillation framework which only relies on RGB videos during inference. For the model training, we first extract the 3D meshes using a body pose estimation model. We use the mesh-based action recognition model as the teacher model because of its viewpoint-invariant representations. The mesh-based model is supervised by the labeled source videos. Then we leverage the I3D model as the student model. The I3D model takes the source videos as well as the unlabeled target videos as inputs. The pseudo labels of the unlabeled target videos come from the teacher model.

We use the cross-entropy loss to supervise the training of mesh-based action recognition

(teacher model) with labeled source data:

$$\mathcal{L}^m = \sum_{i=1}^{N} \text{CE}(cls^i_{mesh}, cls^{*i}), \tag{6.1}$$

where cls^i_{mesh} is the predicted label and cls^{*i} is the ground truth label. N is the number of training data in the labeled source domain. Once the teacher model finishes training, we obtain pseudo labels from the teacher model for the unlabeled target data. Then we use the cross-entropy loss to supervise the training of the RGB-based action recognition model with labeled source data as well as the pseudo-labeled target data:

$$\mathcal{L}^r = \sum_{i=1}^{N} \text{CE}(cls^i_{rgb}, cls^{*i}) + \sum_{i=1}^{M} \text{CE}(cls^i_{rgb}, cls^i_{mesh}), \tag{6.2}$$

where cls^i_{mesh} and cls^i_{rgb} are the predicted label and cls^{*i} is the ground truth label. During inference, we only keep the I3D student model to ensure real-time inference speed.

6.4 Experiments

6.4.1 Dataset

We evaluate our model on the sim-to-real domain adaptation benchmark, named Mixamo→Kinetics [288]. The synthetic dataset (Mixamo) consists of 24, 533 videos which are generated using the 3D characters from Mixamo. The target dataset contains 11, 662 videos from 14 action categories extracted from the Kinetics dataset [130]. The overlapping actions between the two datasets are swing dancing, breakdancing, salsa dancing, throwing, capoeira, jogging, shouting, side kick, clapping, texting, golf putting, squat, punching and backflip.

6.4.2 Baselines

We compare our proposed model with state-of-the-art models in different settings. (1) Domain adaptation baselines. We compare our model with video domain adaptation models. Those models are trained with labeled source videos and unlabeled target videos. The domain adaptation baselines include ADDA [290], TA^{3}N [37], and CO^{2}A [288] as well as their variants. Some of those models are also trained with extra supervision. Those baselines considering a weakly supervised setting, i.e. assuming that annotations are available for 5 randomly selected target instances per class, following [288]. (2) Domain generalization methods. We compare our

Method	Backbone	Weak Supervision	Val Accu (%)
ADDA[290]	I3D		11.2
ADDA[290]	I3D	✓	17.0
TA^{3}N[37]	I3D-TRN		10.0
TA^{3}N[37]	Resnet101-TRN		7.0
TA^{3}N[37]	I3D-TRN	✓	19.1
TA^{3}N[37]	Resnet101-TRN	✓	13.0
CO^{2}A[288]	I3D		16.4
CO^{2}A[288]	I3D	✓	20.1
MMKD (Ours)	I3D		19.4

Table 6.1: Comparison with Video Domain Adaptation Baselines. We report the experimental results on the Mixamo→Kinetics Dataset. The baselines with weak supervision assume that annotations are available for 5 randomly selected target samples per class.

Method	Modality	Backbone	Val Accu (%)
2s-AGCN	3D Skeleton	CNN	9.6
MS-G3D	3D Skeleton	CNN	17.0
CTR-GCN	3D Skeleton	CNN	16.7
STMT	3D Mesh	Transformer	21.3
MMKD (Ours)	RGB	I3D	19.4

Table 6.2: Comparison with Video Generalization Baselines. We report the experimental results on the Mixamo→Kinetics Dataset.

model with domain generalization models, which are trained with labeled source videos only and without access to any target videos. One promising direction is leveraging 3D representations, such as skeletons and meshes. Therefore, we compare our model with state-of-the-art 3D action recognition methods, including 2s-AGCN [264], MS-G3D [188], CTR-GCN [42] and *STMT* [351].

Setting	Pre-training Dataset	Backbone	Val Accu (%)
-	-	ViT-S	7.6
Unsupervised	Kinetics	ViT-S	15.8
Supervised	SomethingSomethingV2	ViT-S	18.4
Supervised	UCF-101	ViT-S	9.1
Supervised	ImageNet	I3D	19.4

Table 6.3: Ablations of Different Pre-training Settings.

6.4.3 Quantitative Results

Comparison with Video Domain Adaptation Baselines. We compare our model with video domain adaptation baselines and report the experimental results in Table 6.1. We observe that our model significantly outperforms all the video domain adaptation baselines under the same setting (without access to the labeled target videos). Besides, our model also outperforms some of the baselines which are trained with extra labeled target videos.

Comparison with Video Domain Generalization Baselines. We compare our model with domain generalization baselines and report the experimental results in Table 6.2. We observe that while our model does not require the 3D body pose estimation step during inference, it still achieves comparable performance to the best 3D action recognition model (*STMT*), which is 19.4 v.s. 21.3. This suggests that the light-weighted student model (I3D) can learn meaningful feature representations from the stronger teacher model.

Ablation Studies of Different Pre-training Settings. We explore different pre-training settings and report the results in Table 6.3. We observe that the I3D model pre-trained with ImageNet [62] achieves the best performance. Besides, the ViT-based [67] methods are sensitive to different pre-training settings, and transformer-based methods need a large amount of data to make the model converge.

6.5 Conclusion

We propose a multimodal knowledge distillation model for domain-adaptive action recognition. The model distills the strong 3D mesh-based action recognition model into the light-weighted I3D model. We empirically validate our model on the Mixamo→Kinetics dataset, which consists of a variety of real and synthetic videos of actions. We show that our model significantly outperforms the state-of-the-art video domain adaptation models. In the future, more work can be done to improve pose estimation quality and explore implicit 3D representations.

Part III

Generatively Pretrained Foundation Models for Open-Vocabulary Perception

In the last part, we leverage generatively pre-trained vision-language models and develop systems that can handle novel vocabularies and text prompts. We propose to evaluate the model on the open-vocabulary 3D scene understanding tasks including 3D semantic segmentation and visual grounding.

Chapter 7

Text-to-image diffusion models for open-vocabulary 3D scene understanding

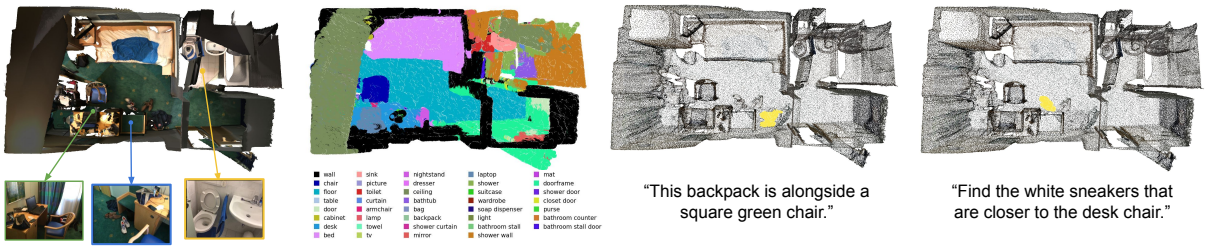

Figure 7.1: **Illustration of open-vocabulary 3D semantic scene understanding**. We propose *Diff2Scene*, a 3D model that performs open-vocabulary semantic segmentation and visual grounding tasks given novel text prompts, without relying on any annotated 3D data. By leveraging discriminative-based and generative-based 2D foundation models, *Diff2Scene* can handle a wide variety of novel text queries for both common and rare classes, like "desk" and "soap dispenser". It can also handle compositional queries, such as "find the white sneakers that are closer to the desk chair."

7.1 Overview

3D semantic scene understanding, with the task of assigning semantics to every 3D point, plays a fundamental role in many computer vision applications, such as robotics [321], autonomous

driving [119], human-computer interaction [76], and augmented reality [97]. Traditional studies in this field usually target solving this problem in a closed-set fashion [50, 253], resulting in models that can only be used to make predictions within the predefined label space.

Recent progress in computer vision have witnessed the emerging interests in solving semantic understanding tasks in open-vocabulary settings [115, 223, 239, 277]. In contrast to closed-set setting, models targeting open-vocabulary tasks must make predictions for any semantics described in text, including object category and fine-grained attributes (*e.g.*, shape, color, material, property) as well as their complicated compositions. However, this is a challenging task due to the wide diversity and complexity of possible queries. Motivated by the advance of aligning text and image embeddings with large-scale foundation models [11, 72, 123, 152], existing methods mitigate this challenge by lifting the image features from foundation models such as CLIP [72] or their descendants [87, 149] to 3D. These lifted feature representations for 3D points can then be used to query with open-vocabulary descriptions, achieving semantic understanding in 3D. Despite these achievements, contrastively trained CLIP-based models exhibit limitations in handling fine-grained classes [72] and novel compositional text queries [193], restricting their performance in open-vocabulary 3D semantic understanding.

The recently developed text-to-image diffusion models have shown outstanding abilities for image generation even with challenging text prompts [183, 241, 252], such as combinational descriptions with multiple attributes (*e.g., A bucket bag made of blue suede with intricate golden paisley patterns.*) The internal visual representation of these models, entangled with text embedding through cross-attention, have proven correlate well with semantic concepts described by language [145, 207, 209, 325]. On the other hand, CLIP-based foundation models have been shown to struggle with compositionality [193]. Moreover, compared with the CLIP model which is optimized for global representation, diffusion models have proven to be superior at local representation [279], which is a key for dense prediction tasks. Specifically, ODISE [318] applied the internal representations of Stable Diffusion [241] to open-vocabulary 2D semantic understanding tasks and achieved promising results.

One of the key challenges in 3D perception is the severe scarcity of point clouds and their dense labels. Several existing methods have been proposed to solve the lack of data issue in a zero-shot fashion by leveraging the CLIP model pre-trained on large-scale text-image data [121, 223, 277]. The prior art [223] extracts dense CLIP features from 2D images and distill the knowledge of their lifted 3D counterpart into a 3D mask predictor. However, CLIP features, as discussed above, struggle to handle fine-grained classes [72] and show worse localization capability compared with diffusion features. We leverage diffusion model as feature backbone

along with a mask-based segmentation head (*e.g.*, Mask2Fromer [44]) for its intrinsic nature that decouples mask and its semantic representations. This is intuitively suitable for leveraging semantically-rich embeddings from 2D foundation models and further learning geometrically-accurate masks from the 3D branch. However, performing multi-modal distillation with mask-based segmentation head is a non-trivial task. The frozen features extracted from the decoder of the U-Net in the diffusion model are trained with generative objectives, and cannot be directly used for the perception task. Therefore, directly distilling knowledge from these features as normally done in prior art [144, 187, 223] is infeasible. Another intuitive way is to leverage a supervised 3D mask proposal network and pool the feature representations from 2D CLIP features for each mask [277]. However, the training of 3D mask proposal network requires labeled 3D data, which may not be feasible in practice.

To mitigate these issues, we propose a novel mask distillation method tailored to distill knowledge from the Mask2Former style 2D branch [44, 318] to the 3D branch, which is shown in Fig. 7.2. Specifically, we design our 3D branch to take a 3D point cloud as input and to predict their 3D features. The semantically meaningful mask embeddings produced from our 2D branch are used as linear classifiers to assign class probability to these 3D features. Their corresponding 2D masks are lifted to 3D based on pixel-point correspondence and used to force the consistency learning of the 2D and 3D branch.

We evaluate *Diff2Scene* quantitatively on ScanNet [53], ScanNet200 [245], Matterport 3D [30] and Replica [272] for open-vocabulary 3D semantic segmentation and qualitatively on Nr3D [6] for visual grounding tasks. Our experimental results show that *Diff2Scene* outperforms state-of-the-art models [223] on all the four semantic segmentation datasets and achieves promising results on visual grounding tasks. In summary, we make the following contributions:

- To the best of our knowledge, we are the first to leverage text-image diffusion to perform open-vocabulary 3D semantic segmentation.
- We propose a novel mask distillation method to train a 3D mask prediction model by distilling knowledge from the Mask2Former style 2D segmentation model.
- The proposed method achieves state-of-the-art performance on several open-vocabulary 3D semantic segmentation and visual grounding benchmarks.

