Weakly Supervised Learning for Joint Image Denoising and Protein Localization in Cryo-Electron Microscopy

Qinwen Huang, Ye Zhou, Hsuan-Fu Liu, Alberto Bartesaghi
Duke University
{qinwen.huang, ye.zhou867, hl325, alberto.bartesaghi}@duke.edu

Abstract

Deep learning-based object detection methods have shown promising results in various fields ranging from autonomous driving to video surveillance where input images have relatively high signal-to-noise ratios (SNR). On low SNR images such as biological electron microscopy (EM) data, however, the performance of these algorithms is significantly lower. Moreover, biological data typically lacks standardized annotations further complicating the training of detection algorithms. Accurate identification of proteins from EM images is a critical task, as the detected positions serve as inputs for the downstream 3D structure determination process. To overcome the low SNR and lack of image annotations, we propose a joint weakly-supervised learning framework that performs image denoising while detecting objects of interest. By leveraging per-pixel soft segmentation and consistency regularization, our framework denoises images without the need of clean images and is able to detect particles of interest even when less than 0.5% *of the data are labeled. We validate our approach on real single-particle cryo-EM and cryo-electron tomography (ET) images which are known to suffer from extremely low SNR, and show that our strategy outperforms existing state-of-the-art (SofA) methods used in the cryo-EM field by a significant margin. We also evaluate the performance of our algorithm under decreasing SNR conditions and show that our method is more robust to noise than competing methods.*

1. Introduction

Deep learning based-algorithms for object detection have witnessed a dramatic improvement over the past few years. Given sufficient amounts of data, a network can easily learn to identify different subjects in images or perform tracking in video sequences. Most of these applications, however, rely on the availability of images that have relatively high signal-to-noise ratios (SNR). Cryo-electron microscopy (EM) is a popular technique for structure determination that can produce 3D reconstructions of proteins by back-projecting a large number of 2D protein projections taken from different orientations. This requires the detection of individual molecular images from large electron micrographs, a process commonly referred to as *particle picking*. The low SNR typical of cryo-EM images is caused by the limited electron doses used during acquisition to prevent radiation damage, making the detection problem very challenging. Recent efforts to tackle particle picking have focused on either improving SNR using image denoising or other pre-processing strategies, followed by automatic or semi-automatic detection algorithms. Under very low SNR conditions, however, the performance of these algorithms is sub-optimal and can result in many missed particles, thus limiting the quality of the downstream 3D reconstructions.

In this paper, we propose a framework that performs image denoising and particle segmentation and identification simultaneously. The information learned from these tasks is complementary and therefore by enabling information sharing, we are able to improve the performance of both tasks. Since noiseless images do not exist in cryo-EM and labeling is very time consuming (a normal dataset usually contains up to millions of particles), we adopt a strategy that does not require any information on clean images and learns to segment particles when only the center pixels of a small fraction of particles are labeled. We validate our approach on three challenging cryo-EM datasets: one from single particle cryo-EM and two from cryo-ET. We show that under increasingly challenging SNR conditions, our method is able to outperform the state-of-the-art by a significant margin. To our knowledge, this is the first example of a method that is able to perform both image denoising and particle segmentation and detection at the same time without the need of ground-truth clean images for denoising and per-pixel labeling for segmentation.

The paper is organized as follows: we present related work in Section 2 and describe our proposed method in Section 3. In Section 3.1, we first present the theoretical framework for joint learning and then detail how we transform our hypothesis into a neural network setting that learns to denoise and segment without clean images or per-pixel labeling. In Section 3.2, we give implementation details of our approach, and in Section 4 we present experimental results on three real cryo-EM datasets and compare the performance of our approach against two commonly used particle picking methods.

2. Related Work

Recent developments in deep learning have led to breakthrough performance in tasks such as image enhancement, object detection and segmentation. In this section, we introduce relevant recent work, including denoising without clean images, semi-supervised object detection and weakly supervised segmentation, and multi-task learning.

2.1. Denoising without clean images

Unlike denoising algorithms based on supervised learning that are trained using noisy-clean image pairs, blind image denoising is usually achieved by leveraging internal data statistics. Traditional methods based on internal statistics include Non-Local Means (NLM) which predicts clean pixel values based on similar local neighborhoods [7], and Block-Matching 3D (BM3D) which similarly relies on data repetitiveness [8]. More recent denoising methods based on convolutional neural networks include: Deep Image Prior which trains a neural network to learn the prior distribution of data from pure noise [46], Noise2Noise that learns to denoise using pairs of independently corrupted training images that share the same underlying signal [25], and Noise2Void that assumes independence of noise corruption for each pixel and trains the denoising network only using the single input noisy image by masking the central pixel [22]. Built on Noise2Void, a more generalized formulation was proposed in [2] and was further improved by incorporating Bayesian statistics in [23]. In addition, a number of related methods have been proposed in the literature [24, 16, 35, 17, 34]. For the case of cryo-EM images, implementations of Noise2Noise have been successfully applied to low SNR biological samples in [3, 31].

