Deep learning based domain adaptation for mitochondria segmentation on EM volumes

Daniel Franco-Barranco[1,2], Julio Pastor-Tronch[1], Aitor Gonzalez-Marfil[1], Arrate Muñoz-Barrutia[3,4], and Ignacio Arganda-Carreras[1,2,5]

[1] Dept. of Computer Science and Artificial Intelligence, University of the Basque Country (UPV/EHU)
[2] Donostia International Physics Center (DIPC)
[3] Universidad Carlos III de Madrid
[4] Instituto de Investigación Sanitaria Gregorio Marañón
[5] Ikerbasque, Basque Foundation for Science
daniel_franco001@ehu.eus

Abstract

Background and Objective: Accurate segmentation of electron microscopy (EM) volumes of the brain is essential to characterize neuronal structures at a cell or organelle level. While supervised deep learning methods have led to major breakthroughs in that direction during the past years, they usually require large amounts of annotated data to be trained, and perform poorly on other data acquired under similar experimental and imaging conditions. This is a problem known as domain adaptation, since models that learned from a sample distribution (or source domain) struggle to maintain their performance on samples extracted from a different distribution or target domain. In this work, we address the complex case of deep learning based domain adaptation for mitochondria segmentation across EM datasets from different tissues and species.
Methods: We present three unsupervised domain adaptation strategies to improve mitochondria segmentation in the target domain based on (1) state-of-the-art style transfer between images of both domains; (2) self-supervised learning to pre-train a model using unlabeled source and target images, and then fine-tune it only with the source labels; and (3) multi-task neural network architectures trained end-to-end with both labeled and unlabeled images. Additionally, to ensure good generalization in our models, we propose a new training stopping criterion based on morphological priors obtained exclusively in the source domain. The code and its documentation are publicly available at https://github.com/danifranco/EM_domain_adaptation
Results: We carried out all possible cross-dataset experiments using three publicly available EM datasets. We evaluated our proposed strategies and those of others based on the mitochondria semantic labels predicted on the target datasets.
Conclusions: The methods introduced here outperform the baseline methods and compare favorably to the state of the art. In the absence of validation labels,

monitoring our proposed morphology-based metric is an intuitive and effective way to stop the training process and select in average optimal models.

1 Introduction

Supervised learning has achieved great success in computer vision leading to the development of robust algorithms that have been successfully applied in diverse research areas. The generalization capability and reliability of these algorithms are based on the assumption that the data used to train them and the data used to test them are drawn from the same distribution or *domain*. Thus, when the training data is not representative enough of the target population, there is a drop in the algorithm's performance [1]. This performance gap is highly significant when the data acquisition changes (i.e., protocol, instrument) even for a similar target domain. In the particular case of biomedical imaging, data distributions are highly biased due to the variety of acquisition techniques and protocols. Therefore, a significant number of annotations is usually required to ensure a good representation of the population.

Nevertheless, collecting and annotating these datasets is extremely expensive in both time and human resources [2]. For that reason, the field of domain adaptation has emerged to tackle both issues: the reduction of the domain gap difference and the generation of annotated data. The purpose of domain adaptation is to learn from labeled data in a source domain to perform well on a different, but related target domain without any annotation [3].

Aiming to reduce source and target domain dissimilarity, many methods have been proposed to create synthetic source images, and therefore, increase the heterogeneity of the data [4]. Some of these approaches generate new images from random noise without any other conditional information for Computed Tomography (CT) data [5,6], Magnetic Resonance (MR) [6,7,8] or chest X-rays [9,10]. Other methods of synthetic data generation aim to create new training samples using target domain samples and labeled source domain knowledge [3]. A large amount of this cross-modality synthesis work has been proposed for adapting MR data to CT [11,12], CT to MR [13,14] and MR to Positron Emission Tomography (PET) [15,16].

Additionally, image generation can be constrained by the appearance of the anatomical structures and segmentation maps. Many approaches have been presented in the literature that generate image-mask pairs, for instance, implementing domain adaptation from CT to MR [17], generating synthetic samples to solve a segmentation task [18,19,20,21] or for one-shot segmentation [22,23,24].

In the particular case of Electron Microscopy (EM) volumes of the brain, its accurate segmentation is essential to characterize the neural structures present in the volume. Several recent works have been presented in the literature that use domain adaptation to segment neuronal structures [25,26,27], vesicles [28], mitochondria [29,30,31,32] and whole-cell organelles [33]. For the specific task of mitochondria segmentation, domain adaptation methods have been introduced to handle the limited availability of labeled data [34,35,36].

In this work, we address the complex case of domain adaptation for mitochondria segmentation across EM datasets from different tissues and species. We assume the absence of target domain annotations to simulate a real scenario. More specifically, we compare three deep learning based strategies to improve mitochondria segmentation in the target dataset based on 1) style transfer between domains, 2) self-supervised learning, and 3) multi-task neural network architectures. To demonstrate the potential of these three strategies, we employed a cross-domain thorough study between three publicly available datasets for mitochondria segmentation. The same initial conditions and basic architectural design choices are maintained across all strategies, which are also compared with the same supervised baseline methods.

In brief, our main contributions are as follows:

- We have presented state-of-the-art style transfer as a solution for domain adaptation for mitochondria segmentation in EM volumes.
- We introduce a self-supervised approach based in a pre-training step using both datasets without annotations and a final fine-tuning with only source annotations.
- We have performed a cross-dataset analysis of state-of-the-art deep multi-task networks for EM datasets in the context of domain adaptation and propose a novel architecture based on one of them.
- As a stopping criterion, we propose a new metric to ensure a good generalization towards the target domain based on the morphology of the resulting mitochondria segmentation.

2 Related work

The work presented here focuses on domain adaptation and style transfer methods for EM image analysis. By domain and style, we refer to the intrinsic feature space and characteristics of a particular dataset and the distribution from where it is drawn. Domain adaptation can be seen as a particular type of transfer learning where instead of trying to transfer the knowledge from task A in domain A to task B in domain B, the tasks are kept the same while the domains are different. On the other hand, style transfer is mainly focused on adapting the domain from one dataset to another.

Existing domain adaptation methods can be divided depending on the label availability during the training process. Thus, they can be supervised, if both source and target domain labels are available; semi-supervised, if source labels and some target labels are available; and unsupervised, if only source labels are available while target data is entirely unlabeled [37]. Moreover, methods can also be categorized based on the learning model used, i.e., either shallow (usually relying on predefined image features and traditional machine learning models) or deep (if they use deep learning architectures). In this paper, we focus on the strategy known as deep unsupervised domain adaptation.

One particular way of addressing this problem is by style transfer. For instance, the Cycle Generative Adversarial Networks (CycleGAN) [38] approach

is becoming an effective method in medical image synthesis. Many variations have been presented addressing cross-domain style transfer problems targeting different sources and target types of data, such as from MR to CT [39,40,17,41], transferring the stain style for histopathological images [42,43,44] or creating target-style data pairs, image and mask, without using any annotation [45,46,47].

More recent approaches to address style transfer exploit contrastive learning [48], where models are trained without labels to learn which data samples are similar or different. Similarity is defined in an unsupervised way, by using different data augmentation techniques to create similar examples to the original image and then maximizing a similarity function (e.g., mutual information) during training. Following this idea, Contrastive Unpaired Translation (CUT) [49] compares unpaired image patches and associates similar patches to each other while disassociating them from others. This way, the model learns to pay attention to the commonalities between domains. For instance, a patch containing a mitochondrion will have a high similarity with a patch in a different tissue containing mitochondria, or at least a higher value than if it is compared with a patch showing other organelles. Thus, a generator learns to change the style of input images to match a target style.

Another way to address this domain problem is by using self-supervised learning (SSL), which consists in establishing a *pretext* task using unlabeled related images that do not require to be annotated by an expert to initially train the model. Then, the model is used as the starting training point for the *downstream* (segmentation) task. The main advantage is that the pretext examples (or pseudo-labels) are automatically generated from existing raw data, not being conditioned to the number of available expert-reviewed images. Therefore, during the pre-training step, models can leverage from all available images to learn useful feature representations.