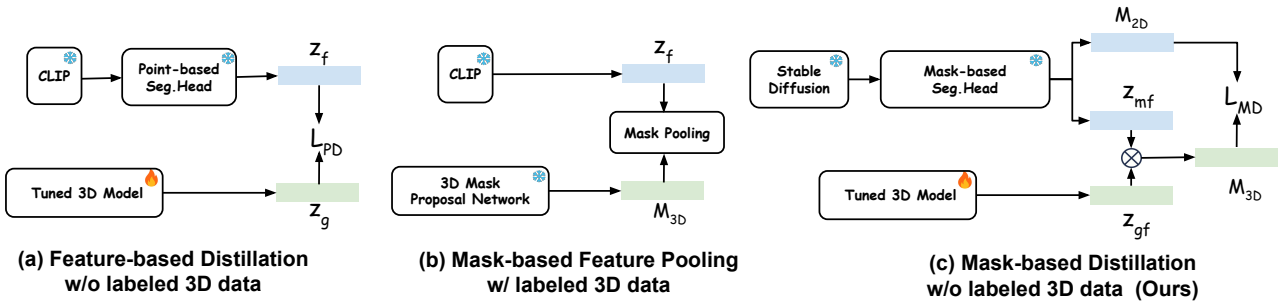

Figure 7.2: **Illustration of open-vocabulary 3D perception methods.** L_{PD} and L_{MD} denote point-based distillation loss and mask-based distillation loss. M_{3D} denote a set of predicted 3D masks; M_{2D} and Z_{mf} denote a set of predicted 2D masks and their semantic embeddings; Z_{gf} denote the high-resolution 3D feature map. (a) Directly minimizing the per-point feature distance between the CLIP-based model and the tuned 3D model [223]. (b) Directly using a 3D mask proposal network trained on labeled 3D data to produce class-agnostic masks, and then pool corresponding representations from the CLIP feature map [277]. (c) The proposed mask distillation approach, namely *Diff2Scene*, that uses Stable Diffusion and performs mask-based distillation. *Diff2Scene* leverages the semantically-rich mask embeddings from 2D foundation models and geometrically accurate masks from the tuned 3D model, and thus achieves superior performance compared to previous methods.

7.2 Related Work

Closed-vocabulary 3D semantic segmentation. In 3D semantic segmentation, a semantic category is assigned for each 3D point. It has been long studied [12, 13, 14, 69, 70, 89, 111, 112, 139, 147, 155, 192, 232, 233, 281, 306, 348] due to its importance in computer vision and robotics applications. One challenge of this task is that 3D point clouds are not in a regular structured format; network architectures that work well for 2D tasks cannot handle 3D point clouds effectively. As a result, most of the early studies focus on designing effective and efficient network architectures that are suitable for 3D point clouds [50, 69, 70, 89, 111, 112, 147, 232, 233, 282]. This line of work achieved great success and significantly improves the results of 3D semantic segmentation. Another challenge is the lack of large scale data with ground truth annotations. Due to the intensive labeling effort and high cost of data annotation [245], the available datasets for 3D semantic segmentation are usually small in scale. In the absence of

large scale data, early studies usually target solving this problem in a closed-vocabulary setting, where the trained model can only predict categories that appear during training. To mitigate the scale limitation of existing datasets, a handful of works [45, 46, 47, 48, 174, 207, 334] have applied zero-shot learning in 3D scene understanding tasks. [45, 46, 47, 48, 334] focused on 3D point classification task and [174, 207] tried to address the 3D semantic segmentation problem. However, these zero-short methods still require ground truth annotations for a certain amount of 3D point clouds.

Open-vocabulary 3D segmentation. The recent progress of large-scale vision and language representation learning [11, 72, 123, 152] has advanced the study of semantic and instance segmentation in an open-vocabulary setting. [87, 149, 163] first explored open-vocabulary 2D semantic segmentation. They proposed aligning per-pixel features [149] or features from mask regions [87, 163] with the corresponding text embedding. Following these works, [32, 64, 82, 113, 121, 200, 223, 257, 258] focus on 3D semantic segmentation in an open-vocabulary setting. Among them, [32, 82, 113, 121, 200, 248, 257] project 3D points to 2D images and solve the 3D problem in the 2D space, instead of targeting the 3D open-vocabulary semantic segmentation directly. As [64] pointed out, the projection from 3D to 2D has information loss and the solution is suboptimal.

To make better use of information from the 3D point cloud, [64, 223] proposed to directly applying semantic segmentation on the 3D point cloud. [64] and its extension [123] proposed associating captions generated for 2D images to corresponding 3D point clouds to build the pseudo-ground truth captions for 3D point clouds. A neural network is trained to associate the 3D point cloud with these pseudo labels through contrastive loss. Similar to the zero-shot setting, [64] evaluated their model in a leave-one-out fashion, which still requires annotations for 3D point cloud. Inspired by the strong open-vocabulary ability of large-scale vision and language models, Peng *et al.* [223] proposed distilling knowledge to a 3D point cloud model. They trained 3D semantic segmentation model by only distilling the knowledge from a CLIP-style [72] 2D open-vocabulary semantic segmentation model [87, 149]. They demonstrated that without training with any ground truth labels, the model can achieve great performance on many open-vocabulary tasks. However, we observe that [223] is strongly limited by the 2D open-vocabulary semantic segmentation models used as the teacher. Its performance on rare classes that are not used in training these models are not satisfactory. Our method follows this idea by distilling the knowledge of 2D open-vocabulary semantic segmentation model to a 3D model. In constrast to the approach in [223], we use a diffusion-based 2D open vocabulary semantic segmentation model [318] as the teacher model.

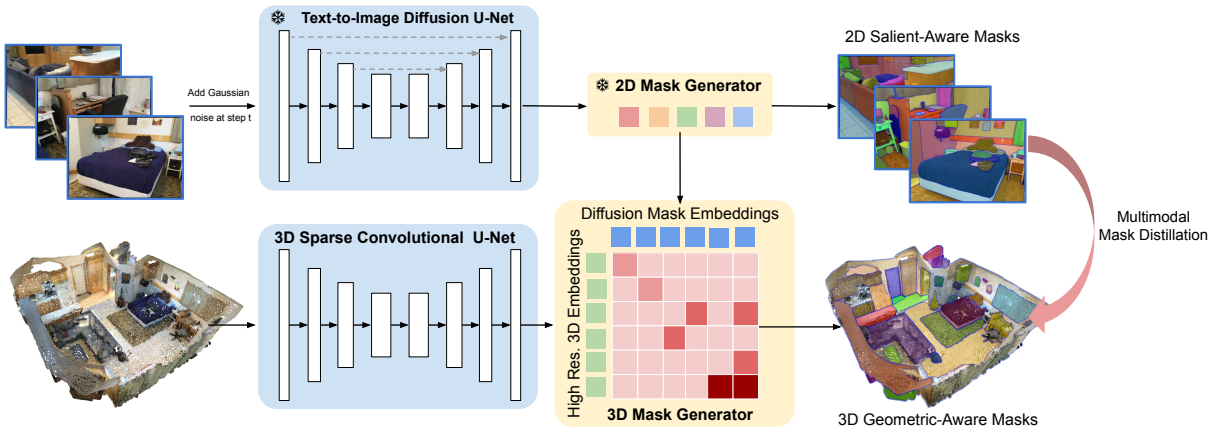

Figure 7.3: **Overview of our method.** We propose *Diff2Scene*, an open-vocabulary 3D semantic understanding model. *Diff2Scene* contains two branches. The 2D branch is designed to be a diffusion-based 2D semantic segmentation model. It accepts a 2D image as input and predicts a set of 2D probabilistic masks with corresponding semantically-rich mask embeddings. The 3D branch utilizes the point cloud and 2D mask embeddings as input. The 2D mask embeddings are used as "semantic queries" to generate corresponding 3D probabilistic masks. The model learns salient patterns from the RGB images and geometric information from the point clouds.

Diffusion models for scene understanding. The last few years have witnessed the success of diffusion models in image generation [241, 248]. Recent studies also observed the diffusion models are strong representation learners [145, 207, 209, 325]. As a result, researchers have applied it to many understanding tasks such as image classification [148], object detection [40], image semantic segmentation [17, 122, 318], instance segmentation [162], human pose estimation [79, 104, 260], action segmentation [176], camera pose estimation [300], to name a few, and achieved great success. Especially, [318] and [162] showed that Stable Diffusion [241], whose internal representation being well correlated with text embedding, has strong open-vocabulary abilities for understanding tasks. Inspired by this, we are the first to apply text-to-image diffusion models to open vocabulary 3D semantic segmentation task.

7.3 Method

We introduce *Diff2Scene*, an open-vocabulary 3D semantic understanding method. Similar to [223], our proposed model operates in a zero-shot fashion, where no ground truth 3D annotations are needed during training.

7.3.1 Overview

An overview of *Diff2Scene* is shown in Fig. 7.3. It takes posed RGB images and the reconstructed 3D point cloud as model inputs. The model predicts the semantic label for each 3D point. *Diff2Scene* has two branches. The 2D branch is designed to be an open-vocabulary 2D semantic segmentation model. It leverages text-to-image generative model [241] which is pre-trained on massive text-image pairs. The model takes a 2D image as input to predict a set of 2D probabilistic masks with their corresponding 2D mask embeddings. Thanks to the generative pre-training process with large-scale text-image pairs, the 2D mask embeddings are semantically rich. The model leverages the salient patterns in RGB images to produce the 2D salient masks. The 3D branch takes the point cloud and the 2D mask embeddings as inputs. The 2D mask embeddings are used as linear classifiers to assign class probabilities to each of 3D features output from the 3D branch, resulting in a 3D probabilistic mask termed as geometric masks. To predict the per-point semantic class, the model first computes the per-mask category logits for both salient and geometric masks. Then we ensemble the per-mask logits for those two types of masks. In the way, the model can learn salient patterns from the RGB images and geometric information from the point clouds.

7.3.2 2D Semantic Understanding Model

One challenge of 3D semantic understanding is the severe scarcity of 3D point clouds with groundtruth labels. To tackle the challenge brought by limited training data, vision-language foundation models have been used to transfer semantically-rich 2D features into the 3D space [121, 223, 277]. [277] used on a model trained on labeled 3D data to produce class-agnostic masks, and then pooled the corresponding 2D representations as the mask embeddings. On the other hand, [223] proposed to leverage a pre-trained 2D semantic segmentation model as feature extractor to perform open-vocabulary 3D segmentation, and no ground truth 3D annotations are needed during training. In this work, we follow the setting in [223] to reduce the 3D annotation efforts.

The 2D segmentation model consists of an image backbone ϕ which is a foundation model pretrained on large-scale text-image pairs; and a segmentation head σ to predict the semantic embedding. There are multiple design options for the 2D backbone ϕ and segmentation head σ. (1) The 2D backbone could either be **contrastively pretrained** or **generatively pre-trianed**. The popular frameworks for contrastive representation learning include CLIP [72] and ALIGN

[123]. On the other hand, a few works [302, 318, 340] have demonstrated promising performance by using generatively pre-trained representations for perception task. The feature representations from Diffusion U-Net blocks are extracted for different downstream tasks.

Once feature representations from text-image foundation models are extracted, a segmentation head is added upon those features to predict the per-point semantic classes. The segmentation problem could be formulated as **pixel-based classification** or **mask-based classification**. For pixel-based classification [87, 149], the intermediate output of segmentation head is of shape $H \times W \times C$, where H and W is image height and width, and C is the dimension of feature embedding. For mask-based classification [43, 44, 318], the segmentation head takes the 2D feature map $\mathbf{F}^{2d}$ and N fixed mask queries $\{q_i\}_{i=1}^{N}$ as input. The intermediate output is N 2D probabilistic masks $\{\mathcal{B}_i^{2d}\}_{i=1}^{N}$ and their corresponding mask embeddings $\{f_i^{2d}\}_{i=1}^{N}$.

In this work, we choose diffusion model as the feature backbone ϕ, considering its strong localization ability brought by generative pre-training. Besides, we leverage mask-based segmentation head for its intrinsic nature that decouples mask and its semantic representations. This is intuitively suitable for leveraging semantically-rich embeddings from 2D foundation models, and further learn geometrically-accurate masks from the 3D branch.

7.3.3 Geometry-Aware 3D Mask Model

While mask-based segmentation has achieved promising performance in fully-supervised setting [43, 44, 253], it has been rarely explored to transfer the learned mask-level representations into another domain. On the other hand, the point-based feature representations from 2D foundation model can be naively distilled by minimizing the per-point feature distance. For example, [223] proposed to train a 3D model to predicts 3D semantic meaningful features by distilling pixel aligned 2D features. However, similar methods are not applicable in our proposed method. First of all, our 2D semantic understanding model uses a mask-based segmentation head which does not provide semantically-rich features in the pixel level. Secondly, the backbone of our 2D semantic understanding model is a frozen stable diffusion model [241] which is designed to generate realistic images with rich details and not tuned for semantic segmentation tasks. The per-pixel features extracted from it are not feasible to supervise the training of our 3D mask model†.[1] In the following, we introduce our proposed mask distillation which is tailored to distill knowledge from the mask-based 2D foundation model to the geometry-aware 3D mask model.

[1] †The 3D mask model trained to distill these features does not converge.