2.2. Semi-supervised object detection and weakly supervised segmentation

The majority of existing semi-supervised object detection methods are either built upon one-stage detectors [39, 29] or two-stage detectors [12, 40]. Most of the semi-supervised learning frameworks incorporate the use of unlabeled data through the use of consistency regularization [51, 54, 44], along with supervised learning of labeled data. Self-supervised sample mining stitches high confidence unlabeled data patches to labeled data and maximizes the consistency of proposed regions [49]. Data augmentation is widely applied in consistency learning, including approaches that maximize consistency between detection and classification outputs of labeled/unlabeled images and their augmented pairs [18], and algorithms that use labeled data to first train a teacher model and then update the model by maximizing the consistency between output pseudo-labels of unlabeled data and strongly augmented pairs [45]. In the cryo-EM field, several deep learning-based object detection methods [4, 48, 1, 47] have been used successfully for particle detection. Topaz [4] is a semi-supervised particle picking method based on positive-unlabeled learning. crYOLO [47] is a fully supervised picking method built upon the popular YOLO model [38]

Unlike fully supervised segmentation methods which use per-pixel supervision, weakly supervised segmentation usually uses weak labels such as bounding box annotations [9, 36, 30], scribbles [26], or image-level labels [21, 10, 33, 50]. Segmentation using image-level labels usually involves training a classifier from which class saliency map or class activation map (CAM) [53] can be obtained. For instance, [10] uses a conditional random field (CRF) to post-process the saliency map for segmentation, and [50] utilizes equivariance constraints to generate a CAM.

2.3. Multi-task learning

Multi-task learning (MTL) which leverages information shared between related tasks to improve the performance of the original task has been widely applied in various image processing tasks [42, 19, 11, 52, 28, 37, 27]. Specific applications of MTL in object detection include Mask-RCNN which simultaneously performs image segmentation and detection [14], joint detection of objects while estimating distance between them [6], learning of segmentation maps as attention to aid detection [32], denoising, segmentation and detection through a cascaded network in a supervised manner [13], and denoising and segmentation of florescence microscopy images [5]. All of these strategies, however, operate on images with relatively high SNRs. Building on the success of these methods, here, we extend their applicability to lower SNR datasets such as those routinely used in cryo-EM.

3. Proposed Method

The overall architecture of our framework is shown in Figure 1. The cascaded network, which includes image denoising and per-pixel labeling, is able to leverage complementary information from these two tasks leading to mutual improvements. The denoising branch estimates the mean and covariance of the underlying noiseless data distribution from noisy inputs and feeds the estimated data statistics into the detection branch. The detection branch identifies particle locations by performing pixel-wise segmentation of the input and in return improves the denoising output as segmentation accuracy increases. Segmentation is performed

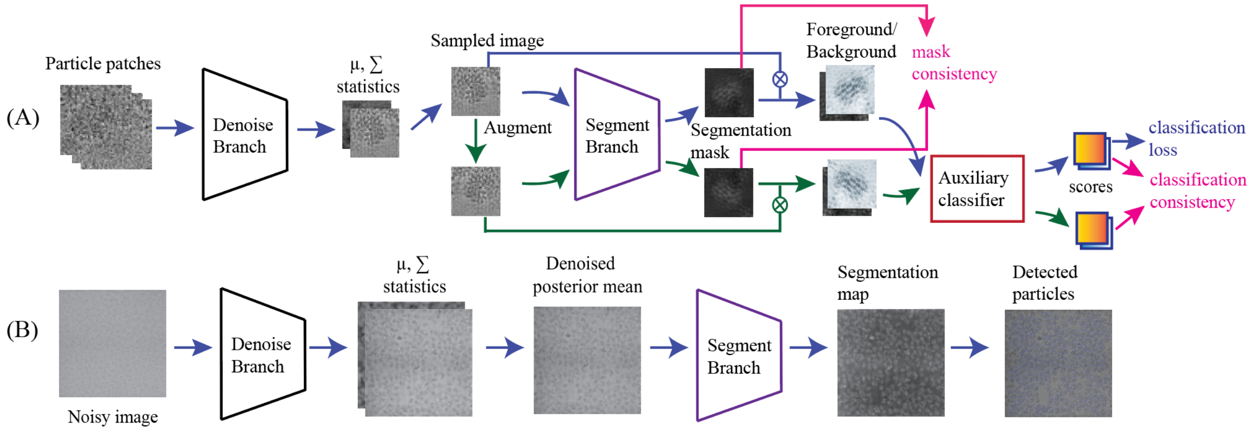

Figure 1: **Overall architecture of our framework for simultaneous image denoising and detection.** A) Network training branch: patches that contain particles are cropped from noisy images serving as input to the network. The denoising branch first outputs relative statistics of the estimated clean image. The desnoised sampled image and its augmented pair are fed into the segmentation branch and a segmentation mask is produced. An auxiliary classifier is used to classify foreground/background images based on the segmentation mask. Segmentation is guided by the classifier. Consistency regularization using augmented pair has two components: mask and classification consistency. B) Network evaluation branch: during inference, the entire image is fed into the network. The network outputs the denoised posterior mean and a segmentation mask of the entire image. The segmentation mask is used to identify particle locations.

in a weakly supervised fashion: only the center location of the particle is provided. We hypothesize that a particle-containing patch can be separated into foreground (which contains the particle) and background (which excludes the particle). To guide the segmentation, we add a classifier to the segmentation module which learns to discriminate between foreground and background based on the assumption that accurate segmentation leads to the correct classification of images. The framework eventually predicts the location of particles based on the segmentation results. We describe the training procedure and details of the framework in the following sections.