In the computer vision literature, related to natural images, the usefulness of this self-supervised pre-training step has been widely explored for several tasks. Namely, the coloring of a grayscale image [48,50,51], the restoration of a distorted or deteriorated image [52,53,54,55], the prediction of the transformation performed in an image [56] or even, the re-ordering of pieces or frames of images [57,58] and videos [59]. However, there is hardly any work applying this methodology to microscopy images. The published works mostly focus on reducing the number of annotated images required for training thanks to a good network initialization achieved by pre-training with denoising [60,61,62,63], jigsaw solving [64,65] and image restoration [66].

Finally, another approach is based on multi-task deep neural network architectures that receive both source and target samples as input. In this case, apart from solving the downstream task for the source (labeled) data, the model aims to exploit the features of the target domain to learn the feature shift between domains. Among these types of unsupervised and semi-supervised domain adaptation methods, we find the Y-Net [35], used for the segmentation of EM images. Its architecture consists of an encoder-decoder such as a U-Net [67], coupled with a second decoder in an autoencoder strategy. While one decoder is trained

for segmentation, using the images with available labels, the second decoder is trained to reconstruct all available images, including the unlabeled ones, in an unsupervised manner. Since both decoders share the same encoder, the features learned by the autoencoder are used for segmentation too. Consequently, the model works with unlabeled (target domain) data features. Following this idea, in combination with adversarial losses, similar models such as Domain Adaptive Multi-Task Learning network (DAMT-Net) [36] have been proposed. This network builds on top of the Y-Net architecture and adds two discriminators during training, following a Generative Adversarial Network (GAN) approach. The first discriminator uses the predicted segmentation, while the second discriminator uses the final feature maps of the network.

3 Methods

To address the problem of domain adaptation between different EM datasets, we present different approaches that reduce the domain shift. Firstly, a cross-domain baseline is introduced using stable state-of-the-art models [32] trained only on source domains. Next, a simple histogram matching between domains is added as pre-processing prior to the use of the baseline models. Finally, more sophisticated domain adaptation approaches are presented based on (1) a modern style-transfer technique, (2) self-supervised pretext tasks, and (3) state-of-the-art domain adaptation multi-task deep neural networks.

3.1 Cross-dataset baseline

As a reference method to compare our results with, we use our recent stable 2D Attention U-Net model [32] trained on the labeled source domain and tested directly on the target domain (without any adaptation method). This network is a modified version of the U-Net [67] including attention gates [68] in the skip connections that has proven to produce consistently robust results in the segmentation of mitochondria on EM volumes [32]. Its architecture is shown in Figure 1.

3.2 Histogram matching

A straightforward approach to make the images of one domain look closer to the images of another domain is histogram matching. Most commonly, this technique is applied to one source image so that its histogram matches the histogram of a target image [69]. Here instead, we use as target histogram the mean histogram of the target domain images, so the histogram of all source images are transformed to match it.

Some images of our datasets present zero-padding surrounding the tissue, which provokes an artificial high pick at the zero value in their histograms. Since we are only interested in matching the histogram of the tissue part of the images, we modified the actual number of zeros with linear regression using the

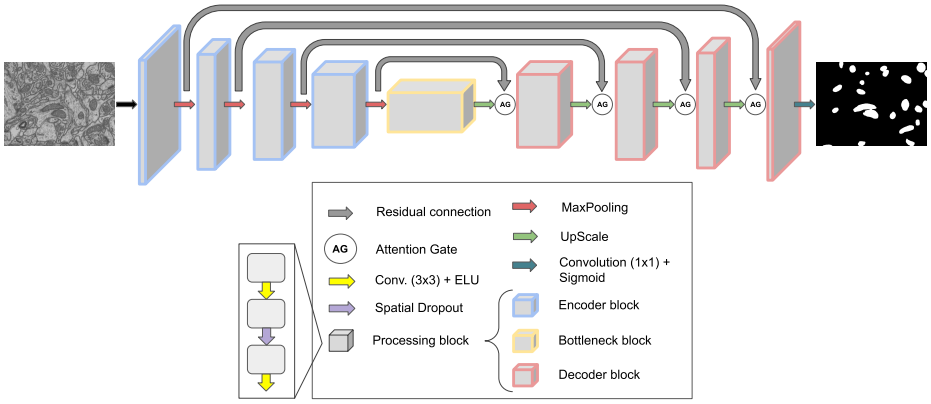


Fig. 1: Architecture of our Attention U-Net [32] used for mitochondria semantic segmentation.

first bins of the original histogram. We set the value to zero in the absence of initial values or when predicting a negative number. This process is done for both target and source histograms. Some example images processed with this histogram matching method can be seen in Figure 2.

3.3 Style transfer approach

As described in the previous section, domain adaptation can be considered a style-transfer problem. In particular, we were motivated by the success of the recent Contrastive Unpaired Translation (CUT) method [49] for the problem of unpaired image-to-image translation. Therefore, we tested it on our EM datasets for mitochondria segmentation and re-analyzed the cross-domain performance of our supervised baseline networks on the translated target datasets.

In order to learn the translation between source and target images, this method randomly crops the images to patches of 512×512 pixels and maximizes the mutual information between the input and output patches using a contrastive learning framework. This way, corresponding patches (positives) are mapped together in feature space and far from other patches (negatives). Results of this method are shown in Figure 3. All cross-dataset stylization results can be found in Section S1.

Following the recommendations of the original paper, we used the default hyperparameter setting as provided in their public implementation, which corresponds with training the method for 400 epochs, with Adam as optimizer and a learning rate of $2e-4$.

3.4 Self-supervised approach

As an alternative approach, we propose a self-supervised framework where we leverage from the use of two sequential training steps: (1) an initial generative

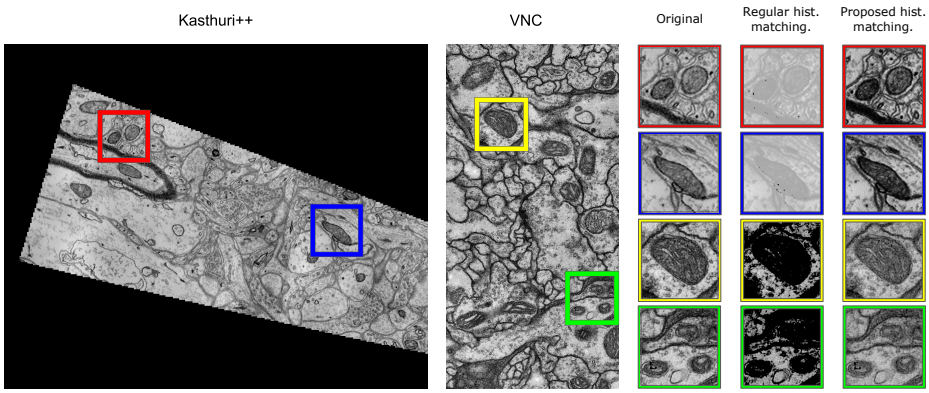


Fig. 2: Examples of histogram matching between source and target domain images. When using a dataset such as Kasthuri++ containing padding (non-tissue) pixels, regular histogram matching methods fail and need to be corrected to focus only on tissue intensities. From left to right: original full-size images from the Kasthuri++ and VNC datasets; four zoomed areas of both images (in red, blue, yellow and green), with their corresponding (Original) pixel values, followed by their histogram-matched versions using the full histogram (Regular hist. matching) and our proposed method to predict the zero values and avoid using padding pixels (Proposed hist. matching). For the red and blue examples, Kasthuri++ is the source domain and VNC is the target domain, while the opposite occurs for the yellow and green examples.