The mask-based 2D foundation model predicts N 2D probabilistic masks $\{\mathcal{B}_i^{2d}\}_{i=1}^N$ and their corresponding mask embeddings $\{f_i^{2d}\}_{i=1}^N$. Specifically, $\mathcal{B}_i^{2d}$ represents a probabilistic map whose elements represent the probability of the corresponding pixel being foreground. We first compute the pixel-point correspondence following [223]. Subsequently, a set of 3D probabilistic masks $\{\mathcal{B}_i^{3d}\}_{i=1}^N$ can be generated by lifting the 2D masks $\{\mathcal{B}_i^{2d}\}_{i=1}^N$ to 3D space based on the pixel-point correspondence. We proposed a novel mask distillation which distills information from both 3D probabilistic masks $\{\mathcal{B}_i^{3d}\}_{i=1}^N$ and the corresponding semantic rich mask embeddings $\{f_i^{2d}\}_{i=1}^N$ generated from the 2D branch. Specifically, we train a Minkowski network [50] as the 3D mask prediction model to generate geometry-aware 3D masks. The 3D point cloud is quantized into voxels by averaging the pixels within each voxel to save memory and reduce computes. The 3D mask prediction model generates a 3D feature to represent each voxel and this feature is assigned to all points within the voxel. This produces a full feature map $\mathbf{F}^{3d} \in \mathbb{R}^{M\times D}$ for the point cloud, where D is the dimension of the 3D feature. The semantic rich 2D mask embeddings $\{f_i^{2d}\}_{i=1}^N$ are used as linear classifiers to compute the logits $\mathcal{S}_i \in \mathbb{R}^M$ of a 3D feature belonging to the corresponding class:

$$\mathcal{S}_i = \langle \mathbf{F}^{3d}, f_i^{2d} \rangle, \tag{7.1}$$

where $\langle \cdot, \cdot \rangle$ denotes inner product. The 3D probabilistic mask $\mathcal{B}'^{3d}_i$ is then generated by applying the sigmoid function on $\mathcal{S}_i$. We propose a multimodal mask distillation loss to train our 3D mask generator:

$$\mathcal{L} = \sum_{i=1}^{N} 1 - \cos(\mathcal{B}'^{3d}_i, \mathcal{B}_i^{3d}). \tag{7.2}$$

The distillation loss aims at forcing the 2D and 3D branch to make consistent predictions. It serves as an implicit distillation objective to make the 3D model learn high-resolution, semantically-rich feature representations.

7.3.4 Open-Vocabulary Inference

During inference, *Diff2Scene* takes a 3D point cloud and its multiview 2D images as inputs. The 2D semantic understanding model consumes the 2D images and generates a set of 2D probabilistic masks $\{\mathcal{B}_i^{2d}\}_{i=1}^N$ with their corresponding mask embeddings $\{f_i^{2\mathrm{d}}\}_{i=1}^N$, where $f_i^{2d} \in \mathbb{R}^D$. The 3D mask model takes the 3D point cloud and the mask embeddings $\{f_i^{2\mathrm{d}}\}_{i=1}^N$ as inputs to predict the 3D probabilistic mask $\{\mathcal{B}'^{3d}_i\}_{i=1}^N$. To ground a semantic label c to the 3D point cloud, we first apply the same idea from [318] to compute the geometric mean (denoted as p_i^c) of

label probabilities from diffusion and discriminative models for each 2D mask $\{\mathcal{B}_i^{2d}\}_{i=1}^{N}$. Next, the label probabilities $\mathbf{p}^c$ are assigned to 3D points via the following equation:

$$\mathbf{p}^c = \lambda \sum_{i=1}^{N} p_i^c * \mathcal{B}_i^{3d} + (1-\lambda) \sum_{i=1}^{N} p_i^c * \mathcal{B}'^{3d}_i, \tag{7.3}$$

where $\lambda = 0.5$. When multiple labels can be assigned to a 3D point, the label with the highest probability from Eq. 7.3 is taken.

7.4 Experiment

We conduct a series of experiments to demonstrate the effectiveness of *Diff2Scene* on a variety of zero-shot 3D scene understanding benchmarks. We first evaluate the proposed model on zero-shot open-vocabulary semantic segmentation tasks following the evaluation protocol of [223]. We then perform comprehensive ablation studies to validate our designs. Finally, we qualitatively demonstrate the strong ability of the proposed model for open-vocabulary 3D segmentation and grounding complicated compositional text queries.

7.4.1 Datasets

We use ScanNet [53], Matterport3D [30], ScanNet200 [245] and Replica [272] for the open-vocabulary 3D semantic segmentation task. We provide qualitative analysis of the visual grounding task on Nr3D [6]. Except for Replica, point clouds and multi-view images in the training split without ground truth annotations are used for model training. As Replica does not provide the training data, we perform training on ScanNet and perform evaluation on Replica, following the setting in [277].

ScanNet is one of the largest 3D semantic segmentation dataset. It provides 80,554 images from 1201 scans for training and 21,300 images from 312 scans for testing with 20 semantic labels.

Matterport3D is a large scale RGB-D dataset containing 10,800 panoramic views from 194,000 RGB-D images of 90 building-scale scenes. It splits 61 scenes for training, 11 scenes for validation and 18 for testing. We train our 3D branch using the images in the training splits and report the results on test split.

ScanNet200 has 200 semantic labels with long-tailed classes. It also provides a grouping of the 200 categories based on the number of labeled surface points in the training set, resulting in 3 subsets: head, common, and tail. This enables us to evaluate the performance of our method on the long-tail distribution, making ScanNet200 a natural choice as an evaluation dataset. We report the mean intersection over union (mIoU) metric on the validation set consisting of 312 scenes following the split in [223, 245, 277].

Replica contains 51 categories, and we further split those categories into head and tail sets based on their appearance frequency. We report the mIoU on the *office0*, *office1*, *office2*, *office3*, *office4*, *room0*, *room1*, and *room2*.

Nr3D is a 3D visual grounding dataset which contains diverse text prompts. To further evaluate the ability of our model to distinguish between objects in the same class but with different attributes, we perform qualitative evaluation on the visual grounding dataset Nr3D [6]. We perform zero-shot evaluation on the validation set without training on any labeled data for the visual grounding task.

7.4.2 Baseline Methods

We compare *Diff2Scene* with the current state-of-the-art fully-supervised 3D semantic segmentation models including TangentConv [280], TextureNet [114], SFSS-MMSI [50], CSC-Pretrain [105], SupCon [343], LGround [245] and MinkowskiNet [50] on the 3D semantic segmentation benchmark. We also compare our model against OpenScene [223] and ConceptFusion [121], the recently proposed open-vocabulary 3D semantic understanding model. For OpenScene [223], we compare with its OpenSeg [87] variant which has the same feature and pre-trained datasets for a fair comparison. We also compare our model with its three different variants (2D Fusion, 3D Distill, and 2D/3D Ensemble). Besides, we adapt the state-of-the-art 3D instance segmentation model OpenMask3D [277] for comparison on the 3D semantic segmentation benchmark.

7.4.3 Implementation Details

We use posed multi-view RGB images and 3D point clouds for all the datasets. ODISE [318], which consists of a diffusion backbone and mask-based segmentation head, is used as the model in our 2D branch. It uses a stable diffusion model [241] pre-trained on Laion-5B [252] as the

Table 7.1: Comparison to state-of-the-art models. We report mIoU for all benchmarks. Best results in zero-shot, open-vocabulary setting are shown in bold.

	ScanNet	**Matterport3D**	**ScanNet200**				**Replica**		
	All	All	Head	Common	Tail	All	Head	Tail	All
Fully-supervised									
TangentConv [280]	40.9	-	-	-	-	-	-	-	-
TextureNet [114]	54.8	-	-	-	-	-	-	-	-
SFSS-MMSI [50]	-	35.9	-	-	-	-	-	-	-
CSC-Pretrain [105]	-	-	45.5	17.1	7.9	24.9	-	-	-
SupCon [343]	-	-	48.6	19.2	10.3	26.0	-	-	-
LGround [245]	-	-	48.5	18.4	10.6	27.2	-	-	-
MinkowskiNet [50]	69.0	54.2	46.3	15.4	10.2	25.3	-	-	-
Zero-shot, open-vocabulary									
MSeg Voting [146]	31.0	33.4	-	-	-	-	-	-	-
ConceptFusion [121]	33.3	-	17.5	6.3	2.8	8.8	11.6	3.5	4.6
OpenMask3D [277]	34.0	-	19.6	7.5	4.5	10.5	13.2	3.4	4.8
OpenScene (2D) [223]	41.4	32.4	21.9	10.8	5.5	12.7	33.4	11.5	14.5
OpenScene (3D) [223]	46.0	41.3	17.6	0.0	0.0	6.3	32.6	7.7	11.1
OpenScene (2D/3D) [223]	47.5	42.6	20.0	9.7	5.1	11.6	34.2	11.9	14.9
Diff2Scene (Ours)	**48.6**	**45.5**	**25.6**	**11.5**	**6.9**	**14.2**	**46.2**	**12.9**	**17.5**

feature backbone. The dimensions for diffusion and CLIP features are 256 and 768 respectively. The number of queries of Mask2Former [44] is 100. Similar to OpenScene [223], we use MinkowskiNet18A [50] as the model in our 3D branch to extract 3D features from the 3D point clouds. Our 3D model is trained for 200 epochs with a batch size of 8. Adam optimizer [135] is used with a learning rate of 0.0001 and polynomial learning rate policy is used as the learning rate scheduler with power 0.9. During inference, text-embeddings are computed by the ViT-L/14 CLIP model [72] for each of the semantic categories and grounding queries. We use the same pre-processing step and pre-trained dataset as OpenScene [223] (OpenSeg [87]) for a fair comparison.

7.4.4 Quantitative Results

Evaluation on zero-shot 3D semantic segmentation. We first compare our method with the state-of-the-art open-vocabulary scene understanding models and fully-supervised 3D seg-

Table 7.2: **Effectiveness of Different Distillation Settings.** We report mIoU of different methods on the Replica [272] dataset.

Setting	Distillation Type	Head	Tail	All
fine-tuned CLIP feature [223]	Point-based	32.6	7.7	11.1
frozen diffusion feature	Point-based	Divergence		
multimodal mask distillation (ours)	Mask-based	**43.3**	**8.0**	**12.8**

mentation models. We report the mIoU for Scannet, Matterport3D, Scannet200, and Replica in Table 7.1. We find that our method achieves better results than the state-of-the-art open-vocabulary models and their variants on all the benchmarks. Besides, although our zero-shot model has noticeable performance drop compared with fully-supervised model, the gap of tail categories between the proposed method and those methods are relatively small (*e.g.*6.9 7.9 from CSC-Pretrain) on Scannet200. This demonstrates the strong potential of the proposed method for long-trailed 3D semantic segmentation tasks.

Generalization to unseen dataset. To test the generalization ability of our proposed model, we evaluate it on an unseen dataset Replica [272] and report the results in Table 7.1. The results shown that our proposed method significantly outperforms the state-of-the-art models on head, tail and all categories in Replica. This demonstrates the strong generalization ability of the proposed model on novel datasets.

Effectiveness of Different Distillation Settings. We compare our mask-based distillation method with point-based ones under different settings and report the performance of the 3D branch on Replica [272] in Table 7.2. The supervisions for point-based method include: (1) Fine-tuned CLIP feature, which follows the same setting as OpenScene [223]; (2) Frozen diffusion feature extracted from the last layer of diffusion U-Net block. We observe that distilling frozen diffusion features does not converge. Our proposed method, by introducing the semantic meaningful mask embedding output from the 2D branch as a fixed classifier, significantly boost the performance of the 3D branch.

Ablation studies. We conduct ablation studies using the Replica dataset [272] and show the results in Table 7.3. We first analyze the effectiveness of combining 2D and 3D masks using equation 7.3. We observe that compared with using salient or geometric mask only, using both types of masks achieves the best performance. This is intuitive as both salient patterns and geometric information are helpful to segment accurate class boundaries. We then analyze the effectiveness of different semantic features. We find that discriminative and diffusion features

Table 7.3: **Performance of different model ablations.** We observe that each component of our model gains consistent improvements.

Method	mIoU
Our full model	17.5
Without 2D (salient) mask	12.8
Without 3D (geometric) mask	16.5
Without discriminative (CLIP) features	15.5
Without generative (Stable Diffusion) features	15.3

serve as strong complementary to each other. We also observe that using those two types of semantic features jointly can significantly outperform using any of them alone.

7.4.5 Qualitative analysis

Visualizations of zero-shot semantic segmentation. In Fig. 7.4, we provide qualitative analysis of our approach and OpenScene for the zero-shot 3D semantic segmentation task. Compared with OpenScene, our model generates coherent and consistent masks (*e.g.*, the table mask in first column and the bed mask in third column) thanks to the mask-instance representations. It predicts accurate semantic labels for both head and tail categories by leveraging both CLIP and diffusion features.

Visualizations of visual grounding results. We provide qualitative analysis of our approach and OpenScene for the zero-shot visual grounding task in Fig. 7.5. We observe that our model can accurately identify objects given complicated text queries. It demonstrates that the proposed method, *Diff2Scene*, has good capability at the following types of queries. Fig. 7.5 (a) describes object shape and color, and even in comparative degree (*It's the shorter, red box*); Fig. 7.5 (b) describes a rare object (*rack*) and its surrounded object with surface appearance descriptions (*wrinkled towel*); Fig. 7.5 (c) describes the relative location of the object (*next to the desk*); Fig. 7.5 (d) describes the usage of the object (*recycling*). In addition, we can see that given vague usage descriptions without common category names like *trash bin* in the text prompts, the model can still accurately identify the object.