3.1. Joint denoising and detection

Consider the prediction of the clean value x and its corresponding label l for a noisy pixel y. As pixels in an image are not independent, we assume that the clean value depends not only on the noisy measurement y, but also on the neighboring context Ω_y. We also assume that the label l of the pixel depends only on its clean value x. From this, performing denoising and detection jointly can be thought of as statistical inference on the probability distribution $p(x, l|y, \Omega_y)$ over the clean pixel value x and its label l, conditioned on the noisy input y and its context Ω_y. In cryo-EM applications, the noise is usually modeled as a Gaussian distribution [43]. We therefore bring in this extra information on the noise corruption so that $p(y|x)$ can be modeled explicitly. With this, we can connect the observed marginal distribution of the noisy labeled training data to the unobserved distribution of the clean data:

$$\begin{aligned} \underbrace{p(y,l|\Omega_y)}_{\text{training data}} &= \int p(y,l,|\Omega_y,x)p(x|\Omega_y)dx \\ &= \int \underbrace{p(y|x)}_{\text{noise model}} \underbrace{p(l|x)}_{\text{label model}} \underbrace{p(x|\Omega_y)}_{\text{unobserved}} dx. \end{aligned} \tag{1}$$

This relationship suggests that even though we only observe corrupted training data, we are able to use the known noise model for the prediction of $p(x|\Omega_y)$ and its label based on the prediction $p(l|x)$. Specifically, we can model $p(x|\Omega_y)$ as a multivariate Gaussian $\mathcal{N}(\mu_x, \Sigma_x)$. Following this assumption, we can train a network to map the context Ω_y to the mean μ_x and convariance Σ_x, and subsequently map the estimated statistics to the label l by maximizing the likelihood under Equation (1).

Information of the noisy measurement y can be included through Bayesian reasoning. Specifically, the posterior probability of the clean value x and its label l, given the noisy y, its surrounding Ω_y, is:

$$\begin{aligned}\underbrace{p(x,l|y,\Omega_y)}_{\text{posterior}} &\propto \underbrace{p(l|x,\Omega_y,y)}_{\text{predicted label}}\ \underbrace{p(x,\Omega_y,y)}_{\text{joint signal model}} \\ &\propto \underbrace{p(y|x)}_{\text{noise model}}\ \underbrace{p(x|\Omega_y)}_{\text{prior}}\ \underbrace{p(l|x)}_{\text{label model}} .\end{aligned} \tag{2}$$

From this point of view, the prior distribution $p(x|\Omega_y)$ encodes our belief of x based on the neighborhood information before observing y. In addition, our belief of x is directly related to its label l. A more accurate belief in x leads to a better estimation of its label l. We therefore perform joint learning through maximization of this posterior distribution. Specifically, the denoising branch of our proposed framework corresponds to the first two terms in Equation (2) and the detection branch corresponds to the label model term.

Self-Supervised Denoising Branch. The denoising branch is based upon a blindspot convolutional neural network proposed in [23] that learns the underlying clean signal by maximizing the posterior likelihood in Equation (2), which includes maximization of the observed noise model $p(y|x)$ subject to prior belief $p(x|\Omega_y) \sim \mathcal{N}(\mu_x, \Sigma_x)$ and its label (we discuss segmentation in the following section). Assuming that y is corrupted by zero-mean Gaussian noise, we have $p(y|x) \sim \mathcal{N}(\mu_y, \Sigma_y)$, where $\mu_y = \mu_x$ and $\Sigma_y = \Sigma_x + \sigma^2 \mathbf{I}$, with σ^2 being the noise variance. Since maximizing $p(y|x)$ is equivalent to minimizing its negative log-likelihood, we can write the denoise loss as:

$$L_{dn} = \frac{1}{2}[(y-\mu_y)^T \Sigma_y^{-1}(y-\mu_y)] + \frac{1}{2}\log|\Sigma_y| + C, \tag{3}$$

where C is a constant that can be discarded. Since σ^2 is unknown, an auxiliary network is used to estimate its value. The output of the denoising branch consists of μ_x and Σ_x. The posterior distribution for x is calculated by multiplying the noise model and the prior Gaussian distribution parameterized by μ_x and Σ_x, which also follows a Gaussian distribution with $E(p(x|y,\Omega_y) = (\Sigma_x^{-1} + \sigma^{-2}\mathbf{I})^{-1}(\Sigma_x^{-1}\mu_x + \sigma^{-2}y)$. The mean of the posterior distribution is the final denoised output.

Weakly-Supervised Detection Branch. As the outputs of the denoising branch are the mean μ_x and covariance Σ_x of the prior belief $p(x|\Omega_y)$, to obtain a tractable approximation of $p(x|\Omega_y)$, input to the detection branch is sampled from this prior. In order to backpropagate the gradient through Σ_x, we adopt the re-parameterization trick proposed in [20] such that the sampled image $x_I = \mu_x + \Sigma_x \odot \mathcal{N}(0, I)$ where $\odot$ is the element-wise multiplication. The detection branch models $p(l|x)$, which assigns a label based on each pixel value. Therefore, the detection branch detects particles by doing binary segmentation. For a sampled input image I containing particles, the detection branch outputs a saliency map, M, which segments the image into two regions: particle (foreground) and background. To simulate binary hard thresholding while preserving differentiability, we add a modified sigmoid layer to the output saliency map:

$$\tilde{M} = \frac{1}{1+\exp[-C(M-t)]}, \tag{4}$$

where C is a constant and t is a threshold value. Segmentation is guided by an auxiliary classifier. If the image is segmented correctly, the classifier will be able to classify segmented images into their corresponding category. To do this, we multiply the sampled image I by $\tilde{M}$ and its complement to get two segmented images:

$$F = \tilde{M}I, \quad B = (1-\tilde{M})I, \tag{5}$$

where F represents the foreground containing the particles and B represents the background. Both F and B are fed into the classifier g and the classifier outputs the probability of the input containing a particle. We adopt the hinge loss to train this classifier:

$$\begin{aligned} L_f &= \min[0, -1+g(F)] \\ L_b &= \min[0, -1-g(B)], \end{aligned} \tag{6}$$

where L_f is the loss for the foreground and L_b is the loss for the background. We also incorporate consistency constraints to further regularize the detection branch. For each I, we randomly apply horizontal and vertical flipping to generate its augmented pair $A(I)$, where $A(\cdot)$ denotes the applied transformation. The augmented image is fed into the same network and the network outputs the segmentation map $M_{A(I)}$ and the classification probabilities $g(F_{A(I)}))$ and $g(B_{A(I)})$. We assume equivariance in segmentation, i.e. $M_{A(I)} = A(M_I)$, and rotation invariance in classification, i.e. $g(F_{A(I)})) = g(F_I)$ and $g(B_{A(I)}) = g(B_I)$. Therefore, to impose consistency regularization, we define the consistency loss as:

$$\begin{aligned} L_{cons} = \left\|M_{A(I)}, A(M_I)\right\|_2^2 + \left\|g(F_{A(I)})), g(F_I)\right\|_2^2 \\ + \left\|g(B_{A(I)}), g(B_I)\right\|_2^2, \end{aligned} \tag{7}$$

where $\|\cdot\|_2^2$ denotes the squared L2 norm error.

In summary, the final loss function for our joint training framework is defined as:

$$L = \alpha L_{dn} + (1-\alpha)(L_f + L_b) + \lambda L_{cons}, \tag{8}$$

where α represents the assigned weights for each task and λ is the weight for the consistency regularization term. L_{dn} is used to denoise the input noisy images, L_f and L_b are used to guide the segmentation, and L_{cons} is used to further refine the segmentation result.

3.2. Implementation

We now provide details about the architecture of the different components of our network and its training procedure. We adopt a U-Net [41] based structure for the denoising branch that uses a shifted convolutional layer as in [23] instead of a normal convolutional layer. The segmentation module of our detection branch is composed of four convolutional layers with two upsampling layers in between and a final max-pooling layer. The auxiliary classifier is composed of three residual blocks [15] and a final convolutional layer. The estimation of the noise variance is also performed using a normal U-Net architecture. During training, only patches that contain particles of interest are fed into the framework. Typically, particles are located at the center of the patch. We use 64×64 for the patch size. The detection branch outputs a segmentation map. Foreground and background images are generated based on the segmentation map and the classifier outputs the probability of a foreground/background image containing a particle. For the modified sigmoid layer in Equation (4), we use $C = 7$ and $t = 0.5$. During the inference process, the entire image is fed into the framework. The auxiliary classifier is removed and the segmentation map of the entire image is used to identify particle locations. The center location for each particle is obtained by applying non-max suppression to the segmentation map. The model is trained on a single NVIDIA Tesla V100 GPU with 32G of RAM. We use a batch size of 32 and a cosine decay learning rate for the scheduler with initial learning rate of 0.001. We adopt the ADAM optimizer and use $\alpha = 0.75$ and $\lambda = 0.1$ for all experiments. Training with $240,000$ iterations takes less than an hour. Inference on a single image takes less than a second.

4. Experiments

4.1. Datasets

In this section, we evaluate our methods on real cryo-EM images of ribosomes, including one single-particle dataset and two cryo-ET datasets available from the Electron Microscopy Public Image Archive (EMPIAR), EMPIAR-10304 and EMPIAR-10499. For the single-particle dataset, 2D projections of the 3D protein sample are collected in movie format, consisting of multiple dose fractionated frames. Particle picking in this case is normally performed on the average of all the frames. In cryo-ET datasets, 2D projections of a 3D protein sample are collected at different rotation angles. These images taken at various projection angles are called tilt series. Tilt series have much lower SNR than single-particle frame averages making the task of particle detection more challenging. Among all three datasets, single particle ribosome has the highest SNR, and EMPIAR 10499 has the lowest SNR. All images are single-channel gray-scale. All datasets are only partially labeled. For each labeled particle, its center location (x, y coordinates) is annotated. Note that we do not include bounding box size in the annotation.

Single-particle cryo-EM ribosome dataset. This dataset contains 1000 movies, each cropped to size 4096×4096, with defocus values ranging from $0.8\,\mu\text{m}$ to $3.0\,\mu\text{m}$. The pixel size is $1.08\,\text{Å}$. Each movie contains 60 frames. Since the ribosome is a relatively large particle, frame averages of this dataset have relatively high SNR compared to most other proteins. We therefore treat the average of all 60 frames as the ground-truth images. Similarly, particles picked on these frame averages are treated as the ground-truth annotations. To simulate increasingly lower SNR conditions, such as those observed for lower molecular weight proteins and lower defocus datasets, we use partial averages calculated from 10%, 20% and $60\%$and of the total number of frames (6, 12, 36 frames). Among these 1000 low SNR partial frame averages, 16 images are used as the training set and the remaining ones are used for testing. All images are further down-sampled by a factor of 8 to size 512×512. Training images are only partially labeled, with 15 to 25 particles identified on each. The entire training set is composed of 500 labeled particles from 16 images, which accounts for around 0.04% of the total number of particles in the entire dataset.