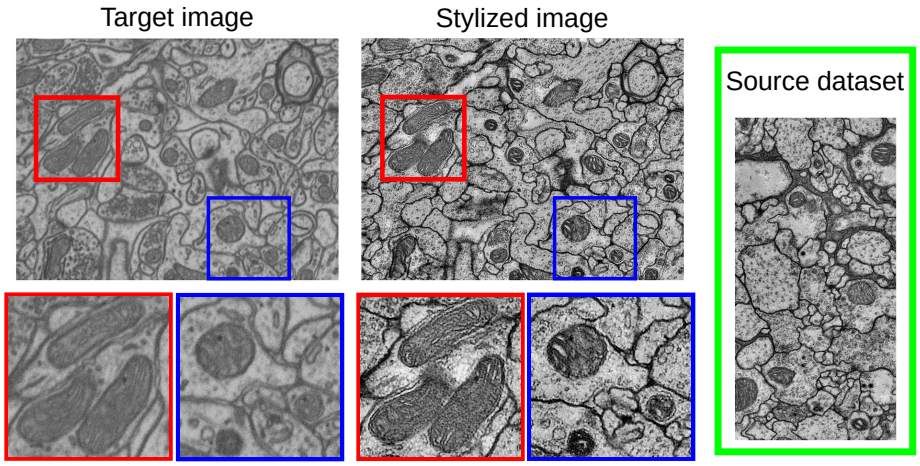


Fig. 3: Stylization made by the CUT [49] method using Lucchi++ images as source and the VNC dataset as reference (target) style. From left to right: Original Lucchi++ sample; its stylized result with the appearance of VNC; and a VNC image sample (green box). Blue and red boxes show zoomed areas from the source and stylized images.

self-supervised step including both (source and target) datasets without annotations, and (2) a fully-supervised fine-tuning step using only the source images and their labels. A summary of our self-supervised workflow is depicted in Figure 4.

Super-resolution pretext task. In this pretext task, our Attention U-Net is trained to enhance the resolution of images from both the source and target datasets. This first step aims to reach a good starting point to solve the downstream task (i.e., supervised mitochondria segmentation). The input images are synthetically generated low-resolution images, while the ground truth is formed by the (high-resolution) original ones. To generate the synthetic input images, the original images are distorted with normally distributed Gaussian noise with $\mu = 0$ and $\sigma = 0.1$ as a fraction of the dynamic range of the image. Next, the images are downsampled by a factor of two in both axes and then upsampled by the same factor to simulate a process where the original resolution is worsened. For both downsampling and upsampling, bilinear interpolation is used.

Source supervised training. Once the model has been pre-trained, the encoder gets frozen. Then, the rest of the network (bottleneck and decoder) are fine-tuned with the available source image annotations to perform semantic segmentation. The source images are pre-processed so their histogram matches that of the target domain. The idea behind freezing the encoder is to enforce the model to remember features learnt during the previous super-resolution step from the target dataset. Thus, allowing for a better generalization and performance in the unlabeled target dataset.

It is worth noting that during the super-resolution step, all available source and target images are used to train the model. That is because the input-label pairs are automatically generated from the raw data but no annotations are used. In the second step, only the training subset from the source dataset and its annotations are used to fine-tune the model.

During the pre-training step, the network is run for 200 epochs, following a one-cycle learning rate policy [70] with a maximum learning rate of $5e-4$, and Adam optimizer. Next, the fine-tuning step is carried out for 60 epochs, using as well a one-cycle learning rate scheduler with a maximum learning rate of $1e-4$ and Adam optimizer. In both cases, the optimal batch size was found to be 1. All training images were randomly cropped to patches of 256×256 pixels, from which 10% was used for validation. A more detailed description of the hyperparameters can be found in Table S3.1 as well as all combinations tested.

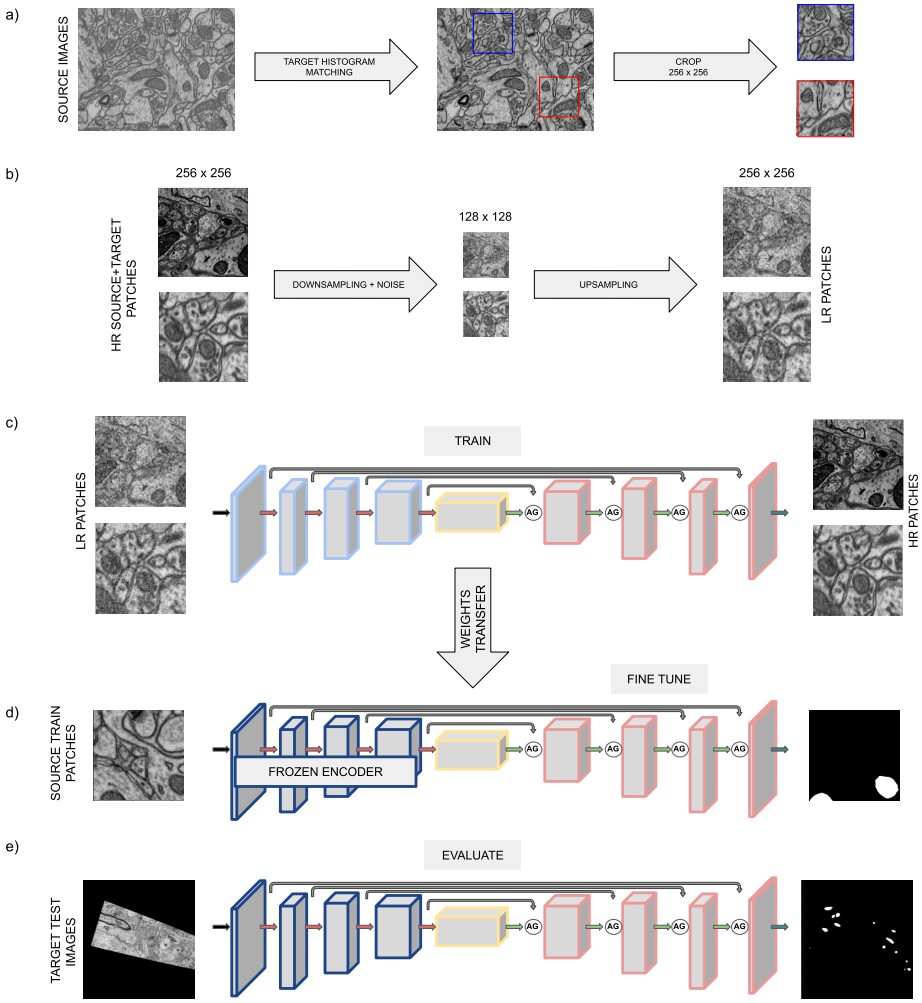


Fig. 4: Diagram of our self-supervised workflow for domain adaptation. From top to bottom: a) The source dataset is adjusted to the target image histogram and cropped into patches of 256×256 pixels; b) crops from both datasets are used to generate low-resolution samples by undersampling them and adding Gaussian noise; c) our Attention U-Net is pre-trained by learning to super-resolve the generated patches to their original versions; d) the encoder of the model is frozen and the rest of the network is fine-tuned for the mitochondria segmentation task using only source training patches and their corresponding binary masks; e) the model is evaluated on the target test dataset.

3.5 Multi-task neural networks

Following the idea behind Y-Net [35], we have built a similar architecture taking as a base model the previously mentioned Attention U-Net [32]. We refer to this network as Attention Y-Net. In short, the architecture consists of the classical encoder-decoder setup, where a new second decoder is placed. We can see the architecture as the combination of the Attention U-Net and an autoencoder, where both parts share the same encoder. The architecture is illustrated in Figure 5.

