Weakly Supervised 3D Semantic Segmentation Using Cross-Image Consensus and Inter-Voxel Affinity Relations

Xiaoyu Zhu[1] Jeffrey Chen[1] Xiangrui Zeng[1] Junwei Liang[1]
Chengqi Li[2] Sinuo Liu[1] Sima Behpour[1] Min Xu[1]*
[1]Carnegie Mellon University [2]University of California San Diego
{xiaoyuz3, jc6, xiangruz, junweil, mxu1}@cs.cmu.edu
{lichengqi0805, liusinuo1994, sima.behpour}@gmail.com

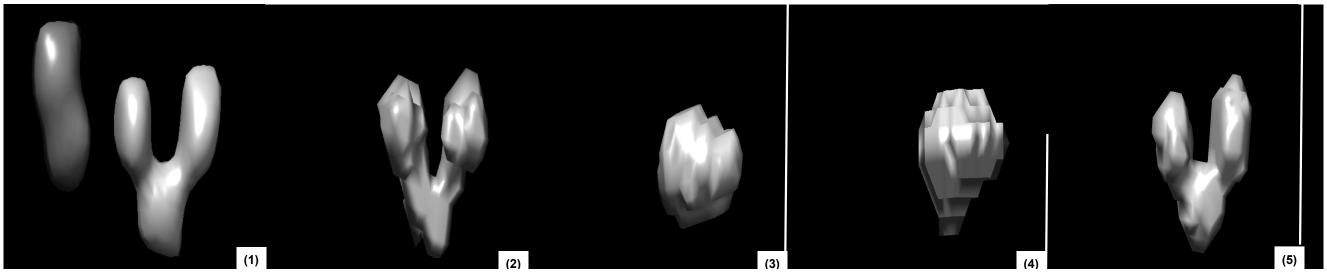

Figure 1: Illustration of 3D semantic segmentation using image-level class labels as supervision. This figure shows: (1) input 3D cryo-ET image; (2) ground truth segmentation; (3) semantic segmentation generated by Grad-CAM baseline. It only covers the most discriminative area; (4) Grad-CAM results augmented 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[1].

Abstract

We propose a novel weakly supervised approach for 3D semantic segmentation on volumetric images. Unlike most existing methods that require voxel-wise densely labeled training data, our weakly-supervised CIVA-Net is the first model that only needs image-level class labels as guidance to learn accurate volumetric segmentation. Our model learns from cross-image co-occurrence for integral region generation, and explores inter-voxel affinity relations to predict segmentation with accurate boundaries. We empirically validate our model on both simulated and real cryo-ET datasets. Our experiments show that CIVA-Net achieves comparable performance to the state-of-the-art models trained with stronger supervision.

1. Introduction

Recently, there has been an increasing interest in semantic segmentation for 3D images [66, 11, 50, 42]. 3D semantic segmentation methods that rely on point-wise annotations have been successfully developed and achieved promising performance [11, 50, 42]. 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 [56, 53, 41, 40] and meshes [47, 7]. 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 [9, 13], bounding boxes [57, 62] or sparse point annotations [43].

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

[1] Note that we use UCSF Chimera for 3D cryo-ET image visualization. For raw 3D images, the software will choose a threshold based on expert knowledge and denote the foreground voxels by light color, and noise/background voxels by black color. For binary segmentation masks, it will denote the foreground voxels by light color, and background voxels by black color.

* Corresponding Author

studying macromolecular complexes and virus structures in single cells [10]. Cryo-ET captures the 3D native structure and spatial distribution of all macromolecular complexes and other subcellular components without disrupting the cell [30]. 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 [24, 36]. 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 1, 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). 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.