EMPIAR-10304. This dataset consists of 12 tilt-series from a sample of purified ribosomes. Each tilt series is composed of 41 projection images ranging from -60 degree to + 60 degree. A single tilt image has size of 4096×5760, with a pixel size of $2.1\,\text{Å}$. We evaluate our framework on the zero-degree tilt images of each tilt series. From the total of 12 zero-degree tilt images, 3 images are used for training and the remaining ones are the testing set. We also down-sampled each image by a factor of 8 to size 512×720. Training images are partially labeled as well, with 40 to 60 particles identified on each. The entire training set is composed of 200 labeled particles, which accounts for around 4% of all particles in the entire dataset. We use manually labeled particle locations as ground truth.

EMPIAR-10499. This is a cryo-ET dataset of ribosomes imaged within cells. This dataset is challenging because particles are observed within a crowded context that includes cell membranes and other sub-cellular components. We use a subset of the entire dataset which consists of 65 tilt series. Each tilt series is composed of 41 projection images ranging from -60 degree to +60 degree. A single tilt image has size of 3838×3710, with pixel size of $1.7\,\text{Å}$. We also evaluate our framework on the zero-degree tilt images. Of the 65 tilt images, 7 are used for training. Each image is

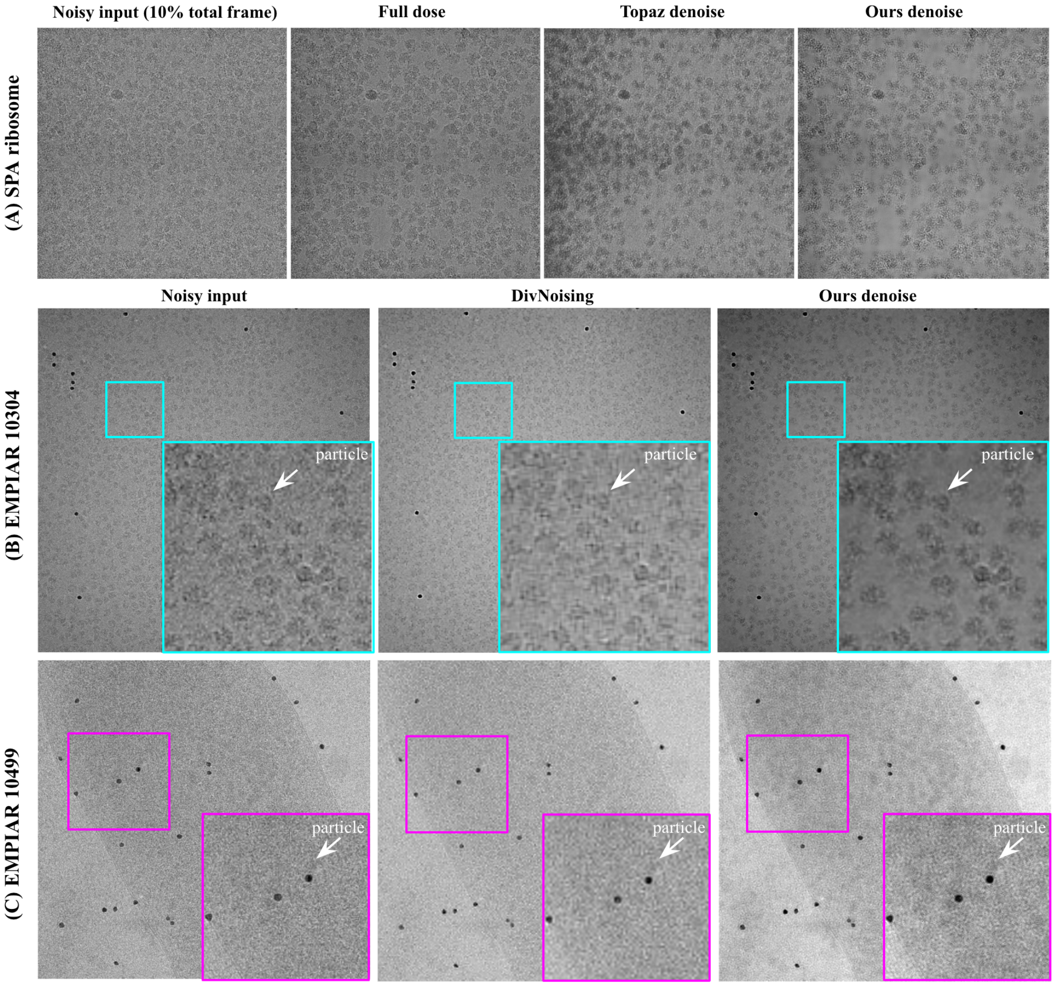

Figure 2: **Visualization of improvement in denoising performance on three real cryo-EM datasets compared to Topaz denoise and DivNoising.** A) single particle ribosome image denoised using topaz and our method. B) EMPIAR-10304 image denoised using DivNoising and our method. C) EMPIAR-10499 denoised using DivNoising and our method. Our method helps to visualize particles better, especially for zoomed in views of 10304 and 10499.

down-sampled by a factor of 8 to 480×464. Partial-labeling is performed and a total of 90 particles are identified, which accounts for less than 1% of all particles in the dataset. We use manually labeled particle locations as ground truth.

4.2. Results

We use PSNR values to evaluate denoising performance. We only calculate PSNR values for the single-particle ribosome dataset as we are able to treat full dose frame averages as ground truth. As full dosage micrographs can still be noisy, we apply a low-pass filter to remove potential high-frequency noise. The average PSNR values for the dataset are shown in Table 1. PSNR1 is calculated against the full dose micrographs and PSNR2 is calculated against low pass-filtered ones. Since low-pass filtering introduces image blurring, we expect the actual PSNR value to lie in between the two reported values. Our joint learning framework is able to achieve significantly higher PSNR val-

Method	PSNR1	PSNR2
Topaz denoise	16.79	19.49
Ours w/o segmentation	19.18	19.33
Ours joint	**19.78**	**21.62**

Table 1: **Denoising performance on single-particle cryo-EM images of ribosomes**. Comparison of PSNR values between topaz, our approach w/o segmentation, and new joint strategy. PSNR1 is obtained using full dose as a reference ans PSNR2 is obtained using low pass-filtered full dose as reference.