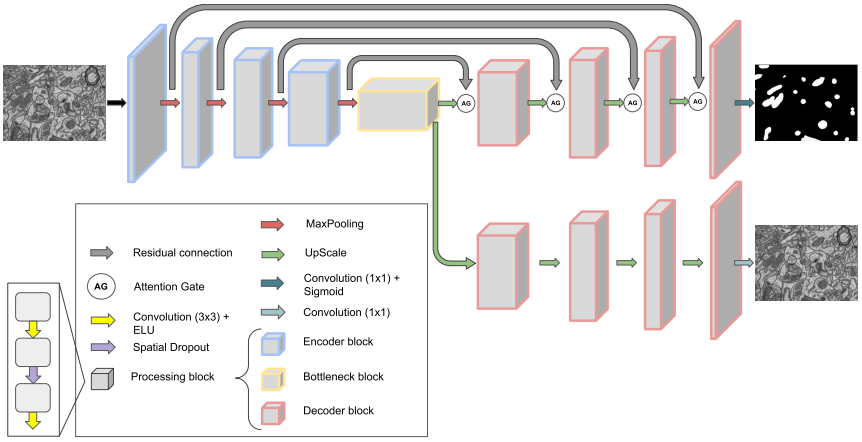


Fig. 5: Architecture of the proposed Attention Y-Net used for domain adaptation. The architecture is formed by one encoder and two decoders: one for image reconstruction (without skip connections) and one for segmentation (with skip connections and attention gates).

The network is trained using a loss function ($\mathcal{L}$) made of two terms: a segmentation term based on the binary cross-entropy between the predicted and ground truth masks ($\mathcal{L}_{BCE}$), and a reconstruction term based on the mean squared error between the predicted and the original grayscale images ($\mathcal{L}_{MSE}$), as given by

$$\mathcal{L} = \alpha\mathcal{L}_{MSE} + (1-\alpha)\mathcal{L}_{BCE}, \tag{1}$$

where the weight α is a numeric value between 0 and 1. For those images without available labels (binary masks), the $\mathcal{L}_{BCE}$ value will be 0.

In its original work, the training of the Y-Net [35] was proposed in two sequential steps. First, the network is trained unsupervised to perform only reconstruction ($\alpha = 1$). Then, the model is fine-tuned to perform segmentation with the available labels ($\alpha = 0$). However, we have experienced instability in this step. Namely, quite often, the predicted reconstruction of the network was a flat grey-value image. Therefore, we propose a new additional step before the

unsupervised pre-training, which combines both tasks using all the available data. We set $\alpha = 0.98$, which was experimentally found to help balancing both loss terms.

With our additional pre-training step, the network consistently outputs improved results, out of the local minimum achieved with the flat grey-value image. Next, we freeze the network encoder (blue blocks in Figure 5). Otherwise, the network forgets the target domain features in the next step. Experimentally, we observed that the network performs better if we let the bottleneck and the two decoders unfrozen. Remarkably, as observed with the self-supervised approach, the performance of the whole process was greatly enhanced thanks to the use of histogram matching after the first step.

The first step was carried out for 50 epochs. We used an initial learning rate of $1e-3$ that got reduced when reaching plateaus, stochastic gradient descent (SGD) as optimizer and a patience of 7 epochs over the monitored validation loss. In the second training step, we train for 40 epochs (with a patience of 6). We use a learning rate of $2e-4$, and a "reduce on plateau" scheduler once again, but this time with Adam optimizer. Finally, in the last training step, we train for 100 epochs (the different stop criteria will be analysed later). We follow a one-cycle learning rate policy [70] with a maximum learning rate of $2e-4$, and use Adam as optimizer. For all training steps, the optimal batch size was found to be 1. The input to the model consists of 1000 random cropped patches of 256×256 pixels, from which 10% is used for validation. This training configuration was empirically found. A more detailed description of the hyperparameters as well as all combinations tested can be found in Table S3.2.

4 Experimental Results

4.1 EM Datasets

All the experiments performed in this work are based on the following publicly available datasets:

EPFL Hippocampus or Lucchi dataset [71]. The original volume represents a $5 \times 5 \times 5$ $(\mu\text{m})^3$ section of the CA1 hippocampus region of a mouse brain, with an isotropic resolution of $5 \times 5 \times 5$ nm per voxel. The volume of $2048 \times 1536 \times 1065$ voxels was acquired using scanning electron microscopes (SEM), specifically with focused ion beam scanning electron microscopy (FIB-SEM). The mitochondria of two sub-volumes formed by 165 slices of 1024×768 pixels were manually labeled by experts, and are used as the official training and test partitions. In particular, we used a more recent version of the labels [30] after two neuroscientists and a senior biologist re-labeled mitochondria by fixing misclassifications and boundary inconsistencies.

Kasthuri++ dataset [30]. This is a re-labeling of the dataset by [72]. The volume corresponds to a part of the somatosensory cortex of an adult mouse and was acquired using scanning electron microscopes (SEM) as Lucchi++, but specifically with serial section electron microscopy (ssEM). The train and test

volume dimensions are $1463 \times 1613 \times 85$ voxels and $1334 \times 1553 \times 75$ voxels, respectively, with an anisotropic resolution of $3 \times 3 \times 30$ nm per voxel.

VNC dataset [73]. This dataset represents a $4.7 \times 4.7 \times 1$ $(\mu\text{m})^3$ serial section transmission electron microscopy (ssTEM), acquired using transmission electron microscopy (TEM), of the *Drosophila* melanogaster third instar larva ventral nerve cord, with an an isotropic resolution of $4.6 \times 4.6 \times 45 - 50$ nm per voxel. Two volumes of $1024 \times 1024 \times 20$ voxels were acquired, but only one of them was labeled. For that reason and following common practice, we use only the later and split the data volume along the x axis into two subsets with equal size ($20 \times 512 \times 1024$ voxels) that constitute our training and test partitions.

For fair comparison with other published work, only the training set labels of the source datasets are used during the supervised or fine-tuning steps of our approaches, while the quantitative evaluation is performed only on the test set of the target datasets.

4.2 Evaluation metrics

Since our downstream task is semantic segmentation, we evaluate all methods using the Jaccard index of the positive class or foreground intersection over union (IoU_F), defined as

$$IoU_F = \frac{TP}{TP + FP + FN} \tag{2}$$

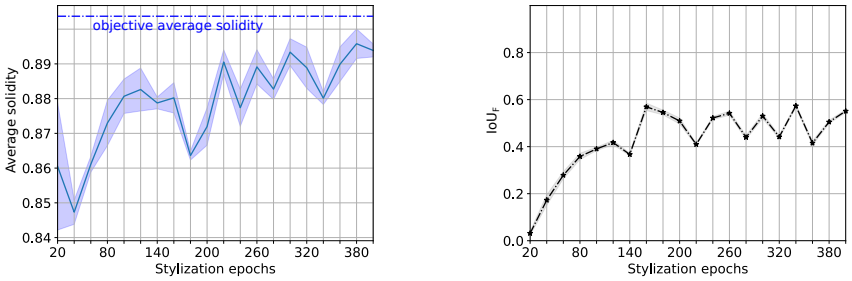
where TP are the true positives, FP the false positives and FN the false negatives. As a convention, the positive class is foreground and the negative class, background. This way, IoU_F values range from 0 to 1, where 0 represents no overlap at all between the ground truth and the predicted mitochondria masks, and 1 means a perfect overlap.

4.3 Stopping criterion

An intrinsic issue of unsupervised domain adaptation methods is blindly deciding when to stop their respective optimization processes since no labels are available from the target domain samples to guide us in such optimization. This problem is common to all our proposed approaches, either to select the number of stylization iterations or to fix the number of epochs to train our self-supervised or multi-task models. For that reason, we have selected a stopping criterion using morphological priors extracted from the source labels. More specifically, we calculate the average *solidity* $\overline{S}$ of each mitochondrion in the image as:

$$\overline{S} = \frac{1}{N} \sum_{n=1}^{N} solidity(n) \tag{3}$$

where N is the total number of objects (in our case mitochondria instances) in the image and $solidity(n)$ is the ratio of pixels in the nth object to pixels of the convex hull of that object. In practice, each instance is found by the connected components algorithm on the binarized outputs of the models.



