Adversarial Mutual Leakage Network for Cell Image Segmentation

Hiroki Tsuda  Kazuhiro Hotta
Meijo University
1-501 Shiogamaguchi, Tempaku-ku, Nagoya 468-8502, Japan
193427019@ccalumni.meijo-u.ac.jp, kazuhotta@meijo-u.ac.jp

Abstract

We propose three segmentation methods using GAN and information leakage between generator and discriminator. First, we propose an Adversarial Training Attention Module (ATA-Module) that uses an attention mechanism from the discriminator to the generator to enhance and leak important information in the discriminator. ATA-Module transmits important information to the generator from the discriminator. Second, we propose a Top-Down Pixel-wise Difficulty Attention Module (Top-Down PDA-Module) that leaks an attention map based on pixel-wise difficulty in the generator to the discriminator. The generator trains to focus on pixel-wise difficulty, and the discriminator uses the difficulty information leaked from the generator for classification. Finally, we propose an Adversarial Mutual Leakage Network (AML-Net) that mutually leaks the information each other between the generator and the discriminator. By using the information of the other network, it is able to train more efficiently than ordinary segmentation models. Three proposed methods have been evaluated on two datasets for cell image segmentation. The experimental results show that the segmentation accuracy of AML-Net was much improved in comparison with conventional methods.

1. Introduction

Automated cell image segmentation has been widely studied due to the large number of cell images and the tedious task of obtaining dense annotations. Overall time and cost savings are expected to be achieved by automated processing without human involvement. Manual segmentation is slow and burdensome to process, and thus there is a significant demand for algorithms that can perform segmentation quickly and accurately without human intervention. However, cell image segmentation is a difficult task because the number of supervised images is smaller and there is not regularity compared to the other datasets such as automatic driving. A large number of supervised images require manual labeling which take a lot of effort and time. Therefore, it is necessary to improve the segmentation ability for pixel-level recognition with small number of training images.

By the advent of Convolutional Neural Networks (CNNs), near-human level performance can be achieved in medical image analysis tasks such as blood vessel extraction from fundus images [11, 24, 33], cancerous lung nodule detection [16], and cell image segmentation [14, 45]. CNNs have become the de facto standard in the field of image recognition, Fully Convolutional Networks (FCNs) and U-Net [28] are the commonly used for segmentation.

Generative Adversarial Network (GAN) [7] is for image generation. It preforms adversarial training between generator and discriminator. The discriminator enchants the difference between real and generated images. The adversarial training allows to generate realistic images. In addition, pix2pix [15] is the extended version of GAN can train image-to-image translation. In particular, pix2pix is also effective for semantic segmentation tasks [20].

We focused on the relationship between the generator and the discriminator. We consider that adversarial training could be more efficient by mutually leaking important information from the generator and the discriminator. In this paper, we propose Adversarial Mutual Leakage Network (AML-Net) which consists of two new attention modules. The first one is Adversarial Training Attention Module (ATA-Module) that creates an attention map from the feature map in discriminator and leaks it to the generator. The second one is Top-Down Pixel-wise Difficulty Attention (Top-Down PDA) Module that creates an attention map based on pixel-wise difficulty of generator and leaks it to the discriminator. AML-Net combines these two attention mechanisms and leaks them to each other. Figure 1 shows the overview of AML-Net, which aims to improve the performance of generator and discriminator by mutually leaking information through two attention mechanisms.

In experiments on two kinds of cell image datasets [6, 45], we evaluate the proposed method. We confirmed that our method gave higher accuracy than conventional methods without information leakage.
2. Related works

2.1. Semantic Segmentation

FCNs based methods have achieved significant results for semantic segmentation. The original FCN [19] finally created high-dimensional feature map with low-resolution. This feature map has semantic information but fine information such as small objects and correct location are lost. Thus, if upsampling is used at the final layer, the accuracy is not sufficient. SegNet [1], U-Net [28] and RefineNet [17] obtained feature map with high-resolution. They are designed to gradually recover the spatial information by decoder to combine various contextual information extracted by encoder. In addition, Attention U-Net [23] introduces attention to U-Net. By adding a per-pixel attention gate similar to the sSE block [29] to the skip connection, it improves the segmentation accuracy. We improved the segmentation accuracy by leaking the information through attention map between the generator and the discriminator.

Alternatively, DeepLabv3+ [3] removes the last two downsampling processes from ResNet [8] and introduces dilated (atrous) convolutions [42] to maintain the receptive field. As a result, it can hold location information with sufficient size. Furthermore, DeepLabv3+ [3] combined the advantages of the encoder-decoder structure of DeepLabv3 [2] to reduce the computational cost. In addition, FastFCN [40] used the original ResNet in encoder and adopted Joint Pyramid Upsampling to reduce the computational cost without degrading the performance.