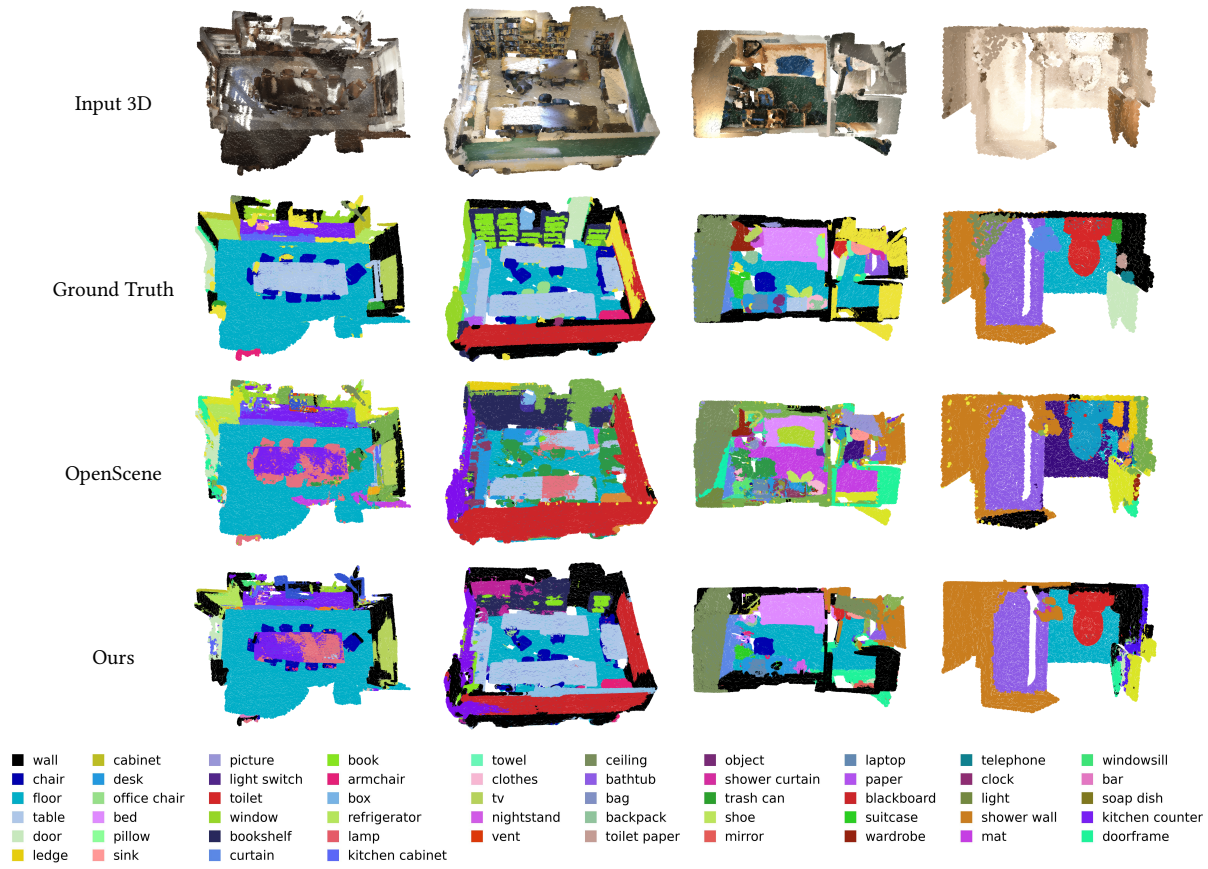

Figure 7.4: **Qualitative results from our model and OpenScene on zero-shot semantic segmentation.** We visualize the segmentation results on the validation set of ScanNet200 [245]. We observe that our model can predict coherent masks with accurate semantic labels compared to OpenScene for both head and tail categories.

7.5 Conclusion

In this chapter, we investigate the problem of leveraging frozen representations from large text-to-image diffusion models for open-vocabulary 3D semantic understanding. *Diff2Scene* sets a new state-of-the-art in the zero-shot 3D semantic segmentation task and shows promising performance in the visual grounding task. Our method also shows outstanding generalization ability towards unseen datasets and novel text queries. It provides a new way to effectively leverage generative text-to-image foundation models for 3D semantic scene understanding tasks.

There are several limitations of the proposed model. First, while our model achieves better performance compared to existing methods in small objects, it still misclassified some small and rare categories (*e.g.*rail). Second, we observe that the model can be easily confused by

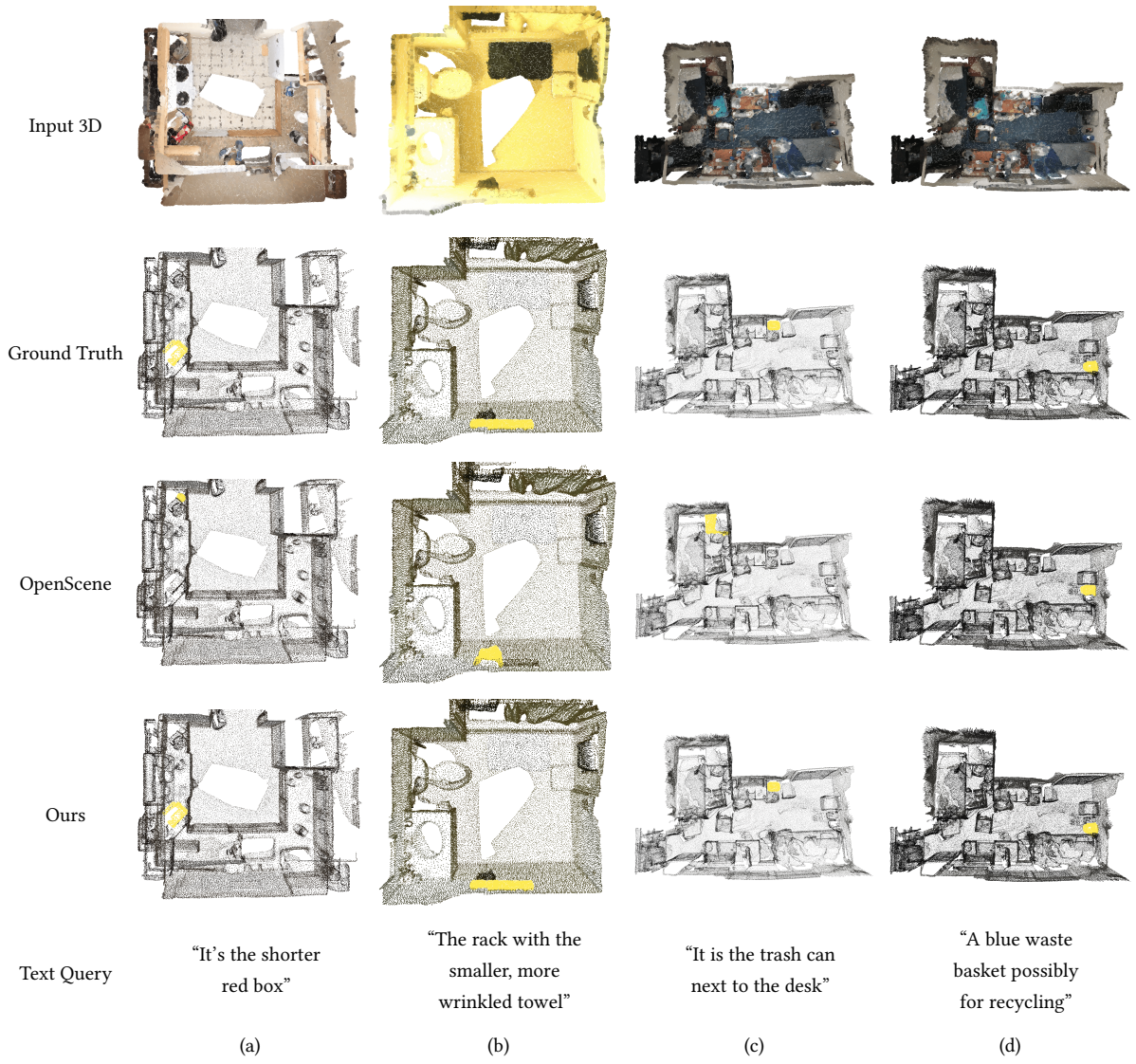

Figure 7.5: **Qualitative results from our model and OpenScene on zero-shot visual grounding.** Our open-vocabulary semantic understanding model is capable of handling different types of novel and compositional queries. Novel object classes as well as objects described by colors, shapes, appearances, locations, and usages are successfully retrieved by our method. Note that the located points are colored in yellow.

fine-grained categories that with similar semantic meaning. For example, the model sometimes wrongly classifies points of windowsill to the window class. In future work, it will be interesting to design models that can accurately distinguish between fine-grained categories in the open-vocabulary setting.

Part IV

Conclusions and Future Directions

Chapter 8

Conclusions

In this thesis, we explore different representation learning methods based on Siamese learning, masked visual modeling, and generatively pre-training to develop systems that can generalize to novel viewpoints, scenes and vocabularies. This thesis consists of three parts. The first part conducts robust semantic instance segmentation for videos and 3D data. We aim to further learn feature representations that are invariant to various viewpoints and noise conditions via Siamese learning. We propose to leverage temporal consistency for videos and spatial consistency for 3D volumetric images, such that the learned feature representations have strong generalization ability. In the second part, we tackle the problem of human action analysis, which requires the model to learn from dynamic cues. We propose representation learning techniques based on masked visual modeling, such that the model can learn better spatial-temporal context. We also exploit both RGB videos and 3D human meshes for robust multi-modal action analysis. Finally, in the third part, we leverage generatively pre-trained vision-language models and develop systems that can handle novel vocabularies and text prompts. Our final goal is to build a robust system that can generalize to novel viewpoints, scenes, and vocabularies. Below, we first summarize the contributions of each part of our work, and then discuss the limitations of the thesis, as well as the short term goals. We provide key insights and long-term future goals in the final part based on all the works we have done.

8.1 Contributions

8.1.1 Part I: Robust Semantic Instance Segmentation

- We propose a novel method that exploits inter-frame consistency for robust instance segmentation from drone videos. From practical perspective, We are the first work that study how we could utilize social media drone videos for natural damage assessment. Our system was deployed by Federal Emergency Management Agency.
- We explore Siamese learning method for semantic segmentation from 3D volumetric images, which is also the first work for leveraging image-level class labels for weakly-supervised 3D segmentation.

8.1.2 Part II: Generalized Human Action Analysis

- We develop a novel masked visual modeling method for recognizing human actions from 3D meshes. This is also the first model that is able to encode temporal mesh sequences.
- We propose a novel generalized system that jointly takes RGB videos and estimated 3D meshes for human action analysis. It achieves the top performance among research projects funded by U.S. Army Research Lab for 2 years.

8.1.3 Part III: Open-Vocabulary Perception

- We are the first to leverage text-image diffusion to perform open-vocabulary 3D semantic segmentation.
- We propose a novel mask distillation method to train a 3D mask prediction model by distilling knowledge from the Mask2Former style 2D segmentation model.

8.2 Limitations

8.2.1 Limitations of High-Quality 3D Datasets

Based on our experiments on 3D datasets like Scannet [53] and Replica [272], we have noticed that those datasets are limited by the scene diversities. This could bring the following problems. First, the model evaluation doesn't comprehensively cover the real-world scenarios. This may

bring extra bias during model deployment. (2) It is not feasible to incorporate the 3D modality in the large-scale pre-training stage, especially when involving multiple modalities. Considering that the paired text-image data is exponentially greater than the number of 3D datasets, directly introducing 3D modality may only bring marginal improvement for 3D perception. Considering the cost of collecting high-quality 3D data, one of the potential solution is leveraging 3D reconstruction models for images and videos. In chapter 6, we find that it is helpful to incorporate estimated 3D meshes during the fine-tuning stage. It will be interesting to see if noisy 3D estimations are helpful for large-scale multimodal pre-training as well.

8.2.2 Limitations of Model Efficiency

In Part II, we have proposed transformer-based architectures for modeling videos and 3D meshes. However, the computational cost of transformers is expensive, particularly when handling high-dimensional mesh representations. Therefore, model distillation or quantization methods are essential for deploying our model on edge devices. In Part III, we apply Stable Diffusion, based on the U-Net architecture [243], for the open-vocabulary 3D perception task. While our model improves segmentation accuracy, we observe that the inference speed becomes a bottleneck. Specifically, inference with Stable Diffusion [241] is significantly slower compared to CLIP-based methods [72]. To meet real-time deployment requirements, one potential approach is to transition to transformer-based diffusion models and adopt optimization techniques such as Flash Attention [58] and KV Cache [229] to improve speed.

8.2.3 Limitations of Efficient 3D Representations Under Limited GPU Constrains

Due to the constrain of GPU memories, it is not feasible to use the entire 3D scene as model input. Therefore, in chapter 7, we split the input 3D scenes into several sub-scenes, and perform model predictions on each of the individual sub-scenes. However, this may lead to inaccurate predictions on the boundaries of the sub-scenes. Besides, for text prompts that require reasoning of the spatial locations, the model needs to jointly consider multiple sub-scenes. Therefore, how to represent the 3D scene in a memory-efficient way is an important problem.

8.2.4 Trade-Off Between Specialized Expert models and General-Purpose Foundation Models

In this thesis, we have explored specialized expert model in chapter 2 and chapter 3. In these chapters, we train a single-purpose segmentation model on domain-specific datasets. On the other hand, we also explored how to fine-tune a general-purpose foundation models for open-vocabulary segmentation task in chapter 7. There are several computation-accuracy trade-offs between specialized models and general-purpose foundations models. In this thesis, we do not provide trade-off analysis for expert models and foundation models in a unified benchmark due to constrains of computational resources. We believe it will be interesting to see the accuracy-cost analysis for expert models and different fine-tuning techniques for foundation models (*e.g.* parameter-efficient fine-tuning, LoRA adapters [110], etc). We believe a well-defined trade-off between computational efficiency and accuracy would be helpful for real-world deployment.