(a) (b)

Fig. 6: Example of connection between average solidity $\overline{S}$ and IoU_F values: (a) Average value of $\overline{S}$ (shaded area represents its standard deviation) as function of the epochs run for the style-transfer method. The source $\overline{S}$ value is depicted with a dashed line. (b) Average IoU_F value (shaded area represents its standard deviation) as function of the same epochs. Values are calculated over the output of ten style-transfer model executions using Kasthuri++ and VNC as source and target domains, respectively.

The main advantage of the average solidity is that it is agnostic of the dataset resolution and easy to implement. As a criterion, we can monitor the $\overline{S}$ value of the predictions in the target dataset and stop optimizing our domain adaptation methods when it moves away from the objective $\overline{S}$ value (measured in the source domain). To select the best model, one can simply take the model producing test masks with the $\overline{S}$ value that is closest to the objective one. Moreover, to increase the robustness of this criterion, we discard very tiny objects (with less than ten pixels) for all datasets.

An example of the connection between the $\overline{S}$ values of test predictions and their respective segmentation results expressed in terms of IoU_F is shown in Figure 6. One can observe that the range of epochs where the test $\overline{S}$ values are closer to the objective $\overline{S}$ (calculated in the source domain) in Figure 6a correspond, overall, to the epochs with higher IoU_F values in Figure 6b. The same plots for all methods and cross-dataset experiments can be found in Section S2.

4.4 Cross-dataset results

All the methods proposed here were applied to all the possible source-target combinations of the three EM datasets introduced in Section 4.1. Moreover, for a more detailed evaluation and comparison with the state of the art, we executed as well the same experiments using the publicly available implementation of DAMT-Net [36]. As it is an extended practice on EM image processing, we also tested all methods on the same image data after preprocessing them using contrast limited adaptive histogram equalization (CLAHE) [74]. Notice CLAHE

is a contrast equalization method, thus not intended to match two intensity distributions. However, its effect on the image contrast may bring the histogram of our datasets closer to each other.

To ensure the robustness of the proposed training configurations and hyperparameters, each experiment was repeated ten times using exactly the same setup. A full description of the search of hyperparameters for each approach can be found in Section S3.

The best results based on the average IoU_F of the predicted mitochondria in the corresponding target test images for each method are shown in Table 1. Furthermore, we explored the impact of stopping the model training by each of the following criteria: (1) monitoring the IoU_F value of the source validation set (and also selecting the best model based on that value); (2) leaving the model train for a fixed number of epochs; and (3) monitoring the average solidity values of the target test set (and selecting the model that better approaches the known source average solidity value).

First, although expected, it is worth mentioning that all tested methods outperform the baseline in all cases, demonstrating the need for a domain adaptation strategy that allows addressing the domain shift problem. Secondly, we can observe an evident boost in performance by simply applying either our histogram matching method to the target images or CLAHE as preprocessing for all images, and re-using the baseline models for inference. Interestingly, on one of the source-target combinations (Lucchi++ as source and Kasthuri++ as target) these strategies provide very good segmentation results ($IoU_F = 0.679$ and 0.620 respectively), but they perform poorly ($IoU_F = 0.268$ and 0.249) on the opposite experiment (Kasthuri++ as source and Lucchi++ as target). This reflects an asymmetric aspect of the problem and the need for solutions that learn more than just simple histogram image features. Moreover, these results show our proposed methods generally perform favourably to the state of the art, represented by DAMT-Net [36]. In particular, our style-transfer based approach provides consistent results across all datasets, followed by our proposed multi-task Attention Y-Net.

Finally, the choice of the stopping criterion seems to play an important role improving the segmentation results depending on the dataset combination. Although the monitoring of the source validation results is a good indicator of the performance in the target domain by the multi-task networks (DAMT-Net and Attention Y-Net), we observe their segmentation can be improved by either leaving the training converge (with a maximum number of epochs) or by monitoring the target average solidity instead.

Some qualitative results of the learning-based methods are shown in Figure 7, where the probability maps of mitochondria masks produced by each method are displayed side by side for the same sample images. More specifically, the predictions shown were obtained using average solidity as stopping criterion. In agreement with the quantitative results of Table 1, we can observe most methods predict reasonable masks when Lucchi++ is used as the target dataset (where the IoU_F values are in the range of $\sim 0.5 - 0.7$), but present different levels of

Stop criteria	Method	Source: Lucchi++		Source: Kasthuri++		Source: VNC	
		Kasthuri++	VNC	Lucchi++	VNC	Lucchi++	Kasthuri++
Source val set	Baseline [32]	0.017±0.008	0.009±0.010	0.000±0.000	0.095±0.013	0.351±0.101	0.288±0.050
	Baseline [32] + CLAHE	0.620±0.051	0.249±0.021	0.433±0.085	0.121±0.045	0.586±0.016	0.534±0.065
	Baseline [32] + HM (ours)	0.679±0.043	0.265±0.028	0.268±0.048	0.111±0.011	0.531±0.019	0.454±0.035
	Attention Y-Net + HM (ours)	0.668±0.020	0.402±0.040	0.704±0.045	0.252±0.048	0.536±0.022	0.389±0.041
	DAMT-Net [36]	0.279±0.078	0.469±0.054	0.569±0.088	0.324±0.038	0.491±0.102	0.162±0.042
	DAMT-Net [36] + HM	0.226±0.037	0.489±0.040	0.438±0.094	0.274±0.080	0.371±0.123	0.170±0.049
	DAMT-Net [36] + CLAHE	0.299±0.099	0.497±0.029	0.547±0.088	0.346±0.047	0.545±0.039	0.221±0.085
Last epoch	Style transfer (ours, [49])	0.515±0.011	0.586±0.009	0.569±0.003	**0.551±0.006**	0.638±0.014	**0.654±0.026**
	SSL + HM (ours)	0.568±0.165	0.327±0.135	0.511±0.145	0.138±0.043	0.582±0.237	0.237±0.191
	SSL + CLAHE (ours)	0.254±0.159	0.149±0.113	0.456±0.189	0.153±0.077	0.205±0.156	0.162±0.111
	SSL + HM + CLAHE (ours)	0.578±0.160	0.187±0.084	0.421±0.177	0.166±0.061	0.270±0.139	0.116±0.091
	Attention Y-Net + HM (ours)	0.669±0.019	0.388±0.026	0.719±0.024	0.232±0.024	0.540±0.014	0.404±0.016
	DAMT-Net [36]	0.261±0.039	0.455±0.066	0.581±0.057	0.295±0.040	0.449±0.082	0.169±0.055
	DAMT-Net [36] + HM	0.258±0.062	0.422±0.169	0.416±0.078	0.276±0.072	0.380±0.077	0.187±0.118
	DAMT-Net [36] + CLAHE	0.284±0.047	0.482±0.050	0.440±0.135	0.319±0.100	0.488±0.061	0.233±0.100
Solidity	Style transfer (ours, [49])	0.703±0.009	0.605±0.032	0.572±0.044	0.509±0.034	0.608±0.017	0.560±0.032
	Style transfer (ours, [49]) + CLAHE	**0.768±0.020**	**0.671±0.009**	0.529±0.051	0.146±0.165	0.581±0.017	0.572±0.078
	SSL + HM (ours)	0.685±0.092	0.394±0.102	0.572±0.109	0.136±0.027	**0.694±0.022**	0.278±0.171
	SSL + CLAHE (ours)	0.204±0.164	0.165±0.164	0.477±0.168	0.170±0.073	0.253±0.150	0.186±0.088
	SSL + HM + CLAHE (ours)	0.649±0.093	0.152±0.069	0.510±0.177	0.177±0.043	0.249±0.105	0.083±0.084
	Attention Y-Net + HM (ours)	0.713±0.029	0.397±0.013	**0.728±0.031**	0.310±0.067	0.508±0.030	0.416±0.026
	Attention Y-Net + CLAHE (ours)	0.729±0.017	0.420±0.041	0.635±0.049	0.350±0.025	0.545±0.031	0.516±0.094
	Attention Y-Net + HM + CLAHE (ours)	0.731±0.047	0.396±0.023	0.678±0.029	0.360±0.017	0.551±0.030	0.565±0.020
	DAMT-Net [36]	0.223±0.091	0.441±0.140	0.608±0.070	0.224±0.073	0.502±0.048	0.180±0.063
	DAMT-Net [36] + HM	0.230±0.038	0.497±0.059	0.655±0.037	0.308±0.035	0.551±0.045	0.172±0.114
	DAMT-Net [36] + CLAHE	0.244±0.043	0.506±0.037	0.625±0.066	0.291±0.100	0.554±0.053	0.288±0.097