2. Related Work

Weakly Supervised Semantic Segmentation on 2D Images. Recent studies [25, 48, 34, 8] 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 [25, 48], scribble [34], and point annotation [8]. 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 [63] 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 [28, 6, 5] 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 [17, 21, 16, 65]. Many 2D co-segmentation approaches were trained with strong pixel-level masks [12, 31]. Some weakly supervised methods used co-segmentation for initial seeds generation or incorporated the co-segmentation module to an end-to-end framework [46, 16]. 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

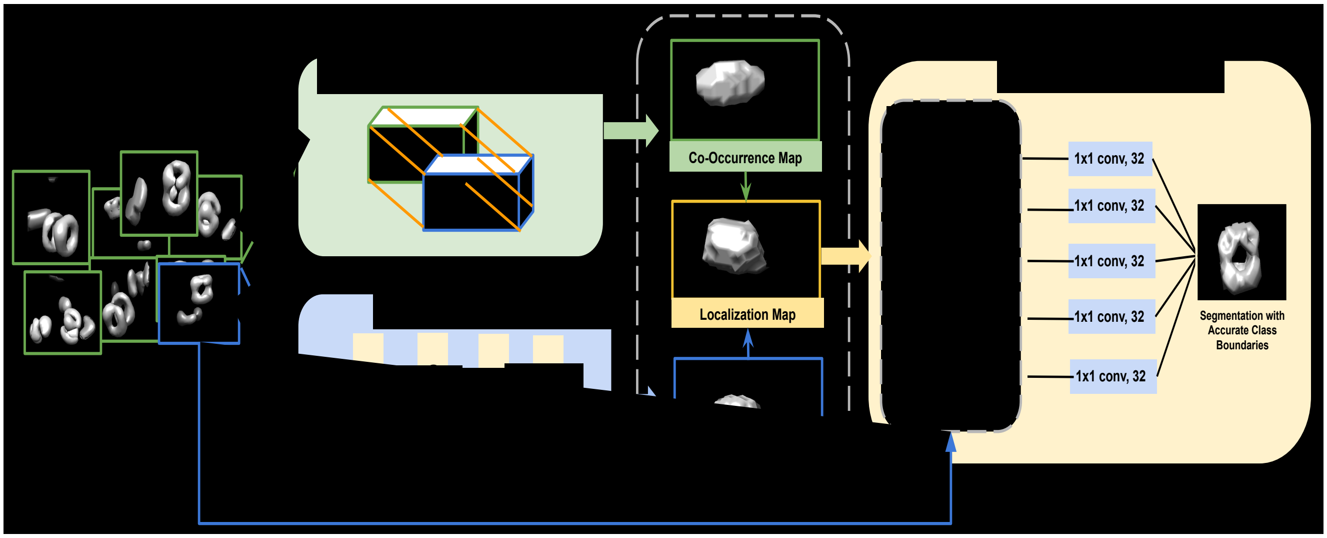

Figure 2: Network architecture of *CIVA-Net*. The left part includes the initial seeds generation module, which is combined with a cross-image co-occurrence learning module to generate more integral seed areas. For the initial seed generation, Grad-CAM is used to take single images as input to train 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.

learning. Supervised learning approaches have gained popularity in recent years [13, 11]. Cicek et al. proposed 3D U-Net [13] which extended previous U-Net architecture by replacing all 2D operations with their 3D counterparts. Chen et.al proposed VoxResNet [11] 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 [13, 43, 18]. For example, 2D slices were proposed as supervision to predict full object segmentation [13]. Point annotations were also adapted to reduce human annotation costs [43]. 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 [32]. Several unsupervised learning methods were based on learning anatomical prior [14] or training adversarial networks [23]. 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. Our Weakly Supervised Setting

In this section, we introduce image-level supervision for our weakly-supervised setting.

Among weak labels for 3D volumetric images, image-level annotation is the cheapest way. Image-level class labels refer to the classes that appeared in our input samples. Although researchers developed many successful approaches on 2D weakly supervised segmentation using image-level labels [6, 5, 51], there are three major challenges for using image-level class labels in 3D weakly supervised segmentation compared to 2D images: (1) 3D data are reconstructed from multiple 2D image sequences, which usually contain much more information than a single image. Thus, a single label for a 3D volumetric image is considerably coarse. (2) 3D biomedical data, especially for cryo-ET, have heavy noise due to imaging limits, which is not common for 2D images in the wild [35, 64]. (3) Some objects represented in 3D spaces usually have visible holes, while most 2D objects are solid. Therefore, state-of-the-art 2D models in weakly-supervised segmentation are not suitable for 3D volumetric images.