8.2.5 Generalization Ability Towards Novel Tasks

In this thesis, we have investigated the generalization ability of computer vision models towards novel viewpoints (Part I), scenes (Part II), and vocabularies (Part III). While our systems achieve substantial performance improvement compared to state-of-the-art models, it still fails behind considering the generalization ability of Large Language Models (LLMs). Specifically, LLMs demonstrate impressive performance in terms of generalization towards novel tasks. This is mainly due to the NLP tasks can be represented in a unified format (*i.e.* next token prediction). In the future, more work can be done to train different vision tasks (including both perception and generation) in a unified format, and investigate the model's generalization ability towards novel tasks.

8.3 Key Insights and Future Directions

8.3.1 Graphics Engines for Generative Models

In this thesis, we have investigated the use of graphics engines for visual perception tasks in Part II. Specifically, we have explored how 2D and 3D data generated from graphics engines can be utilized to learn domain-invariant feature representations. Future work could further explore the potential of graphics engines in generative models, which is also a key area in the computer

vision field. We present the results of our initial exploration below. Given an input image and a text description specifying the object movement and its destination, we first estimate the 3D scene from the image and use a graphics engine to render the movement trajectory of the object. A pre-trained video generative model is then employed to generate the video conditioned on the trajectory predicted by the physics engine. Our results shown in Figure 8.1 demonstrate that incorporating physics priors into video diffusion models improves text-video alignment, physical realism, and photorealism. Building on this approach, future research could utilize graphics engines to generate synthetic datasets, enabling a single diffusion model to learn directly from physics-realistic videos in an end-to-end manner. This would represent a significant advancement in the field of physics-aware video generation.

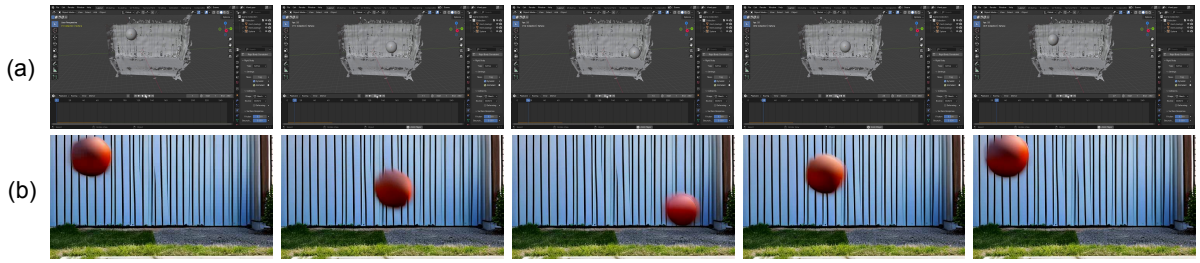

Figure 8.1: Visualizations of (a) physics simulation of ball bouncing from the graphics engine and (b) video generated from diffusion model which conditions on the object motion.

8.3.2 Language-Driven Graphics Engines

One of the fundamental differences between visual and text data lies in the complex compositions inherent in visual representations, which pose additional challenges for model learning. For example, videos consist of intricate combinations of textures, styles, object movements, and camera dynamics over time. The diversity of these inputs makes it difficult for models to consistently capture the underlying patterns needed for various downstream tasks. A potential solution is to leverage 3D graphics engines to render realistic videos by arranging 3D assets in accordance with physical laws. However, current graphics engines require substantial human effort to create even a single scene. Future research could focus on developing language-driven graphics engines, where users simply provide textual descriptions of the spatial relationships between key objects, and the engine autonomously arranges and renders the scene. Additionally, such systems could be extended to simulate dynamic scenarios, such as object and human movements within a 3D environment, ensuring natural and plausible motions. By doing so,

graphics engines could become a powerful tool for scaling up datasets used to train foundational models, especially as real-world datasets become increasingly scarce.

8.3.3 Self-Improved Computer Vision System and Embodied AI

In this thesis, we focus primarily on visual perception tasks, which serve as foundational components for various downstream applications, such as robotics. Looking ahead, it would be valuable to explore the system's generalization ability on real robots. Specifically, if it is able to generalize towards novel tasks specified by users. Additionally, we anticipate that the system could improve over time through a self-learning process. For example, if a user provides a new task, such as "fold the wheelchair," the system could employ a policy generator to control the robot, alongside a discriminator to assess the success of the task. The underlying idea is that while collecting real-world manipulation data for every possible task is impractical, it is feasible to use a discriminator—such as a multimodal large language model—that evaluates the final manipulation results (e.g., video) and determines whether the task was successfully completed. If the task is not deemed successful, the policy generator could then generate a new policy. Through this iterative process, the system could gradually develop the ability to generalize to new tasks via self-learning.

Bibliography

[1] https://github.com/xulabs/projects/tree/master/respond-cam/. 3.4.3

[2] https://github.com/wolny/pytorch-3dunet/. 3.4.3

[3] https://github.com/txin96/VoxResNet/. 3.4.3

[4] Temporal relational crosstransformers for few-shot action recognition. In *Computer Vision and Pattern Recognition*, 2021. 6.2

[5] Felix Achilles, Alexandru-Eugen Ichim, Huseyin Coskun, Federico Tombari, Soheyl Noachtar, and Nassir Navab. Patient mocap: Human pose estimation under blanket occlusion for hospital monitoring applications. In *International conference on medical image computing and computer-assisted intervention*, pages 491–499. Springer, 2016. 5.1, 5.2

[6] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, 2020. 7.1, 7.4.1, 7.4.1

[7] Eirikur Agustsson, Jasper R. R. Uijlings, and Vittorio Ferrari. Interactive full image segmentation by considering all regions jointly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2019. 3.4.2

[8] Sheharyar Ahmad, Kashif Ahmad, Nasir Ahmad, and Nicola Conci. Convolutional neural networks for disaster images retrieval. In *MediaEval*, 2017. 2.2

[9] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018. 3.2, 3.3.3

[10] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019. 3.2, 3.3.3, 3.3.4, 3.3.4

[11] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 7.1, 7.2

[12] Abhishek Anand, Hema Swetha Koppula, Thorsten Joachims, and Ashutosh Saxena. Contextually guided semantic labeling and search for 3d point clouds. In *IJRR*, 2011. 7.2

[13] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, 2016. 7.2

[14] Matan Atzmon, Haggai Maron, and Yaron Lipman. Point convolutional neural networks by extension operators. *ACM Transactions on Graphics*, 2018. 7.2

[15] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Weakly-supervised domain adaptation via gan and mesh model for estimating 3d hand poses interacting objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3.1

[16] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. 2022. 4.2, 5.2

[17] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *ICLR*, 2022. 7.2

[18] Mathieu Barnachon, Saïda Bouakaz, Boubakeur Boufama, and Erwan Guillou. Ongoing human action recognition with motion capture. *Pattern Recognition*, 47(1):238–247, 2014. ISSN 0031-3203. 5.1

[19] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016. 3.2

[20] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. Dynamic image networks for action recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3034–3042, 2016. 5.2

[21] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-nms – improving object detection with one line of code. In *The International Conference on Computer Vision*, 2017. 2.2

[22] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *The International Conference on Computer Vision*, 2019. 2.4.3

[23] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 5.3.4

[24] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. In Yoshua Bengio and Yann LeCun, editors, *International Conference on Learning Representations*, 2014. 5.3.3

[25] Jinzheng Cai, Youbao Tang, Le Lu, Adam P Harrison, Ke Yan, Jing Xiao, Lin Yang, and Ronald M Summers. Accurate weakly-supervised deep lesion segmentation using large-scale clinical annotations: Slice-propagated 3d mask generation from 2d recist. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 396–404. Springer, 2018. 3.1

[26] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2.2

[27] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. L. Rojo-Alvarez, and M. Martinez-Ramon. Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 46(6):1822–1835, 2008. 2.2

[28] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 6.2

[29] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and

the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 6.1, 6.3, 6.1, 6.3.1

[30] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *3DV*, 2017. 7.1, 7.4.1

[31] Juan Chang, Xiangan Liu, Ryan Rochat, Matthew Baker, and Wah Chiu. Reconstructing virus structures from nanometer to near-atomic resolutions with cryo-electron microscopy and tomography. *Advances in experimental medicine and biology*, 726:49–90, 2012. 3.1

[32] Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S. Ryoo, Austin Stone, and Daniel Kappler. Open-vocabulary queryable scene representations for real world planning. *arXiv prepreint arXiv:2209.09874*, 2022. 7.2

[33] Hao Chen, Qi Dou, Lequan Yu, and Pheng-Ann Heng. Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation. *arXiv preprint arXiv:1608.05895*, 2016. 3.1, 3.2, 3.3.4, 3.4.7

[34] Hong Chen, Yifei Huang, and Hideki Nakayama. Semantic aware attention based deep object co-segmentation. In C.V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Asian Conference on Computer Vision*, pages 435–450, Cham, 2019. Springer International Publishing. ISBN 978-3-030-20870-7. 3.2

[35] Junjie Chen, Li Niu, Liu Liu, and Liqing Zhang. Weak-shot fine-grained classification via similarity transfer. *CoRR*, abs/2009.09197, 2020. 6.2

[36] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2.2

[37] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *International Conference on Computer Vision (ICCV)*, October 2019. 4.2, 6.1, 6.2, 6.4.2, ??, ??, ??, ??

[38] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal Attentive Alignment for Large-Scale Video Domain Adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6321–6330, 2019. 4.1, 4.3.1, 4.3.3, 4.4.2, 4.4.3, ??

[39] Sean Andrew Chen, Andrew Escay, Christopher Haberland, Tessa Schneider, Valentina Staneva, and Youngjun Choe. Benchmark dataset for automatic damaged building detection from post-hurricane remotely sensed imagery. *Arxiv*, 2018. 2.1, 2.2

[40] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *ICCV*, 2023. 7.2

[41] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3D pose estimation. In *3DV*, 2016. 6.2

[42] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368, 2021. **??**, 5.4.2, **??**, 6.4.2

[43] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 7.3.2, 7.3.3

[44] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 7.1, 7.3.2, 7.3.3, 7.4.3

[45] A. Cheraghian, S. Rahman, D. Campbell, and L. Petersson. Transductive zero-shot learning for 3d point cloud classification. In *WACV*, 2020. 7.2

[46] Ali Cheraghian, Shafin Rahman, Dylan Campbell, and Lars Petersson. Mitigating the hubness problem for zero-shot learning of 3d objects. In *BMVC*, 2019. 7.2

[47] Ali Cheraghian, Shafin Rahman, and Lars Petersson. Zero-shot learning of 3d point cloud objects. In *MVA*, 2019. 7.2

[48] Ali Cheraghian, Shafin Rahman, Townim F. Chowdhury, Dylan Campbell, and Lars Petersson. Zero-shot learning on 3d point cloud objects and beyond. *IJCV*, 2022. 7.2

[49] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 5.3.4

[50] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 7.1, 7.2, 7.3.3, 7.4.2, 7.1, 7.4.3

[51] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 3.2

[52] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14755–14764, 2022. 5.3.4

[53] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 7.1, 7.4.1, 8.2.1

[54] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *The Conference and Workshop on Neural Information Processing Systems*, 2016. 2.2

[55] Adrian V. Dalca, John Guttag, and Mert R. Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. 3.2

[56] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 4.1, 4.4.1

[57] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 4.1

[58] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022. 8.2.2

[59] Avijit Dasgupta, CV Jawahar, and Karteek Alahari. Overcoming label noise for source-free unsupervised video domain adaptation. In *Proceedings of the Thirteenth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–9, 2022. 4.4.3, **??**

[60] César Roberto De Souza, Adrien Gaidon, Yohann Cabon, and Antonio M. López Peña. Procedural generation of videos to train deep action recognition networks. In *CVPR*,

2017. 6.2

[61] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects, 2023. 1.1

[62] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR)*, 2009. 1.1, 5.3.4, 6.4.3

[63] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 4.2, 5.2

[64] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *CVPR*, 2023. 7.2

[65] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 6.2

[66] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021. 4.2

[67] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2021. 5.2, 6.4.3

[68] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 2015. 5.2, 6.2

[69] Francis Engelmann, Theodora Kontogianni, Alexander Hermans, and Bastian Leibe. Exploring spatial context for 3d semantic segmentation of point clouds. In *ICCV workshop*,

2017. 7.2

[70] Francis Engelmann, Theodora Kontogianni, Jonas Schult, and Bastian Leibe. Know what your neighbors do: 3d semantic segmentation of point clouds. In *ECCV workshop*, 2019. 7.2

[71] Michael F. Schmid and Steven J. Ludtke esús G. Galaz-Montoya, John Flanagan. Single particle tomography in eman2. In *Journal of structural biology*, 2015. 3.4.1

[72] A. Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 7.1, 7.2, 7.3.2, 7.4.3, 8.2.2

[73] EyesJapan. Eyes Japan. `https://mpcapdata.com`, 2018. 5.1, 5.2

[74] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2021. **??**, 5.4.2, **??**

[75] Hehe Fan, Xin Yu, Yuhang Ding, Yi Yang, and Mohan S. Kankanhalli. Pstnet: Point spatio-temporal convolution on point cloud sequences. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 5.2, **??**, 5.4.2, **??**, 6.2