Table 1: Cross-dataset domain adaptation methods evaluation. Results are shown based on the mean IoU_F value (± standard deviation) obtained in the test partition of the target datasets under the three possible stopping criteria: (1) performance on the validation partition of the source dataset, (2) maximum number of epochs (experimentally found for each method), and (3) the proposed average solidity metric. The best results of each column are shown in bold. CLAHE and HM refer to the use of contrast limited adaptive histogram equalization [74] and histogram matching as pre-processing methods, respectively.

performance when predicting the mitochondria of the two other datasets used as target. Remarkably, all methods except our style-transfer approach struggle with the VNC/Kasthuri++ and Kasthuri++/VNC combinations, suggesting a larger domain shift between those two datasets.

5 Conclusions and Discussion

In this paper, we address the problem of domain adaptation for the challenging task of semantic segmentation of EM volumes. More specifically, we propose three novel solutions that built on top the deep-learning based state of the art by means of (1) unsupervised style transfer to transform the target domain images into the "style" of the source domain and then reuse robust models trained on annotated data; (2) self-supervised learning to pre-train our segmentation models without annotations and then fine-tune them using the source labels; and (3) a multi-task deep architecture able to learn from both labeled and unlabeled data. All methods have been evaluated under the same setups using three publicly available EM datasets of different modalities (FIB-SEM, ssEM and ssTEM) and each of their possible source-target combinations. In addition, we propose a novel unsupervised metric to avoid blindly selecting the best model during training.

First of all, quantitative and qualitative results prove that learning-based methods are needed to deal with the domain shift in five out of the six cross-

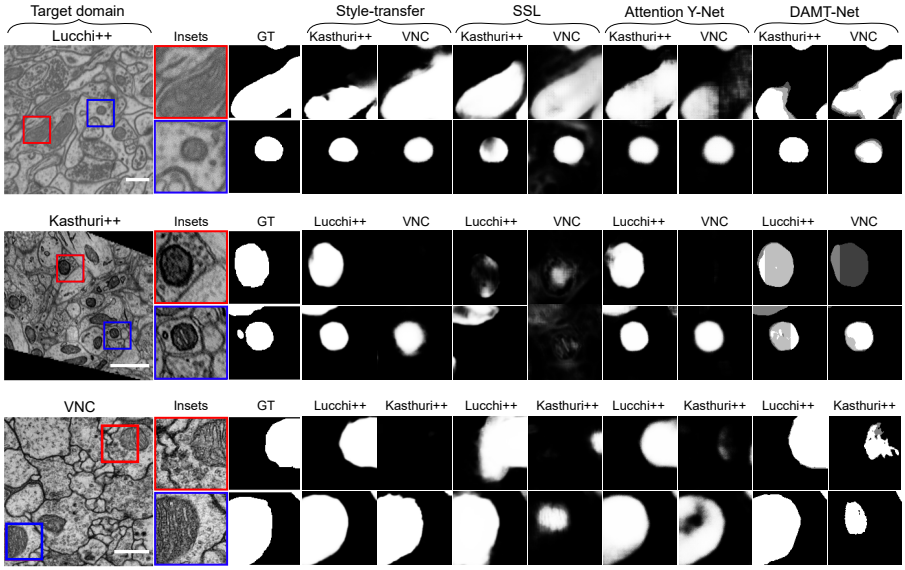


Fig. 7: Cross-dataset segmentation results using average solidity as stopping criterion for all learning based methods. From top to bottom: segmentation results when using Lucchi++, Kasthuri++ and VNC as target datasets. From left to right: image sample from the target dataset, two crops of that sample (in red and blue), their corresponding ground-truth (GT) binary masks, and probability maps produced by each method (Style-transfer, SSL, Attention Y-Net and DAMT-Net). The white scale bar represents 500 nm.

dataset experiments. Only in one combination (Lucchi++ as source domain and Kasthuri++ as target domain) an ad-hoc histogram matching method has been able to reduce the shift at the level of the learning approaches.

Regarding the proposed approaches, the style-transfer based method produces segmentation results with consistently medium-high IoU_F values ($\sim 0.5-0.6$), specially when the stylization is run for a large number of epochs (> 200, see Section S2). The performance of our SSL and Attention Y-Net methods also gets stabilized after a fixed number of training epochs (60 and 100, respectively) as can be seen in Section S2. However, their results are not as consistent as those of the style-transfer approach, oscillating between low ($0.1-0.2$) and high ($0.6-0.7$) values of IoU_F depending on the specific source and target dataset combination. Nevertheless, we have been able to estimate the correct number of epochs to train the models thanks to the availability of target labels (although they are not used at all during training). In a real scenario, monitoring the proposed average solidity metric is an intuitive and effective way to stop the training process in the absence of validation labels, and select (in average) models of similar or better accuracy. Although other morphological and area measurements were initially tested, the average solidity correlates better with the IoU_F value

of the test labels. Nevertheless, the performance of this metric depends on how close its value is the source and target domains.

It is also interesting to note that TEM and SEM images are different, with TEM images usually having higher resolution. Consequently, Lucchi++ and Kasthuri++ datasets (SEM) are -in principle- in closer domains compared to VNC (TEM) as reflected by the baseline results in Table 1. When Lucchi++ or Kasthuri++ are used as sources, the results obtained with VNC are clearly lower than with Kasthuri++ and Lucchi++, respectively. However, when VNC is used as source, the results obtained with Lucchi++ or Kasthuri++ are similar. As similar discussion is applicable to the figures presented in Section S1: going to lower resolution (i.e., from VNC as a source, to Lucchi++ or Kasthuri++) in principle, could be easier than the opposite (from Lucchi++ or Kasthuri++ as a source, to VNC). Apart from the intrinsic variability due to the modality, we need to acknowledge also the variability due to the differences in the samples itself, their preparation and the acquisition protocol.

In summary, from a practical point of view, the style-transfer approach appears as both the safest and simplest way of addressing the domain shift in EM volumes for semantic segmentation. Nevertheless, using self-supervised or multi-task models may provide better results on specific datasets at the cost of more complex training setups and a larger set of hyperparameters.

The present work is an initial assessment of the three competing approaches running under the same conditions and compared with the same supervised baseline methods. In a future work, we plan to explore the performance of meaningful combinations of the proposed strategies. Namely, the outputs of the style transfer method could be used as inputs or the self-supervised learning and the multi-tasks neural network architectures. We expect the combined strategies to outperform the histogram matching approach.

Moreover, current initiatives (e.g., volume EM, http://www.volumeem.org) are developing massive databases of heterogeneous 3DEM data. These initiatives promise to facilitate deep-learning-based model building for automated segmentation [75]. In our view, the style-transfer strategies could be more effective when pre-trained in massive databases of heterogeneous 3DEM data than in a small dataset of well-defined characteristics.

Finally, it is important to highlight that even the best results among all our proposed domain adaptation strategies lie much lower than the fully supervised approaches. As a reference, the average IoU_F values obtained by our baseline models trained on the target annotated images are 0.9066 for Lucchi++, 0.9154 for Kasthurhi++, and 0.8041 for VNC. This leaves plenty of room for improvement and future lines of research. In particular, we will explore the use of massive databases of heterogeneous 3D EM data, with the combination of some of our proposed strategies and the exploitation of segmentation-specific pretext tasks.