Table 2: Detection performance measured on averages using 10% of frames compared against particles detected in full dose micrographs.

		Topaz Pick w/o Denoise	*Topaz Denoise + Pick*	*CrYOLO*	*Ours w/o Consistency*	*Ours w/ Consistency*
single-particle ribosome	Precision	0.615	0.656	0.501	0.681	**0.735**
	Recall	0.525	0.645	0.211	0.702	**0.825**
	F_1	0.567	0.648	0.290	0.693	**0.778**
EMPIAR-10304	Precision	0.56	N/A	0.259	0.605	**0.618**
	Recall	0.448	N/A	0.146	0.443	**0.683**
	F_1	0.497	N/A	0.184	0.510	**0.648**
EMPIAR-10499	Precision	0.289	N/A	0.102	0.216	**0.356**
	Recall	0.217	N/A	0.016	0.235	**0.326**
	F_1	0.248	N/A	0.026	0.225	**0.341**

ues. For the other two cryo-ET datasets, we are not able to provide quantitative measurements of performance because clean images do not exist. We therefore provide visualization of denoised images obtained using our approach. In addition, since Topaz requires pairs of noisy images, we compare the performance of our method against DivNoising [34][1], which is a self-supervised denoising method that has been proven to perform well on high-contrast fluorescence microscopy images. More qualitative visualizations are provided in the supplementary material.

Even though we perform particle detection through segmentation, our main focus is on the detection task, not on segmentation. In addition, since the bounding box for each particle used for downstream processing is usually much larger than the particle size, we do not use intersection over union (IoU) as evaluation criteria. Instead, we calculate precision, recall and F1 values. True positive, false negative and false positive values are obtained by comparing against particles detected on full-dose images (single-particle ribosome dataset) and manually labeled particles (EMPIAR-10304 and 10499). To account for small variations in the detected particle centers, instead of looking at a single pixel, we also look at pixels located within a certain radius from the center. If the detected particle position is within a certain radius of a ground truth particle position, we consider it as a true positive match. Similarly, if there is no ground truth particle within a certain radius of a detected particle position, we consider it as a false positive. We use a radius of 5. With this, precision and recall scores are calculated as:

$$\begin{aligned} \text{Precision} &= \frac{\text{\# of TP/matches}}{\text{\# of predicted particles}} \\ \text{Recall} &= \frac{\text{\# of TP/matches}}{\text{\# of target particles}} \end{aligned} \tag{9}$$

Results are presented in Table 2. We compare the performance of our method against two of the most commonly used particle picking methods in cryo-EM: topaz [4][2] and crYOLO [47][3]. Topaz is a semi-supervised particle picking method and crYOLO uses a fully supervised approach. For both Topaz and crYOLO, we use the code provided by the authors and we adjusted hyper-parameters to get the best possible results. Training of Topaz is performed using the same training set as our approach. crYOLO has a generalized pre-trained model and we further fine-tune it with the same training set. For all three datasets, our joint framework is able to obtain higher precision, recall and F1 values than topaz and crYOLO. We also perform an ablation study on the effectiveness of consistency regularization and show that our framework is able to perform significantly better with the addition of the consistency loss term. Notably, the improvement is more significant under lower SNR conditions. Visualization of our particle picking results is presented in Figure 3. Since detection is based on the segmentation output, we also show the soft segmentation map (before the modified sigmoid layer). Segmentation maps show a clear distinction between particles of interest and the background. The background/foreground contrast is less obvious on EMPIAR-10499 as this dataset has the lowest SNR among all three. However, regions with particles still have brighter values than the surrounding background, which enabled us to perform the subsequent particle-picking step. In addition, regions corresponding to contamination and gold beads (small black circles) have either extremely high or low pixel values, which allow us to avoid these regions during the picking step. Our method shows significant improvement in picking performance under low SNR conditions when compared to current state-of-the-art methods (Topaz and crYOLO). We provide more examples and choice of parameters in supplementary mate-