[76] Junming Fan, Pai Zheng, and Shufei Li. Vision-based holistic scene understanding towards proactive human–robot collaboration. *Robotics and Computer-Integrated Manufacturing*, 75:102304, 2022. 7.1

[77] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. CIAN: cross-image affinity net for weakly supervised semantic segmentation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 10762–10769. AAAI Press, 2020. 3.2, 3.3.2

[78] Ruochen Fan, Qibin Hou, Ming-Ming Cheng, Gang Yu, Ralph R. Martin, and Shi-Min Hu. Associating inter-image salient instances for weakly supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, September 2018. 3.2

[79] Runyang Feng, Yixing Gao, Tze Ho Elden Tse, Xueqing Ma, and Hyung Jin Chang. Diffpose: Spatiotemporal diffusion model for video-based human pose estimation. In *ICCV*, 2023. 7.2

[80] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 2017. 6.2

[81] NOAA National Centers for Environmental Information (NCEI). U.s. billion-dollar weather and climate disasters, 2019. 2.1

[82] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *CVPR*, 2023. 7.2

[83] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. 05 2016. 6.2

[84] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 4.1, 4.3.1, 4.3.3, 4.3.3, 4.4.2, 4.4.3, **??**, **??**, **??**

[85] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *AAAI*, 2019. 6.2

[86] Mona Fathollahi Ghezelghieh, Rangachar Kasturi, and Sudeep Sarkar. Learning camera viewpoint using CNN to improve 3D body pose estimation. In *3DV*, 2016. 6.2

[87] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022. 7.1, 7.2, 7.3.2, 7.4.2, 7.4.3

[88] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. 2014. 4.3.3

[89] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018. 7.2

[90] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant

Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2022. 4.1, 4.4.1

[91] L. Gueguen, P. Soille, and M. Pesaresi. Change detection based on information measure. *IEEE Transactions on Geoscience and Remote Sensing*, 49(11):4503–4515, 2011. 2.2

[92] Lionel Gueguen and Raffay Hamid. Large-scale damage detection using satellite imagery. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2015. 2.2

[93] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R. Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, Apr 2021. ISSN 2096-0662. 5.3.3, 5.3.3

[94] Qiang Guo, Carina Lehmer, Antonio Martínez-Sánchez, Till Rudack, Florian Beck, Hannelore Hartmann, Manuela Pérez-Berlanga, Frédéric Frottin, Mark S. Hipp, F. Ulrich Hartl, Dieter Edbauer, Wolfgang Baumeister, and Rubén Fernández-Busnadiego. In situ structure of neuronal c9orf72 poly-ga aggregates reveals proteasome recruitment. *Cell*, 172(4):696 – 705.e12, 2018. ISSN 0092-8674. 3.4.1

[95] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xbd: A dataset for assessing building damage from satellite imagery. In *Arxiv*, 2019. 2.1, 2.2, 2.3.1, 2.4.1

[96] Charles A Hall and W.Weston Meyer. Optimal error bounds for cubic spline interpolation. *Journal of Approximation Theory*, 16(2):105 – 122, 1976. ISSN 0021-9045. 3.3.2

[97] Lei Han, Tian Zheng, Yinheng Zhu, Lan Xu, and Lu Fang. Live semantic 3d perception for immersive augmented reality. *IEEE Trans. visualization and computer graphics*, 26(5): 2012–2022, 2020. 7.1

[98] Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. Meshcnn: A network with an edge. *ACM Transactions on Graphics (TOG)*, 38(4): 90:1–90:12, 2019. 5.3.1

[99] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018. 6.1

[100] Bradley Hayes and Julie A Shah. Interpretable models for fast activity recognition and anomaly explanation during collaborative robotics tasks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6586–6593. IEEE, 2017. 5.1

[101] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *International Conference on Computer Vision*, 2017. 2.2, 2.4.1, 2.4.3, 2.4.4, 2.5.1, 2.5.2

[102] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 4.2, 4.3.3, 5.2, 5.3.4

[103] Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021. 1.1

[104] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *ICCV*, 2023. 7.2

[105] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *CVPR*, 2021. 7.4.2, 7.1

[106] Yonghong Hou, Zhaoyang Li, Pichao Wang, and Wanqing Li. Skeleton optical spectra-based action recognition using convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(3):807–811, 2018. 5.2, 6.2

[107] Lukas Hoyer, Mauricio Munoz, Prateek Katiyar, Anna Khoreva, and Volker Fischer. Grid saliency for context explanations of semantic segmentation. 2019. 4.3.4

[108] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. MIC: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), 2023. 4.2, 4.3.4

[109] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Co-attention cnns for unsupervised object co-segmentation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 748–756. International Joint Conferences on Artificial Intelligence Organization, 7 2018. 3.2

[110] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 8.2.4

[111] Zeyu Hu, Xuyang Bai, Jiaxiang Shang, Runze Zhang, Jiayu Dong, Xin Wang, Guangyuan Sun, Hongbo Fu, and Chiew-Lan Tai. Vmnet: Voxel-mesh network for geodesic-aware 3d semantic segmentation. In *ICCV*, 2021. 7.2

[112] Binh-Son Hua, Minh-Khoi Tran, and Sai-Kit Yeung. Pointwise convolutional neural networks. In *CVPR*, 2018. 7.2

[113] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *ICRA*, 2023. 7.2

[114] Jingwei Huang, Haotian Zhang, Li Yi, Thomas Funkhouser, Matthias Nießner, and Leonidas J Guibas. Texturenet: Consistent local parametrizations for learning from high-resolution signals on meshes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4440–4449, 2019. 7.4.2, 7.1

[115] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *CVPR*, 2022. 7.1

[116] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2702–2719, 2019. 1.1

[117] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7014–7023, 2018. 3.3.2

[118] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2.2

[119] Zhiyu Huang, Chen Lv, Yang Xing, and Jingda Wu. Multi-modal sensor fusion-based

deep neural network for end-to-end autonomous driving with scene understanding. *IEEE Sensors Journal*, 21(10):11781–11790, 2020. 7.1

[120] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep domain adaptation in action space. In *British Machine Vision Conference (BMVC)*, 2018. 4.2, 6.1, 6.2

[121] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. In *Robotics science and systems*, 2023. 7.1, 7.2, 7.3.2, 7.4.2, 7.1

[122] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. Ddp: Diffusion model for dense visual prediction. In *ICCV*, 2023. 7.2

[123] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 7.1, 7.2, 7.3.2

[124] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2.2

[125] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 5.4.6

[126] J Adam Jones, J Edward Swan, Gurjot Singh, Eric Kolstad, and Stephen R Ellis. The effects of virtual reality, augmented reality, and motion parallax on egocentric depth perception. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization*, pages 9–14, 2008. 4.1

[127] Thomas Joyce, Agisilaos Chartsias, and Sotirios A. Tsaftaris. Deep multi-class segmentation without ground-truth labels. 2018. 3.2

[128] Mohammad Kakooei and Yasser Baleghi. Fusion of satellite, aircraft, and uav data for automatic disaster damage assessment. *International Journal of Remote Sensing*, 38(8-10): 2511–2534, 2017. 2.2

[129] Guoliang Kang, Lu Jiang, Yunchao Wei, Yi Yang, and Alexander Hauptmann. Contrastive adaptation network for single- and multi-source domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1793–1804, 2022. 4.3.4

[130] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 4.4.2, 6.4.1

[131] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 4.2

[132] Oton J. Qu K. et al. Ke, Z. Structures and distributions of sars-cov-2 spike proteins on intact virions. *Nature*, 2020. 3.1

[133] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 3.2

[134] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13618–13627, 2021. 4.2

[135] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2.5.1, 3.4.4, 5.4.3, 7.4.3

[136] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 5.1, 5.2, 5.5, 5.4.6, 6.3.2

[137] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European conference on computer vision*, pages 695–711. Springer, 2016. 3.2

[138] Yu Kong, Zhengming Ding, Jun Li, and Yun Fu. Deeply learned view-invariant features for cross-view action recognition. *Transactions on Image Processing*, 26(6), 2017. 4.2, 6.1, 6.2

[139] Hema Koppula, Abhishek Anand, Thorsten Joachims, and Ashutosh Saxena. Semantic labeling of 3d point clouds for indoor scenes. In *NeurIPS*, 2011. 7.2

[140] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. Empirically analyzing the effect of dataset biases on deep face recognition systems. In *CVPRW*, 2018. 6.2

[141] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011. 3.3.6

[142] Alyssa Kubota, Tariq Iqbal, Julie A Shah, and Laurel D Riek. Activity recognition in manufacturing: The roles of motion capture and semg+ inertial wearables in detecting fine vs. gross motion. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6533–6539. IEEE, 2019. 5.1

[143] Werner Kühlbrandt. The resolution revolution. *Science*, 343(6178):1443–1444, 2014. ISSN 0036-8075. 3.1

[144] Jiahui Zhang MUYU XU Yingchen Yu Abdulmotaleb El Saddik Christian Theobalt Eric Xing Shijian Lu Kunhao Liu, Fangneng Zhan. Weakly supervised 3d open-vocabulary segmentation. In *NeurIPS*, 2023. 7.1

[145] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *ICLR*, 2023. 7.1, 7.2

[146] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. MSeg: A composite dataset for multi-domain semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 7.1

[147] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *CVPR*, 2018. 7.2

[148] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *ICCV*, 2023. 7.2

[149] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 7.1, 7.2, 7.3.2

[150] Haoxin Li, Yijun Cai, and Wei-Shi Zheng. Deep dual relation modeling for egocentric interaction recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7932–7941, 2019. 4.1

[151] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan Kankanhalli. Unsupervised learning of view-invariant action representations. In *Advances in Neural Information Processing Systems (Neurips)*, 2018. 4.2, 6.1, 6.2

[152] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 7.1, 7.2

[153] Min Li, Shizhou Dong, Kun Zhang, Zhifan Gao, Xi Wu, Heye Zhang, Guang Yang, and Shuo Li. Deep learning intra-image and inter-images features for co-saliency detection. In *BMVC*, page 291, 09 2018. 3.2

[154] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 9–14, 2010. 5.1

[155] Wen Li, Zheng Xu, Dong Xu, Dengxin Dai, and Luc Van Gool. Domain generalization and adaptation using low rank exemplar SVMs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1114–1127, 2018. 7.2

[156] Xiangyu Li, Yonghong Hou, Pichao Wang, Zhimin Gao, Mingliang Xu, and Wanqing Li. Trear: Transformer-based rgb-d egocentric action recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):246–252, 2021. 4.2

[157] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2020. 3.2

[158] Xing Li, Qian Huang, Zhijian Wang, Zhenjie Hou, and Tianjin Yang. Sequentialpointnet: A strong frame-level parallel point cloud sequence network for 3d action recognition, 2021. **??**, 5.4.2, **??**

[159] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021. 1.1

[160] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 287–295, 2015. 4.2

[161] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carry-

ing location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 5.4.6

[162] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *ICCV*, 2023. 7.2

[163] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, 2023. 7.2

[164] Hui Liang, Junsong Yuan, Daniel Thalmann, and Nadia Magnenat Thalmann. Ar in hand: Egocentric palm pose tracking and gesture recognition for augmented reality applications. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 743–744, 2015. 4.1

[165] Junwei Liang. From recognition to prediction: Analysis of human action and trajectory prediction in video. *arXiv preprint arXiv:2011.10670*, 2020. (document)

[166] Junwei Liang, Lu Jiang, and Alexander Hauptmann. Temporal localization of audio events for conflict monitoring in social media. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1597–1601. IEEE, 2017. 2.1

[167] Junwei Liang, Liangliang Cao, Xuehan Xiong, Ting Yu, and Alexander Hauptmann. Spatial-temporal alignment network for action recognition and detection. *arXiv preprint arXiv:2012.02426*, 2020. 1.1

[168] Junwei Liang, Lu Jiang, and Alexander Hauptmann. Simaug: Learning robust representations from simulation for trajectory prediction. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 275–292, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58601-0. 6.2

[169] Junwei Liang, Lu Jiang, Kevin Murphy, Ting Yu, and Alexander Hauptmann. The garden of forking paths: Towards multi-future trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 6.2

[170] X. Liao, W. Li, Q. Xu, X. Wang, B. Jin, X. Zhang, Y. Wang, and Y. Zhang. Iteratively-refined interactive 3d medical image segmentation with multi-agent reinforcement learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9391–9399, 2020. 3.4.1

[171] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised

convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3159–3167, 2016. 3.2

[172] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 2.3.3, 2.5, 2.5.1

[173] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2.4.1, 2.4.3

[174] Bo Liu, Shuang Deng, Qiulei Dong, and Zhanyi Hu. Language-level semantics conditioned 3d point cloud segmentation. *arXiv prepreint arXiv:2107.00430*, 2022. 7.2

[175] Chuang Liu, Luiza Mendonça, Yang Yang, Yuanzhu Gao, Chenguang Shen, Jiwei Liu, Tao Ni, Bin Ju, Congcong Liu, Xian Tang, Jinli Wei, Xiaomin Ma, Yanan Zhu, Weilong Liu, Shuman Xu, Yingxia Liu, Jing Yuan, Jing Wu, Zheng Liu, Zheng Zhang, Lei Liu, Peiyi Wang, and Peijun Zhang. The architecture of inactivated sars-cov-2 with postfusion spikes revealed by cryo-em and cryo-et. *Structure*, 28(11):1218 – 1224.e4, 2020. ISSN 0969-2126. 3.1