Code Availability

The developed software that support the findings of this study are publicly available at https://github.com/danifranco/EM_domain_adaptation.

Data availability

The Lucchi++ and Kasthuri++ datasets can be downloaded from https://sites.google.com/view/connectomics/. The VNC dataset can be downloaded from https://github.com/unidesigner/groundtruth-drosophila-vnc.

Acknowledgments

I. Arganda-Carreras would like to acknowledge the support of the 2020 Leonardo Grant for Researchers and Cultural Creators, BBVA Foundation. This work is supported in part by the University of the Basque Country UPV/EHU grant GIU19/027 and by Ministerio de Ciencia, Innovación y Universidades, Agencia Estatal de Investigación, under grant PID2019-109820RB-I00, MCIN/AEI /10.13039/501100011033/, cofinanced by European Regional Development Fund (ERDF), "A way of making Europe."

References

1. V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE signal processing magazine*, vol. 32, no. 3, pp. 53–69, 2015. 2
2. M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018. 2
3. G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 5, pp. 1–46, 2020. 2
4. X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, vol. 58, p. 101552, 2019. 2
5. M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018. 2
6. C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert, "Gan augmentation: Augmenting training data using generative adversarial networks," *arXiv preprint arXiv:1810.10863*, 2018. 2
7. A. K. Mondal, J. Dolz, and C. Desrosiers, "Few-shot 3D multi-modal medical image segmentation using generative adversarial learning," *arXiv preprint arXiv:1810.12241*, 2018. 2

8. C. Bermudez, A. J. Plassard, L. T. Davis, A. T. Newton, S. M. Resnick, and B. A. Landman, "Learning implicit brain MRI manifolds with deep learning," in *Medical Imaging 2018: Image Processing*, vol. 10574, p. 105741L, International Society for Optics and Photonics, 2018. 2
9. A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood, "Chest x-ray generation and data augmentation for cardiovascular abnormality classification," in *Medical Imaging 2018: Image Processing*, vol. 10574, p. 105741M, International Society for Optics and Photonics, 2018. 2
10. A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood, "Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation," in *2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018)*, pp. 1038–1042, IEEE, 2018. 2
11. H. Emami, M. Dong, S. P. Nejad-Davarani, and C. K. Glide-Hurst, "Generating synthetic CTs from magnetic resonance images using generative adversarial networks," *Medical physics*, vol. 45, no. 8, pp. 3627–3636, 2018. 2
12. D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with deep convolutional adversarial networks," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 12, pp. 2720–2730, 2018. 2
13. J. Jiang, Y.-C. Hu, N. Tyagi, P. Zhang, A. Rimner, G. S. Mageras, J. O. Deasy, and H. Veeraraghavan, "Tumor-aware, adversarial domain adaptation from CT to MRI for lung cancer segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 777–785, Springer, 2018. 2
14. C.-B. Jin, H. Kim, M. Liu, W. Jung, S. Joo, E. Park, Y. S. Ahn, I. H. Han, J. I. Lee, and X. Cui, "Deep CT to MR synthesis using paired and unpaired data," *Sensors*, vol. 19, no. 10, p. 2361, 2019. 2
15. W. Wei, E. Poirion, B. Bodini, S. Durrleman, N. Ayache, B. Stankoff, and O. Colliot, "Learning myelin content in multiple sclerosis from multimodal MRI through adversarial training," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 514–522, Springer, 2018. 2
16. Y. Pan, M. Liu, C. Lian, T. Zhou, Y. Xia, and D. Shen, "Synthesizing missing PET from MRI with cycle-consistent generative adversarial networks for Alzheimer's disease diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 455–463, Springer, 2018. 2
17. Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pp. 9242–9251, 2018. 2, 4
18. J. T. Guibas, T. S. Virdi, and P. S. Li, "Synthetic medical images from dual generative adversarial networks," *arXiv preprint arXiv:1709.01872*, 2017. 2
19. P. Costa, A. Galdran, M. I. Meyer, M. D. Abramoff, M. Niemeijer, A. M. Mendonça, and A. Campilho, "Towards adversarial retinal image synthesis," *arXiv preprint arXiv:1701.08974*, 2017. 2
20. A. Beers, J. Brown, K. Chang, J. P. Campbell, S. Ostmo, M. F. Chiang, and J. Kalpathy-Cramer, "High-resolution medical image synthesis using progressively grown generative adversarial networks," *arXiv preprint arXiv:1805.03144*, 2018. 2
21. Y. Ma, Y. Hua, H. Deng, T. Song, H. Wang, Z. Xue, H. Cao, R. Ma, and H. Guan, "Self-supervised vessel segmentation via adversarial learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7536–7545, October 2021. 2

22. A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transformations for one-shot medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8543–8553, 2019. 2
23. S. Wang, S. Cao, D. Wei, R. Wang, K. Ma, L. Wang, D. Meng, and Y. Zheng, "Lt-net: Label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9162–9171, 2020. 2
24. D. Tomar, B. Bozorgtabar, M. Lortkipanidze, G. Vray, M. S. Rad, and J.-P. Thiran, "Self-supervised generative style transfer for one-shot medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1998–2008, 2022. 2
25. T. Liu, C. Jones, M. Seyedhosseini, and T. Tasdizen, "A modular hierarchical approach to 3D electron microscopy image segmentation," *Journal of neuroscience methods*, vol. 226, pp. 88–102, 2014. 2
26. A. Fakhry, T. Zeng, and S. Ji, "Residual deconvolutional networks for brain electron microscopy image segmentation," *IEEE transactions on medical imaging*, vol. 36, no. 2, pp. 447–456, 2016. 2
27. C. Xiao, J. Liu, X. Chen, H. Han, C. Shu, and Q. Xie, "Deep contextual residual network for electron microscopy image segmentation in connectomics," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 378–381, IEEE, 2018. 2
28. K. V. Kaltdorf, K. Schulze, F. Helmprobst, P. Kollmannsberger, T. Dandekar, and C. Stigloher, "Fiji macro 3D ART VeSElecT: 3D automated reconstruction tool for vesicle structures of electron tomograms," *PLoS computational biology*, vol. 13, no. 1, p. e1005317, 2017. 2
29. I. Oztel, G. Yolcu, I. Ersoy, T. White, and F. Bunyak, "Mitochondria segmentation in electron microscopy volumes using deep convolutional neural network," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1195–1200, IEEE, 2017. 2
30. V. Casser, K. Kang, H. Pfister, and D. Haehn, "Fast mitochondria detection for connectomics," in *Medical Imaging with Deep Learning*, pp. 111–120, PMLR, 2020. 2, 11
31. A. Khadangi, T. Boudier, and V. Rajagopal, "EM-net: Deep learning for electron microscopy image segmentation," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 31–38, IEEE, 2021. 2
32. D. Franco-Barranco, A. Muñoz-Barrutia, and I. Arganda-Carreras, "Stable deep neural network architectures for mitochondria segmentation on electron microscopy volumes," *Neuroinformatics*, Dec 2021. 2, 5, 6, 10, 15
33. L. Heinrich, D. Bennett, D. Ackerman, W. Park, J. Bogovic, N. Eckstein, A. Petruncio, J. Clements, S. Pang, C. S. Xu, *et al.*, "Whole-cell organelle segmentation in volume electron microscopy," *Nature*, vol. 599, no. 7883, pp. 141–146, 2021. 2
34. R. Bermúdez-Chacón, P. Márquez-Neila, M. Salzmann, and P. Fua, "A domain-adaptive two-stream U-net for electron microscopy image segmentation," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 400–404, IEEE, 2018. 2
35. J. Roels, J. Hennies, Y. Saeys, W. Philips, and A. Kreshuk, "Domain adaptive segmentation in volume electron microscopy imaging," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1519–1522, IEEE, 2019. 2, 4, 10