[1] https://github.com/juglab/DivNoising

[2] https://github.com/tbepler/topaz

[3] https://cryolo.readthedocs.io/en/stable/installation.html

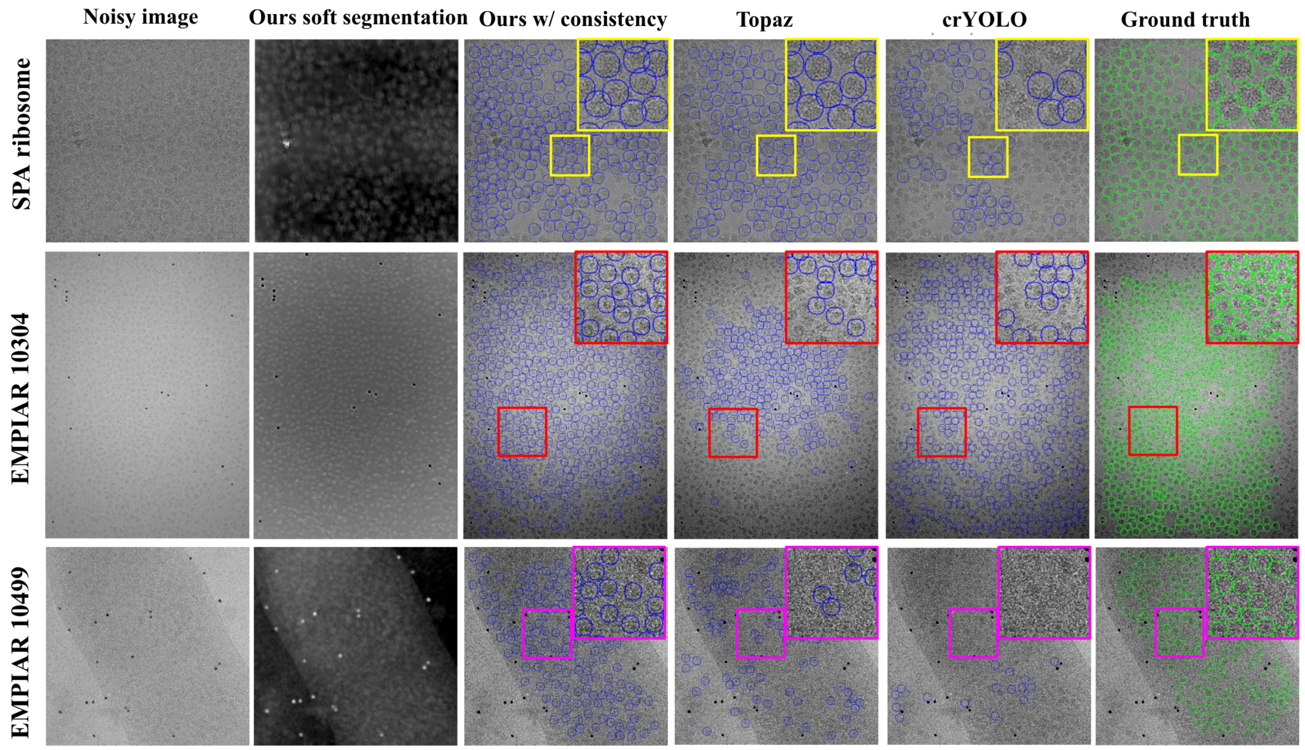

Figure 3: **Visualization of particle detection results on the three datasets**. Detected particles are circled in blue. The first row is an example of detection results for the single-particle dataset. The second row is the result for EMPIAR-10304 and the last row is for EMPIAR-10499. We show both, the soft segmentation map (2nd column) and the detection output based on this segmentation (3rd column), obtained using our proposed method. We also show results obtained using Topaz picking (without denoise), and crYOLO in the next two columns, respectively. The last column shows the ground truth. Note that for the single-particle dataset, ground truth was obtained using the Topaz particle picking method on the full dose micrograph. Ground truth for EMPIAR-10304 and 10499 were obtained using manual picking.

rials.

Overall, our method is able to show a significant performance under challenging SNR conditions. When SNR is relatively higher, the improvement is less significant. Currently, our model only works on images that are corrupted by Gaussian noise (or can be approximated by Gaussian). We will leave the extension to other noise types to future work. When SNR is so low that it is almost impossible to separate signal from noise (e.g. SNR < 0.0001), our model will mostly likely fail, which is why we did not try denoising single frame. It is theoretically impossible to denoise such images.

5. Conclusion

In this paper, we present a novel joint training framework that performs image denoising and segmentation simultaneously without the need of noiseless images or per-pixel annotated datasets. We show that the complementary information shared between the two tasks allows us to improve the performance of both tasks, especially under extremely low SNR conditions. We validated our approach on real single-particle cryo-EM and cryo-ET datasets and showed that our model is able to outperform SotA methods. Our future work will focus on handling more complex and diverse datasets, including datasets with multiple proteins of interest. We will also extend the applicability of our work to 3-D data to enable protein identification on 3D tomograms. We hope that our algorithm will facilitate structural analysis of challenging biomedical targets such as low molecular weight complexes or thick specimens imaged in their native environments using cryo-ET. In addition, our approach could be applied to low SNR signals in other fields of study such as astronomy imaging to improve object recognition.

References

[1] Adil Al-Azzawi, Anes Ouadou, Highsmith Max, Y. Duan, J. Tanner, and J. Cheng. Deepcryopicker: Fully automated deep neural network for single protein particle picking in

cryo-em. *bioRxiv*, 2020.

[2] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. 01 2019.

[3] Tristan Bepler, K. Kelley, A. Noble, and Bonnie Berger. Topaz-denoise: general deep denoising models for cryoem and cryoet. *Nature Communications*, 11, 2020.

[4] Tristan Bepler, Andrew Morin, A. Noble, J. Brasch, Lawrence Shapiro, and B. Berger. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nature Methods*, pages 1–8, 2019.

[5] Tim-Oliver Buchholz, M. Prakash, Alexander Krull, and Florian Jug. Denoiseg: Joint denoising and segmentation. In *ECCV Workshops*, 2020.

[6] Y. Chen, D. Zhao, L. Lv, and Qichao Zhang. Multi-task learning for dangerous object detection in autonomous driving. *Inf. Sci.*, 432:559–571, 2018.

[7] Bartomeu Coll and Jean-Michel Morel. Non-local means denoising. *Image Processing On Line*, 1, 09 2011.

[8] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising with block-matching and 3d filtering. *Proceedings of SPIE - The International Society for Optical Engineering*, 6064:354–365, 02 2006.

[9] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1635–1643, 2015.

[10] Thibaut Durand, Taylor Mordan, Nicolas Thome, and M. Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5957–5966, 2017.