[176] Daochang Liu, Qiyue Li, Anh-Dung Dinh, Tingting Jiang, Mubarak Shah, and Chang Xu. Diffusion action segmentation. In *ICCV*, 2023. 7.2

[177] Jian Liu, Naveed Akhtar, and Ajmal Mian. Temporally coherent full 3D mesh human pose recovery from monocular video. *CoRR*, abs/1906.00161, 2019. 6.2

[178] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C. Kot. Global context-aware attention lstm networks for 3d action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 5.2, 6.2

[179] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2020. 5.1

[180] Jun Liu, Amir Shahroudy, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Skeleton-based online action prediction using scale selection network. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(6):1453–1467, 2020. 5.2, 6.2

[181] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view

invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017. 4.2, 6.1, 6.2

[182] Miao Liu, Siyu Tang, Yin Li, and James Rehg. Forecasting human object interaction: Joint prediction of motor attention and actions in first person video. In *ECCV*, 2020. 4.1

[183] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022. 7.1

[184] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015. 3.4.4

[185] Xingyu Liu, Mengyuan Yan, and Jeannette Bohg. Meteornet: Deep learning on dynamic 3d point cloud sequences. In *ICCV*, 2019. 5.2, 6.2

[186] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022, June 2022. 4.1

[187] Zhengzhe Liu, Xiaojuan Qi, and Chi-Wing Fu. 3d-to-2d distillation for indoor scene parsing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4464–4474, 2021. 7.1

[188] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 143–152, 2020. **??**, 5.4.2, **??**, 6.4.2

[189] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *SIGGRAPH Asia*, 2015. 5.3.1, 6.3.2

[190] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 4.4.2

[191] Minlong Lu, Ze-Nian Li, Yueming Wang, and Gang Pan. Deep attention network for egocentric action recognition. *IEEE Transactions on Image Processing*, 28(8):3703–3713, 2019. 4.2

[192] Yan Lu and Christopher Rasmussen. Simplified markov random fields for efficient semantic labeling of 3d point clouds. In *ICIRS*, 2012. 7.2

[193] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? *arXiv preprint*

arXiv:2212.07796, 2023. 7.1

[194] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019. 5.1, 5.3.1

[195] Adrien Malaisé, Pauline Maurice, Francis Colas, François Charpillet, and Serena Ivaldi. Activity recognition with multiple wearable sensors for industrial applications. In *ACHI 2018-Eleventh International Conference on Advances in Computer-Human Interactions*, 2018. 5.1

[196] Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed, Fahad Shahbaz Khan, and Ling Shao. Out-of-distribution detection for generalized zero-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9985–9993, 2019. 6.2

[197] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The kit whole-body human motion database. In *International Conference on Advanced Robotics (ICAR)*, pages 329–336, 2015. 5.1, 5.4.1

[198] Javier Marin, David Vazquez, David Geronimo, and Antonio M. Lopez. Learning appearance in virtual scenarios for pedestrian detection. In *CVPR*, 2010. 6.2

[199] Iacopo Masi, Anh Tuan Tran, Tal Hassner, Gozde Sahin, and Gérard Medioni. Face-specific data augmentation for unconstrained face recognition. *International Journal of Computer Vision (IJCV)*, 2019. 6.2

[200] Kirill Mazur, Edgar Sucar, and Andrew Davison. Feature-realistic neural fusion for real-time, open set scene understanding. In *ICRA*, 2023. 7.2

[201] Rahil Mehrizi, Xi Peng, Xu Xu, Shaoting Zhang, Dimitris N. Metaxas, and Kang Li. A computer vision based method for 3d posture estimation of symmetrical lifting. *Journal of biomechanics*, 69:40–46, 2018. 5.1

[202] Jilin Mei, Biao Gao, Donghao Xu, Wen Yao, Xijun Zhao, and Huijing Zhao. Semantic segmentation of 3d lidar data in dynamic scene using semi-supervised learning. *IEEE Transactions on Intelligent Transportation Systems*, 21(6):2496–2509, 2019. 3.1

[203] Celso Melo, Brandon Rothrock, Prudhvi Gurram, Oytun Ulutan, and B. Manjunath. Vision-based gesture recognition in human-robot teams using synthetic data. pages 10278–10284, 10 2020. 5.2

[204] Qinghao Meng, Wenguan Wang, Tianfei Zhou, Jianbing Shen, Luc Van Gool, and

Dengxin Dai. Weakly supervised 3d object detection from lidar point cloud. In *European Conference on computer vision*, pages 515–531. Springer, 2020. 1.1

[205] Matteo Menolotto, Dimitrios-Sokratis Komaris, Salvatore Tedesco, Brendan O'Flynn, and Michael Walsh. Motion capture technology in industrial applications: A systematic review. *Sensors*, 20(19):5687, 2020. 5.1

[206] G. Mercier, G. Moser, and S. B. Serpico. Conditional copulas for change detection in heterogeneous remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 46(5):1428–1441, 2008. 2.2

[207] Björn Michele, Alexandre Boulch, Gilles Puy, Maxime Bucher, and Renaud Marlet. Generative zero-shot learning for semantic segmentation of 3D point cloud. In *3DV*, 2021. 7.1, 7.2

[208] Ashish Mishra, Vinay Kumar Verma, M. Shiva Krishna Reddy, Arulkumar S., Piyush Rai, and Anurag Mittal. A generative approach to zero-shot and few-shot action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 372–380, 2018. 6.2

[209] Sarthak Mittal, Korbinian Abstreiter, Stefan Bauer, Bernhard Schölkopf, and Arash Mehrjou. Diffusion based representation learning. In *ICML*, 2023. 7.1, 7.2

[210] MocapClub. Motion Capture Club. `http://www.mocapclub.com/`, 2009. 5.1, 5.2

[211] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4.2, 4.3.1, 4.4.2, 4.4.3, **??**, 6.1

[212] Jonathan Munro and Dima Damen. Multi-modal Domain Adaptation for Fine-grained Action Recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 4.1, 4.3.1, 4.3.3

[213] Meinard Müller, Tido Röder, Michael Clausen, Bernhard Eberhardt, Björn Krüger, and Andreas Weber. Documentation mocap database hdm05, 2007. 5.1, 5.2

[214] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 163–172, 2020. 4.1

[215] Dat T. Nguyen, Ferda Ofli, Muhammad Imran, and Prasenjit Mitra. Damage assessment from social media imagery data during disasters. In *Proceedings of the 2017*

IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, ASONAM '17, page 569–576, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349932. 2.2, 2.2

[216] A. A. Nielsen. The regularized iteratively reweighted mad method for change detection in multi- and hyperspectral data. *IEEE Transactions on Image Processing*, 16(2):463–478, 2007. 2.2

[217] Department of Homeland Security Federal Emergency Management Agency (FEMA). Damage assessment operations manual, 2016. 2.1

[218] OSU. ACCAD. https://accad.osu.edu/research/motion-lab/system-data, 2018. 5.1, 5.2

[219] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. *AAAI Conference on Artificial Intelligence*, 2020. 4.2, 6.1, 6.2

[220] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015. 3.1

[221] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. 2016. 4.2

[222] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 5.3.1, 6.3.2

[223] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, 2023. 7.1, 7.2, 7.2, 7.3, 7.3.2, 7.3.3, 7.4, 7.4.1, 7.4.2, 7.1, 7.4.3, 7.2, 7.4.4

[224] Yan Peng, Chenjun Shi, Yiming Zhu, Min Gu, and Songlin Zhuang. Terahertz spectroscopy in biomedical field: a review on signal-to-noise ratio improvement. *PhotoniX*, 1:1–18, 2020. 1.1

[225] Juan-Manuel Pérez-Rúa, Xiatian Zhu, Timothy M. Hospedales, and Tao Xiang. Incremental few-shot object detection. In *2020 IEEE/CVF Conference on Computer Vision and*

Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 13843–13852. Computer Vision Foundation / IEEE, 2020. 6.2

[226] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormählen, and Bernt Schiele. Articulated people detection and pose estimation: Reshaping the future. In *CVPR*, 2012. 6.2

[227] Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and Barbara Caputo. Domain generalization through audio-visual relative norm alignment in first person action recognition. *arXiv preprint arXiv:2110.10101*, 2021. 4.2

[228] Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, and Barbara Caputo. E2 (go) motion: Motion augmented event stream for egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19935–19947, 2022. 4.2

[229] Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Anselm Levskaya, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference, 2022. 8.2.2

[230] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. VirtualHome: Simulating household activities via programs. In *CVPR*, 2018. 6.2

[231] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, June 2021. 5.1, 5.2, 5.4.1, 5.4.1, 5.4.2

[232] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv:1612.00593*, 2016. 5.2, 5.3.2, 6.2, 7.2

[233] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 3.1, 5.2, 5.3.2, 6.2, 7.2

[234] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *ECCV*, 2018. 6.2

[235] Hui Qu, Pengxiang Wu, Qiaoying Huang, Jingru Yi, Zhennan Yan, Kang Li, Gregory M.

Riedlinger, Subhajyoti De, Shaoting Zhang, and Dimitris N. Metaxas. Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. *IEEE Transactions on Medical Imaging*, page 1–1, 2020. ISSN 1558-254X. 3.1, 3.2

[236] Hossein Rahmani and Ajmal Mian. Learning a non-linear knowledge transfer model for cross-view action recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 4.2, 6.1, 6.2

[237] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. 2021. 4.2, 5.2

[238] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Conference on Neural Information Processing Systems*, 2015. 2.2, 2.4.4

[239] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. Languagerefer: Spatial-language model for 3d visual grounding. In *Conference on Robot Learning*, pages 1046–1056. PMLR, 2022. 7.1

[240] Vito Romaniello, Alessandro Piscini, Christian Bignami, Roberta Anniballe, and Salvatore Stramondo. Earthquake damage mapping by using remotely sensed data: the Haiti case study. *Journal of Applied Remote Sensing*, 11(1):1 – 16, 2017. 2.1, 2.2

[241] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 7.1, 7.2, 7.3.1, 7.3.3, 7.4.3, 8.2.2

[242] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017. 5.1, 5.3.1, 6.3.2

[243] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. 2015. 1.1, 4.3.3, 8.2.2

[244] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. 2016. 6.2

[245] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *ECCV*, 2022. 7.1, 7.2, 7.4.1, 7.4.1, 7.4.2, 7.1, 7.4

[246] Tim G. J. Rudner, Marc Rußwurm, Jakub Fil, Ramona Pelich, Benjamin Bischke, Veronika

Kopackova, and Piotr Bilinski. Multi^{3}net: Segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery. In *Conference on Artificial Intelligence*, 2019. 2.2

[247] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019. 6.2

[248] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 7.2

[249] Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, and Abir Das. Contrast and mix: Temporal contrastive video domain adaptation with background mixing. *arXiv preprint arXiv:2110.15128*, 2021. 4.2, 4.3.1, 4.3.3, 4.4.1, 4.4.2, 4.4.3, ??, ??, ??, ??

[250] Adrian Sanchez-Caballero, Sergio de López Diz, David Fuentes-Jiménez, Cristina Losada-Gutiérrez, Marta Marrón Romera, David Casillas-Perez, and Mohammad Ibrahim Sarker. 3dfcnn: Real-time action recognition using 3d deep neural networks with raw depth information. *CoRR*, abs/2006.07743, 2020. 5.2, 6.2

[251] Adrian Sanchez-Caballero, David Fuentes-Jiménez, and Cristina Losada-Gutiérrez. Exploiting the convlstm: Human action recognition using raw depth video-based recurrent neural networks. *CoRR*, abs/2006.07744, 2020. 5.2, 6.2

[252] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1.1, 7.1, 7.4.3

[253] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. In *ICRA*, 2023. 7.1, 7.3.3

[254] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct 2019. ISSN 1573-1405. 3.3.2, 3.4.3

[255] A. Senior, G. Heigold, M. Ranzato, and K. Yang. An empirical study of learning rates in deep neural networks for speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6724–6728, 2013. 3.4.4

[256] SFU. SFU Motion Capture Database. http://mocap.cs.sfu.ca/. 5.1, 5.2

[257] Nur Muhammad Mahi Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. CLIP-fields: Weakly supervised semantic fields for robotic memory. In *CoRL Workshop on Language and Robotics*, 2022. 7.2

[258] Dhruv Shah, Blazej Osinski, Brian Ichter, and Sergey Levine. LM-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Annual Conference on Robot Learning*, 2022. 7.2

[259] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *CVPR*, 2016. 5.1

[260] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *ICCV*, 2023. 7.2

[261] Nicholas Sharp, Souhaib Attaiki, Keenan Crane, and Maks Ovsjanikov. Diffusionnet: Discretization agnostic learning on surfaces. *ACM Trans. Graph.*, 41(3), mar 2022. ISSN 0730-0301. 5.3.1

[262] Tong Shen, Guosheng Lin, Lingqiao Liu, Chunhua Shen, and Ian Reid. Weakly supervised semantic segmentation based on web image co-segmentation. *arXiv preprint arXiv:1705.09052*, 2017. 3.2

[263] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. https://github.com/abhinanda-punnakkal/BABEL/tree/main/action_recognition, 2019. 5.1, 6.2

[264] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019. **??**, **??**, 5.4.2, **??**, **??**, 6.4.2