4. Method

4.1. 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 relations 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 2 for a high-level summary of the model, and the sections below for more details.

4.2. Initial Seed Generation

Following previous weakly-supervised methods [54, 16, 22], we choose the CAM-based method to generate initial localization clues on 3D volumetric data. We use the Grad-CAM [44] with a 3D convolutional neural network as the model backbone. Grad-CAM plays three important 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 4.3. Third, Grad-CAM is used to produce image-level class labels during model inference.

We first train a classification network using the image-level labels and then obtain the pseudo segmentation label for certain classes via Grad-CAM. Specifically, given an image, in order to obtain the localization map $G^c \in \mathbb{R}^{T\times U\times V}$ of depth T, width U, and height V for class c, we compute the gradient of the score for class c, y^c (layer before the softmax), with respect to feature map activations A^m. These gradients are average-pooled over the width, height and depth dimensions, which is denoted as:

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

Then a weighted combination of activation maps along with a ReLU layer are used to obtain the initial localization map:

$$G_s^c = \text{ReLU}\left(\sum_m \alpha_m^c A^m\right). \tag{2}$$

Then we perform spline interpolation [20] 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.

4.3. Cross-Image Co-Occurrence Learning

Unlike most of the existing weakly-supervised methods which learned from independent images [6, 5, 53], 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 [60] in a high-dimensional space with a learned feature encoder. This feature space represents the common patterns of the image group that shares the same class label. Then the model 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:

$$\hat{e} = \text{Softmax}\left(\sum_{n=1}^N e_n\right). \tag{3}$$

$\hat{e}$ 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 4.2 by:

$$M_{ijk}^c = w_1 G_{ijk}^c + w_2 P_{ijk}^c, \tag{4}$$

where M_{ijk}^c is the voxel-level element in the merged localization map M^c. Note that we apply rank normalization [49] to G^c and P^c before the combination.

4.4. 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 [54, 55]. However, as

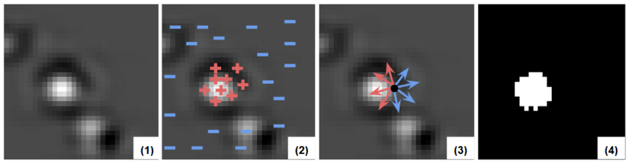

Figure 3: Semantic segmentation under the supervision of voxel affinity pairs. The figure shows: (1) the input biomedical image with heavy noise; (2) pseudo labels generated by localization map M^c; (3) voxel affinity pairs (S$^-$) sampled from M^c; (4) semantic segmentation generated by *CIVA-Net*.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

the pseudo labels are not accurate, especially at the object boundaries, the model may not be able to learn from those inaccurate labels in an ordinary full segmentation manner. Inspired by [5], 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 4.3. 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 [5], 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 voxel pairs from the pseudo class-map $\hat{M}$, and categorize them into two sets $\mathcal{S}^-$ and $\mathcal{S}^+_{bg}$:

$$\mathcal{S} = \{(p,q) \mid \|\mathbf{x}_p - \mathbf{x}_q\| < \gamma, \forall p \neq q\}, \tag{5}$$

$$\mathcal{S}^- = \{(p,q) \mid \hat{M}(\mathbf{x}_p) \neq \hat{M}(\mathbf{x}_q), (p,q) \in \mathcal{S}\}, \tag{6}$$

$$\mathcal{S}^+_{bg} = \{(p,q) \mid \hat{M}(\mathbf{x}_p) = \hat{M}(\mathbf{x}_q) = \mathbf{0}, (p,q) \in \mathcal{S}\}, \tag{7}$$

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 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 2. It uses VoxResNet [11] as the backbone network. Similar to the network structure used in [5], 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. Different from [5] which first predicted class boundaries and then used a separate propagation step to obtain the segmentation, our model directly predicts the object segmentation $\mathcal{O} \in [0,1]^{D\times H\times W}$ and can be optimized in an end-to-end manner. 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. 6 must contain at least one foreground voxel (denotes as 1); any path between positive pairs in Eqn. 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:

$$a_{pq} = 1 - \max_{k\in\Pi_{pq}} \mathcal{O}(\mathbf{x}_k), \tag{8}$$

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

The learning of a_{pq} is supervised by the sampled voxel affinity pairs, which is equivalent to minimize the cross-entropy between the one-hot vector of the binary affinity and the predicted affinity:

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

4.5. Training

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

$$\mathcal{L}_{\mathcal{B}} = \sum_{i=1}^{N} \text{CE}(cls^i, cls^{*i}), \tag{9}$$

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 4.4. The final loss of our proposed approach is calculated using:

$$\mathcal{L} = \mathcal{L}_{\mathcal{B}} + \mathcal{L}_{\mathcal{O}}. \tag{10}$$

4.6. 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 [29] with 3D counterparts to refine the segmentation results.

5. 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.

5.1. Dataset

We follow common practice in cryo-ET analysis to evaluate our method on subtomograms [58, 59]. 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 [15], which takes into account the tomographic reconstruction process with missing wedges and contrast transfer function [37, 38]. 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 [59, 33], we resize each subtomogram to 32^3 due to GPU memory constraints. 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 [19]. 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.

5.2. Evaluation Metrics

Following prior work [52, 4], 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.

5.3. Baseline Methods

Image-level Baselines. Following existing work [53], CAM-based methods are chosen as image-level baselines when there is no existing literature on 3D segmentation using image-level supervision. We choose Grad-CAM [44] and Respond-CAM [61] 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.

5.4. Implementation Details

Grad-CAM and Respond-CAM use the same network structure in [61] and share the same hyper-parameter settings and training configurations. The models are trained with a learning rate of 0.001. Adam [27] 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 [26] 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 [45]. Adam [26] is used as the optimizer with batch size 16. The model converges after about 200 epochs.

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

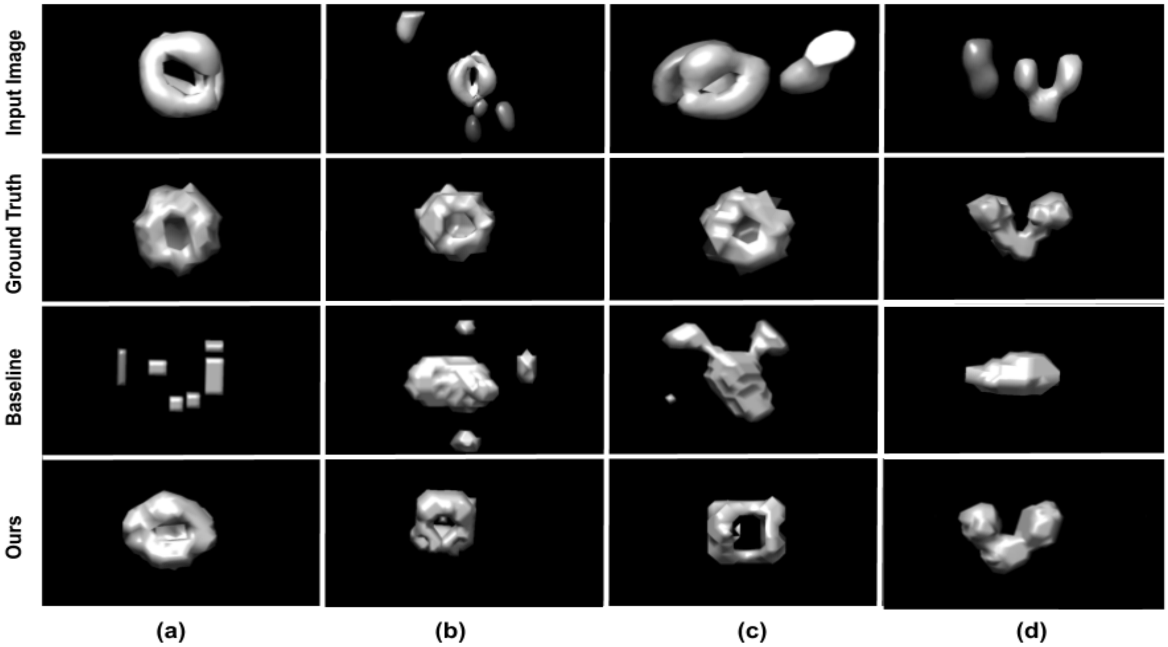

Figure 4: Visualization of the semantic segmentation results. We use Grad-CAM as the visualization baseline.