36. J. Peng, J. Yi, and Z. Yuan, "Unsupervised mitochondria segmentation in EM images via domain adaptive multi-task learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1199–1209, 2020. 2, 5, 13, 14, 15
37. H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173–1185, 2022. 3
38. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017. 3
39. H. Yang, J. Sun, A. Carass, C. Zhao, J. Lee, Z. Xu, and J. Prince, "Unpaired brain MR-to-CT synthesis using a structure-constrained CycleGAN," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 174–182, Springer, 2018. 4
40. Y. Huo, Z. Xu, H. Moon, S. Bao, A. Assad, T. K. Moyo, M. R. Savona, R. G. Abramson, and B. A. Landman, "Synseg-net: Synthetic segmentation without target modality ground truth," *IEEE transactions on medical imaging*, vol. 38, no. 4, pp. 1016–1025, 2018. 4
41. Q. Dou, C. Ouyang, C. Chen, H. Chen, and P.-A. Heng, "Unsupervised cross-modality domain adaptation of CONVNETs for biomedical image segmentations with adversarial loss," *arXiv preprint arXiv:1804.10916*, 2018. 4
42. H. Cho, S. Lim, G. Choi, and H. Min, "Neural stain-style transfer learning using GAN for histopathological images," *arXiv preprint arXiv:1710.08543*, 2017. 4
43. T. d. Bel, M. Hermsen, J. Kers, J. v. d. Laak, and G. Litjens, "Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology," in *International Conference on Medical Imaging with Deep Learning – Full Paper Track*, (London, United Kingdom), 08–10 Jul 2019. 4
44. M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni, "Staingan: Stain style transfer for digital histological images," in *2019 Ieee 16th international symposium on biomedical imaging (Isbi 2019)*, pp. 953–956, IEEE, 2019. 4
45. F. Mahmood, R. Chen, and N. J. Durr, "Unsupervised reverse domain adaptation for synthetic medical images via adversarial training," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2572–2581, 2018. 4
46. D. Wang, T. Zhao, N. Zheng, and Z. Gong, "Two-stage generative models of simulating training data at the voxel level for large-scale microscopy bioimage segmentation.," in *IJCAI*, pp. 4781–4787, 2019. 4
47. S. Kim, B. Kim, and H. Park, "Synthesis of brain tumor multicontrast MR images for improved data augmentation," *Medical Physics*, vol. 48, no. 5, pp. 2185–2198, 2021. 4
48. T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020. 4
49. T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *European Conference on Computer Vision*, pp. 319–345, Springer, 2020. 4, 6, 7, 15
50. I. Katircioglu, H. Rhodin, V. Constantin, J. Spörri, M. Salzmann, and P. Fua, "Self-supervised segmentation via background inpainting," *arXiv*, pp. 1–12, 2020. 4
51. S. Jenni, H. Jin, and P. Favaro, "Steering self-supervised feature learning beyond local pixel statistics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6408–6417, 2020. 4

52. S. Lee, D. Cho, J. Kim, and T. H. Kim, "Self-supervised fast adaptation for denoising via meta-learning," *arXiv preprint arXiv:2001.02899*, 2020. 4
53. V. Dewil, J. Anger, A. Davy, T. Ehret, G. Facciolo, and P. Arias, "Self-supervised training for blind multi-frame video denoising," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2724–2734, 2021. 4
54. S. Laine, T. Karras, J. Lehtinen, and T. Aila, "High-quality self-supervised deep image denoising," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019. 4
55. X. Li, G. Zhang, J. Wu, Y. Zhang, Z. Zhao, X. Lin, H. Qiao, H. Xie, H. Wang, L. Fang, *et al.*, "Reinforcing neuron extraction and spike inference in calcium imaging using deep self-supervised denoising," *Nature Methods*, vol. 18, no. 11, pp. 1395–1400, 2021. 4
56. M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Computer Vision – ECCV 2016* (B. Leibe, J. Matas, N. Sebe, and M. Welling, eds.), (Cham), pp. 69–84, Springer International Publishing, 2016. 4
57. Y. Li, J. Chen, X. Xie, K. Ma, and Y. Zheng, "Self-Loop Uncertainty: A Novel Pseudo-Label for Semi-supervised Medical Image Segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, eds.), (Cham), pp. 614–623, Springer International Publishing, 2020. 4
58. A. Taleb, C. Lippert, T. Klein, and M. Nabi, "Multimodal self-supervised learning for medical image analysis," in *International Conference on Information Processing in Medical Imaging*, pp. 661–673, Springer, 2021. 4
59. J. Jiao, R. Droste, L. Drukker, A. T. Papageorghiou, and J. A. Noble, "Self-supervised representation learning for ultrasound video," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1847–1850, IEEE, 2020. 4
60. A. Krull, T. Vičar, M. Prakash, M. Lalit, and F. Jug, "Probabilistic Noise2Void: Unsupervised Content-Aware Denoising," *Frontiers in Computer Science*, vol. 2, p. 5, feb 2020. 4
61. T.-O. Buchholz, M. Prakash, A. Krull, and F. Jug, "DenoiSeg: Joint Denoising and Segmentation," tech. rep. 4
62. M. Prakash, T.-O. Buchholz, M. Lalit, P. Tomancak, F. Jug, and A. Krull, "Leveraging self-supervised denoising for image segmentation," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 428–432, IEEE, 2020. 4
63. V. Alex, K. Vaidhya, S. Thirunavukkarasu, K. Chandrasekharan, and G. Krishnmurthi, "Semi-supervised Learning using Denoising Autoencoders for Brain Lesion Detection and Segmentation," tech. rep., 2017. 4
64. A. Taleb, C. Lippert, T. Klein, and M. Nabi, "Multimodal self-supervised learning for medical image analysis," in *Information Processing in Medical Imaging* (A. Feragen, S. Sommer, J. Schnabel, and M. Nielsen, eds.), (Cham), pp. 661–673, Springer International Publishing, 2021. 4
65. A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, and C. Lippert, "3d self-supervised methods for medical imaging," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18158–18172, 2020. 4

66. L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Medical Image Analysis*, vol. 58, p. 101539, 2019. 4
67. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, 2015. 4, 5
68. J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, pp. 197–207, 2019. 5
69. R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Upper Saddle River, NJ: Pearson, 3 ed., 2008. 5
70. L. N. Smith and N. Topin, "Super-convergence: very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications* (T. Pham, ed.), vol. 11006, pp. 369 – 386, International Society for Optics and Photonics, SPIE, 2019. 8, 11
71. A. Lucchi, K. Smith, R. Achanta, G. Knott, and P. Fua, "Supervoxel-based segmentation of mitochondria in EM image stacks with learned shape features," *IEEE Transactions on Medical Imaging*, vol. 31, no. 2, pp. 474–486, 2011. 11
72. N. Kasthuri, K. J. Hayworth, D. R. Berger, R. L. Schalek, J. A. Conchello, S. Knowles-Barley, D. Lee, A. Vázquez-Reina, V. Kaynig, T. R. Jones, *et al.*, "Saturated reconstruction of a volume of neocortex," *Cell*, vol. 162, no. 3, pp. 648–661, 2015. 11
73. S. Gerhard, J. Funke, J. Martel, A. Cardona, and R. Fetter, "Segmented anisotropic ssTEM dataset of neural tissue," *figshare*, pp. 0–0, 2013. 12
74. K. Zuiderveld, "Contrast limited adaptive histogram equalization," *Graphics gems*, pp. 474–485, 1994. 13, 15
75. R. Conrad and K. Narayan, "CEM500K, a large-scale heterogeneous unlabeled cellular electron microscopy image dataset for deep learning," *Elife*, vol. 10, p. e65894, 2021. 17