[265] Yuge Shi, N Siddharth, Philip Torr, and Adam R Kosiorek. Adversarial masking for self-supervised learning. In *International Conference on Machine Learning*, pages 20026–20040. PMLR, 2022. 4.2

[266] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011. 6.2

[267] Zhenyu Shu, Xiaoyong Shen, Shiqing Xin, Qingjun Chang, Jieqing Feng, Ladislav Kavan, and Ligang Liu. Scribble based 3d shape segmentation via weakly-supervised learning. *IEEE transactions on visualization and computer graphics*, 2019. 3.1

[268] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 4.2, 6.1, 6.2

[269] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3136–3145, 2019. 3.2

[270] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 1.1

[271] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9787–9795, 2021. 4.2, 4.4.3, **??**

[272] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 7.1, 7.4.1, 7.2, 7.4.4, 7.4.4, 8.2.1

[273] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9954–9963, 2019. 4.1, 4.2

[274] Jiankai Sun, Bolei Zhou, Michael J Black, and Arjun Chandrasekaran. Locate: End-to-end localization of actions in 3d with transformers. *arXiv preprint arXiv:2203.10719*, 2022. 5.4.1

[275] J Edward Swan, Adam Jones, Eric Kolstad, Mark A Livingston, and Harvey S Smallman. Egocentric depth judgments in optical, see-through augmented reality. *IEEE transactions*

on visualization and computer graphics, 13(3):429–442, 2007. 4.1

[276] A Szabo, K Boucher, WL Carroll, LB Klebanov, AD Tsodikov, and AY Yakovlev. Variable selection and pattern recognition with gene expression data generated by the microarray technology. *Mathematical Biosciences*, 176(1):71–98, 2002. 3.3.3

[277] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *NeurIPS*, 2023. 7.1, 7.2, 7.3.2, 7.4.1, 7.4.1, 7.4.2, 7.1

[278] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa. Joint optimization framework for learning with noisy labels. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5552–5560, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society. 6.2

[279] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023. 7.1

[280] Maxim Tatarchenko*, Jaesik Park*, Vladlen Koltun, and Qian-Yi Zhou. Tangent convolutions for dense prediction in 3D. *CVPR*, 2018. 7.4.2, 7.1

[281] Lyne Tchapmi, Christopher Choy, Iro Armeni, JunYoung Gwak, and Silvio Savarese. Segcloud: Semantic segmentation of 3d point clouds. In *2017 international conference on 3D vision (3DV)*, pages 537–547. IEEE, 2017. 3.1, 7.2

[282] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, Francois Goulette, and Leonidas J. Guibas. KPConv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 7.2

[283] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 4.2, 4.4.2

[284] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, 2011. 6.1

[285] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 6.1

[286] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 4.2, 6.1, 6.2

[287] Nikolaus F. Troje. Decomposing biological motion: a framework for analysis and synthesis of human gait patterns. *Journal of vision*, 2 5:371–87, 2002. 5.1

[288] Victor G. Turrisi da Costa, Giacomo Zara, Paolo Rota, Thiago Oliveira-Santos, Nicu Sebe, Vittorio Murino, and Elisa Ricci. Dual-head contrastive domain adaptation for video action recognition. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2234–2243, 2022. 6.4.1, 6.4.2, **??**, **??**

[289] Lachlan Tychsen-Smith and Lars Petersson. Improving object localization with fitness nms and bounded iou loss. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2.2

[290] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial Discriminative Domain Adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 4.1, 4.3.1, 4.3.3, 4.3.3, 4.4.2, 4.4.3, **??**, 6.4.2, **??**, **??**

[291] C. Vaduva, T. Costachioiu, C. Patrascu, I. Gavat, V. Lazarescu, and M. Datcu. A latent analysis of earth surface dynamic evolution using change map time series. *IEEE Transactions on Geoscience and Remote Sensing*, 51(4):2105–2118, 2013. 2.2

[292] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 6.2

[293] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *ECCV*, 2018. 6.2

[294] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic humans for action recognition from unseen viewpoints. *International Journal of Computer Vision*, 129, 07 2021. 6.2

[295] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 5.3.3

[296] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 839–847, 2017. 6.2

[297] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. 2008. 4.2

[298] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. 2010. 4.2, 5.2

[299] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matthew J. Kusner. Unsupervised point cloud pre-training via occlusion completion. In *International Conference on Computer Vision, ICCV*, 2021. 5.2, 5.3.4

[300] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *ICCV*, 2023. 7.2

[301] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2.2

[302] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *arXiv preprint arXiv:2309.02773*, 2023. 7.3.2

[303] Pichao Wang, Zhaoyang Li, Yonghong Hou, and Wanqing Li. Action recognition based on joint trajectory maps using convolutional neural networks. *CoRR*, abs/1611.02447, 2016. 5.2, 6.2

[304] Pichao Wang, Wanqing Li, Zhimin Gao, Chang Tang, and Philip O. Ogunbona. Depth pooling based large-scale 3-d action recognition with convolutional neural networks. *IEEE Trans. Multim.*, 20(5):1051–1061, 2018. 5.2, 6.2

[305] Shuihua Wang, M Emre Celebi, Yu-Dong Zhang, Xiang Yu, Siyuan Lu, Xujing Yao, Qinghua Zhou, Martinez-Garcia Miguel, Yingli Tian, Juan M Gorriz, et al. Advances in data preprocessing for biomedical data fusion: An overview of the methods, challenges, and prospects. *Information Fusion*, 76:376–421, 2021. 1.1

[306] Tianyi Wang, Jian Li, and Xiangjing An. An efficient scene semantic labeling approach for 3d point cloud. In *ITSC*, 2015. 7.2

[307] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *IEEE International Conference on Computer Vision*, 2016. 2.4.3, 2.5.1

[308] Y. Wang, Y. Xiao, F. Xiong, W. Jiang, Z. Cao, J. Zhou, and J. Yuan. 3dv: 3d dynamic voxel for action recognition in depth video. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 508–517, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. 5.2, 6.2

[309] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020. 3.4.2

[310] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. 2022. 4.2

[311] Jiacheng Wei, Guosheng Lin, Kim-Hui Yap, Tzu-Yi Hung, and Lihua Xie. Multi-path region mining for weakly supervised 3d semantic segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4384–4393, 2020. 3.1, 3.3.2, 3.3.3, 3.3.4, 3.4.3, 3.4.5

[312] Pengfei Wei, Lingdong Kong, Xinghua Qu, Yi Ren, Zhiqiang Xu, Jing Jiang, and Xiang Yin. Unsupervised video domain adaptation for action recognition: A disentanglement perspective. 2023. 4.4.3, **??**

[313] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S. Huang. Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3.3.4

[314] Yan Xia, Xudong Cao, Fang Wen, Gang Hua, and Jian Sun. Learning discriminative reconstructions for unsupervised outlier removal. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1511–1519, 2015. 6.2

[315] Yang Xiao, Jun Chen, Yancheng Wang, Zhiguo Cao, Joey Tianyi Zhou, and Xiang Bai. Action recognition for depth video using multi-view dynamic images. *Inf. Sci.*, 480:287–304, 2019. 5.2, 6.2

[316] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2.5.1, 2.5.2

[317] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. 2022. 4.2

[318] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023. 7.1, 7.2, 7.3.2, 7.3.4, 7.4.3

[319] Xun Xu and Gim Hee Lee. Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3.1

[320] Yanwu Xu, Mingming Gong, Junxiang Chen, Ziye Chen, and Kayhan Batmanghelich. 3d-boxsup: Positive-unlabeled learning of brain tumor segmentation networks from 3d bounding boxes. *Frontiers in Neuroscience*, 14:350, 2020. 3.1

[321] Zhenjia Xu, Zhanpeng He, Jiajun Wu, and Shuran Song. Learning 3d dynamic scene representations for robot manipulation. *arXiv preprint arXiv:2011.01968*, 2020. 7.1

[322] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7444–7452. AAAI Press, 2018. 5.2, 6.2

[323] Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14722–14732, 2022. 4.4.3, **??**

[324] Tong Yang, Xiangyu Zhang, Zeming Li, Wenqiang Zhang, and Jian Sun. Metaanchor: Learning to detect objects with customized anchors. In *Conference on Neural Information Processing Systems*, 2018. 2.2

[325] Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *ICCV*, 2023. 7.1, 7.2

[326] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8808–8817, 2020. 6.2

[327] Yuehao Yin, Bin Zhu, Jingjing Chen, Lechao Cheng, and Yu-Gang Jiang. Mix-dann and dynamic-modal-distillation for video domain adaptation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3224–3233, 2022. 4.4.3, **??**

[328] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5.2

[329] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *2018 International Conference on 3D Vision (3DV)*, pages 728–737, 2018. 5.3.5

[330] Xiangrui Zeng and Min Xu. Gum-net: Unsupervised geometric matching for fast and accurate 3d subtomogram image alignment and averaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020. 3.4.1, 3.4.1

[331] Xiangrui Zeng, Miguel Ricardo Leung, Tzviya Zeev-Ben-Mordehai, and Min Xu. A convolutional autoencoder approach for mining features in cellular electron cryo-tomograms and weakly supervised coarse segmentation. *Journal of Structural Biology*, 202(2):150 – 160, 2018. ISSN 1047-8477. 3.4.1

[332] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022. 1.1

[333] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *European Conference on Computer Vision*, pages 127–145. Springer, 2022. 4.1

[334] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *CVPR*, 2022. 7.2

[335] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Single-shot refinement neural network for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2.2

[336] Songyang Zhang, Xiaoming Liu, and Jun Xiao. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 148–157, 2017. 5.2, 6.2

[337] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency

detection. *arXiv preprint arXiv:2004.13364*, 2020. 3.3.3

[338] Ziyi Zhang, Weikai Chen, Hui Cheng, Zhen Li, Siyuan Li, Liang Lin, and Guanbin Li. Divide and contrast: Source-free domain adaptation via adaptive contrastive learning. 2022. 4.3.4

[339] Guannan Zhao, Bo Zhou, Kaiwen Wang, Rui Jiang, and Min Xu. Respond-cam: Analyzing deep models for 3d imaging data by visualizations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 485–492. Springer, 2018. 3.4.3, 3.4.4

[340] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *ICCV*, 2023. 7.3.2

[341] Wenping Zhao, Jinxiang Chai, and Ying-Qing Xu. Combining marker-based mocap and rgb-d camera for acquiring high-fidelity hand motion data. In *Proceedings of the ACM SIGGRAPH/eurographics symposium on computer animation*, pages 33–42, 2012. 5.1, 5.2

[342] Zhuo Zhao, Lin Yang, Hao Zheng, Ian H Guldner, Siyuan Zhang, and Danny Z Chen. Deep learning based instance segmentation in 3d biomedical images using weak annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 352–360. Springer, 2018. 3.1

[343] Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly supervised contrastive learning. In *ICCV*, 2021. 7.4.2, 7.1

[344] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 3.2

[345] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *European Conference on Computer Vision (ECCV)*, 2018. 4.2, 4.4.3, 6.1, 6.2

[346] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image bert pre-training with online tokenizer. 2022. 4.2

[347] Chun Zhu and Weihua Sheng. Motion-and location-based online human daily activity recognition. *Pervasive and Mobile Computing*, 7(2):256–269, 2011. 5.1

[348] Xiaoyu Zhu, Jeffrey Chen, Xiangrui Zeng, Junwei Liang, Chengqi Li, Sinuo Liu, Sima Behpour, and Min Xu. Weakly supervised 3d semantic segmentation using cross-image

consensus and inter-voxel affinity relations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2834–2844, 2021. 1.1, 1.3, 7.2

[349] Xiaoyu Zhu, Junwei Liang, and Alexander Hauptmann. Msnet: A multilevel instance segmentation network for natural disaster damage assessment in aerial videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2023–2032, January 2021. 1.1, 1.3, 6.2

[350] Xiaoyu Zhu, Celso M de Melo, and Alexander Hauptmann. Leveraging body pose estimation for gesture recognition in human-robot interaction using synthetic data. In *Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications*, volume 12529, pages 242–250. SPIE, 2023. 1.1, 1.3, 4.2, 6.3, 6.1, 6.3.2, 6.3.3

[351] Xiaoyu Zhu, Po-Yao Huang, Junwei Liang, Celso M De Melo, and Alexander G Hauptmann. Stmt: A spatial-temporal mesh transformer for mocap-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1526–1536, 2023. 1.3, 4.2, 6.1, 6.2, 6.3, 6.1, 6.3.2, 6.4.2

[352] Xiaoyu Zhu, Junwei Liang, Po-Yao Huang, and Alex Hauptmann. Adversarially masked video consistency for unsupervised domain adaptation. *arXiv preprint arXiv:2403.16242*, 2024. 1.1, 1.3, 6.1

[353] Xiaoyu Zhu, Wenhe Liu, Celso M de Melo, and Alexander Hauptmann. Multi-modal knowledge distillation for domain-adaptive action recognition. In *Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications*. SPIE, 2024. 1.3

[354] Xiaoyu Zhu, Hao Zhou, Pengfei Xing, Long Zhao, Hao Xu, Junwei Liang, Alexander Hauptmann, Ting Liu, and Andrew Gallagher. Open-vocabulary 3d semantic segmentation with text-to-image diffusion models. In *European Conference on Computer Vision*, pages 357–375, 2024. 1.3