[11] Theodoros Evgeniou and M. Pontil. Regularized multi–task learning. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004.

[12] Ross B. Girshick. Fast r-cnn. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[13] I. Gubins and R. Veltkamp. Deeply cascaded u-net for multi-task image processing. *ArXiv*, abs/2005.00225, 2020.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018.

[15] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[16] A. Hendriksen, D. M. Pelt, and K. J. Batenburg. Noise2inverse: Self-supervised deep convolutional denoising for linear inverse problems in imaging. *ArXiv*, abs/2001.11801, 2020.

[17] T. Huang, Songjiang Li, Xu Jia, Huchuan Lu, and J. Liu. Neighbor2neighbor: Self-supervised denoising from single noisy images. *ArXiv*, abs/2101.02824, 2021.

[18] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *NeurIPS*, 2019.

[19] Alex Kendall, Yarin Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018.

[20] Diederik P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.

[21] Alexander Kolesnikov and Christoph H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. *ArXiv*, abs/1603.06098, 2016.

[22] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void - learning denoising from single noisy images. pages 2124–2132, 06 2019.

[23] Samuli Laine, Tero Karras, Jaakko Lehtinen, and Timo Aila. High-quality self-supervised deep image denoising. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[24] Kanggeun Lee and Won-Ki Jeong. Noise2kernel: Adaptive self-supervised blind denoising using a dilated convolutional kernel architecture. 12 2020.

[25] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. 03 2018.

[26] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3159–3167, 2016.

[27] Pengfei Liu, Xipeng Qiu, and X. Huang. Recurrent neural network for text classification with multi-task learning. *ArXiv*, abs/1605.05101, 2016.

[28] Pengfei Liu, Xipeng Qiu, and X. Huang. Adversarial multi-task learning for text classification. *ArXiv*, abs/1704.05742, 2017.

[29] W. Liu, Dragomir Anguelov, D. Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and A. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.

[30] Yuanpeng Liu, Qing lei Hui, Zhiyi Peng, S. Gong, and D. Kong. Automatic ct segmentation from bounding box annotations using convolutional neural networks. *ArXiv*, abs/2105.14314, 2021.

[31] Eugene Palovcak, Daniel Asarnow, Melody G. Campbell, Zanlin Yu, and Yifan Cheng. Enhancing the signal-to-noise ratio and generating contrast for cryo-EM images with convolutional neural networks. *IUCrJ*, 7(6):1142–1150, Nov 2020.

[32] Hughes Perreault, Guillaume-Alexandre Bilodeau, N. Saunier, and Maguelonne H'eritier. Spotnet: Self-attention multi-task network for object detection. *2020 17th Conference on Computer and Robot Vision (CRV)*, pages 230–237, 2020.

[33] Pedro H. O. Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1713–1721, 2015.

[34] M. Prakash, Alexander Krull, and F. Jug. Divnoising: Diversity denoising with fully convolutional variational autoencoders. *ArXiv*, abs/2006.06072, 2020.

[35] Y. Quan, Mingqin Chen, T. Pang, and H. Ji. Self2self with dropout: Learning self-supervised denoising from single image. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1887–1895, 2020.

[36] Martin Rajchl, M. J. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, Wenjia Bai, Bernhard Kainz, and D. Rueckert. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE Transactions on Medical Imaging*, 36:674–683, 2017.

[37] Rajeev Ranjan, V. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:121–135, 2019.

[38] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. pages 779–788, 06 2016.

[39] Joseph Redmon, S. Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[40] Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

[41] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015.

[42] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *ArXiv*, abs/1706.05098, 2017.

[43] S. Scheres. A bayesian view on cryo-em structure determination. *Journal of Molecular Biology*, 415:406 – 418, 2012.

[44] Kihyuk Sohn, David Berthelot, C. Li, Zizhao Zhang, N. Carlini, E. D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *ArXiv*, abs/2001.07685, 2020.

[45] Kihyuk Sohn, Zizhao Zhang, C. Li, Han Zhang, Chen-Yu Lee, and T. Pfister. A simple semi-supervised learning framework for object detection. *ArXiv*, abs/2005.04757, 2020.

[46] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *International Journal of Computer Vision*, 128, 07 2020.

[47] T. Wagner, F. Merino, M. Stabrin, T. Moriya, Claudia Antoni, Amir Apelbaum, Philine Hagel, Oleg Sitsel, Tobias Raisch, Daniel Prumbaum, D. Quentin, D. Roderer, S. Tacke, Birte Siebolds, E. Schubert, T. Shaikh, Pascal Lill, Christos Gatsogiannis, and S. Raunser. Sphire-cryolo is a fast and accurate fully automated particle picker for cryo-em. *Communications Biology*, 2, 2019.

[48] Feng Wang, Huichao Gong, Gaochao Liu, Meijing Li, Chuangye Yan, Tian Xia, Xueming Li, and Jianyang Zeng. Deeppicker: A deep learning approach for fully automated particle picking in cryo-em. *Journal of Structural Biology*, 195(3):325–336, 2016.

[49] Keze Wang, Xiaopeng Yan, Dongyu Zhang, Lei Zhang, and L. Lin. Towards human-machine cooperation: Self-supervised sample mining for object detection. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1605–1613, 2018.

[50] Yude Wang, J. Zhang, Meina Kan, S. Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12272–12281, 2020.

[51] Qizhe Xie, Zihang Dai, E. Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. *arXiv: Learning*, 2020.

[52] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014.

[53] Bolei Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.

[54] Dengyong Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, 2003.