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

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 [39]. 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.

Method	$\mathrm{mIoU}^{003}_{\mathrm{snr}}$	$\mathrm{mIoU}^{005}_{\mathrm{snr}}$	$\mathrm{mIoU}_{\mathrm{real}}$
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: 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.

5.5. Quantitative Results

Comparison to Image-Level Baselines. Table 1 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 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

Grad-CAM	CO	IVA	3D-CRF	$\mathrm{mIoU}_{\mathrm{snr}}^{003}$	$\mathrm{mIoU}_{\mathrm{snr}}^{005}$	$\mathrm{mIoU}_{\mathrm{real}}$
✓				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 4: Performance of ablated versions of our model.

classes and predict precise segmentation. 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, 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 [53], 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.

5.6. Qualitative Analysis

We qualitatively demonstrate the advantages of our model in Figure 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.

5.7. 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 4. We observe that each component of our model gains consistent improvements on all datasets.

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 [11]. The results are reported in Table 5. The first row shows the mIoU of the pseudo segmentation labels. The second row shows the performance of 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.

Setting	$\mathrm{mIoU}_{\mathrm{snr}}^{003}$	$\mathrm{mIoU}_{\mathrm{snr}}^{005}$	$\mathrm{mIoU}_{\mathrm{real}}$
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 5: Ablation Study of Inter-Voxel Affinity Learning.

6. 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. Currently, our model is validated on low-resolution cryo-ET images. In the future, more work can be done to extend *CIVA-Net* to large-size input and different data modalities. 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.

Acknowledgment

This work was supported in part by U.S. NIH grants R01GM134020 and P41GM103712, NSF grants DBI-1949629 and IIS-2007595, and Mark Foundation For Cancer Research 19-044-ASP. We thank the computational resources support from AMD COVID-19 HPC Fund. Xiangrui Zeng was supported in part by a fellowship from Carnegie Mellon University Center for Machine Learning and Health. We would like to thank the anonymous reviewers for their useful comments, and thank Gregory Howe for proof-reading the paper.

References

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

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[14] 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.

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

[16] 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.

[17] 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.

[18] Pierre-Antoine Ganaye, Michaël Sdika, and Hugues Benoit-Cattin. *Semi-supervised Learning for Segmentation Under Semantic Constraint: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III*, pages 595–602. 09 2018.

[19] 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.

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

[21] 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.

[22] 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.

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

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

[25] 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.

[26] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[28] 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.

[29] 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.

[30] Werner Kühlbrandt. The resolution revolution. *Science*, 343(6178):1443–1444, 2014.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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, 2015.

[36] 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.

[37] Sinuo Liu, Xiaojuan Ban, Xiangrui Zeng, Fengnian Zhao, Yuan Gao, Wenjie Wu, Hongpan Zhang, Feiyang Chen, Thomas Hall, Xin Gao, et al. A unified framework for packing deformable and non-deformable subcellular structures in crowded cryo-electron tomogram simulation. *BMC bioinformatics*, 21(1):1–24, 2020.

[38] Sinuo Liu, Yan Ma, Xiaojuan Ban, Xiangrui Zeng, Vamsi Nallapareddy, Ajinkya Chaudhari, and Min Xu. Efficient cryo-electron tomogram simulation of macromolecular crowding with application to sars-cov-2. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 80–87. IEEE, 2020.

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

[40] 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.

[41] 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.

[42] 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*, pages 5099–5108, 2017.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mecha-

nism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[52] 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.

[53] 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, 2020.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

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

[59] 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.

[60] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. *arXiv preprint arXiv:2004.13364*, 2020.

[61] 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.

[62] 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.

[63] 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.

[64] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[65] Xiaoyu Zhu, Junwei Liang, and Alexander Hauptmann. Msnet: A multilevel instance segmentation network for natural disaster damage assessment in aerial videos. In *Proceedings of Winter Conference on Applications of Computer Vision*, 2021.

[66] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation, 2016.