Increasing the receptive field without decreasing the resolution by dilated convolution is that the resolution of the feature map is relatively large. This increases the computational cost, and the processing time for training and inference is long. Therefore, we studied the method for training important features based on U-Net, which is easy to use with simple computational resources and is widely used in medical and biological imaging.

2.2. Adversarial Training

GANs [7] has achieved success in image generation tasks, including image-to-image translation [15, 18, 25, 44, 41, 46], domain adaptation [4, 10, 20, 35] and text-to-image synthesis [27, 44]. Adversarial training improved image generation by training generator and discriminator to compete with each other. The discriminator trains to classify whether an input is a real or a generated image. On the other hand, the generator trains so that the generated image and the real image are not classified by the discriminator. Adversarial training network gave superior result on segmentation in comparison with non-adversarial deep networks [20].

Sawada et al. [31] has proposed a method to train more efficiently using the framework of adversarial training. By sending the feature maps in discriminator to generator, the segmentation accuracy of generator was improved. They concatenated the feature maps in discriminator to generator because the feature maps in discriminator include the difference between generated results and ground truth. However, the training parameters of the method are very large due to the usage of multiple generators and discriminators. In addition, since they did not select the information from discriminator, the effect of leakage is not fully exploited. We extract only the important information from the feature maps in generator and discriminator, and leak them each other between the generator and discriminator.

2.3. Attention Mechanism

Attention mechanism is used in computer vision and natural language processing. In image recognition, important parts or channels are enhanced. Residual Attention Network [37] introduced a stacked network structure composed of multiple attention components, and attention residual training used residual training [8] in attention mechanism. Squeeze-and-Excitation Network (SENet) [12] introduced an attention mechanism that enhances the channels of feature maps. Accuracy booster blocks [32] and efficient channel attention module [38] made further improvements by changing the fully-connected layer in SENet [12]. Transformer [36] performed the translation task only with the attention mechanism in the natural language processing. There are Self-Attention that uses the same tensor, and Source-Target-Attention that uses two different tensors. Several networks have been proposed that use Self-Attention to train the similarity between pixels in feature maps [5, 13, 26, 39, 43]. Pixel-wise Difficulty Attention (PDA) Module [21] is an attention mechanism that uses the confidence at each pixel as difficulty level. It trains efficiently by reinforcing pixels with low confidence as high difficulty. However, PDA-Module has a problem that the difficulty level becomes low when mis-classification occurs with high confidence.

In this paper, we have designed two kinds of atten-
tion mechanisms for better information leakage. The discriminator-to-generator attention mechanism is designed based on self-attention, and we aim to select effective information to the generator from the discriminator. The generator-to-discriminator attention mechanism creates and leaks an attention map with the probability of the correct class as the difficulty level by referring to ground truth. This attention mechanism is used in only training. Since pixel-wise difficulty becomes small through training, we do not need to use it in test phase.

3. Proposed Method

We present two attention mechanisms that leak information: the first mechanism leaks the information from the discriminator to the generator, and the second one leaks the information from the generator to the discriminator. Finally, we combine the two attention mechanisms based on information leakage between the generator and discriminator.

3.1. Adversarial Training Attention Module

Conventional method leaks from the first discriminator to the second generator to improve accuracy [31], but it uses multiple generators, and the feature map in the discriminator is just concatenated to the feature map in the second generator. The good information for improving the generator is not extracted. In addition, the number of parameters increases in the method due to more than two generators. In order to increase the effect of leakage from the discriminator to the generator, we enhance the important pixels from the discriminator’s feature map and then transfer them to the generator for efficient training. We propose an Adversarial Training Attention Module (ATA-Module) that enhances the important pixels of the feature map in the discriminator and then leaks them to the generator.

The discriminator determines whether the generated image is real or fake. The feature map in the discriminator has the information about the difference between the generated image and ground truth. The feature map in the discriminator contains the information on what is wrong with the generated image when it is judged as fake. The segmentation accuracy can be improved by ATA-Module using the information that enhances the evidence of being a fake as determined by the discriminator.

Adversarial loss is one of the influences from the discriminator to the generator, but it is not enough to transfer the information in discriminator to the generator. This is because adversarial loss only transmits the result of whether the discriminator was deceived or not to the generator as a loss function. It is not possible to know the regions that the discriminator was not deceived. By using the ATA-Module, the pixels that were not deceived can be transmitted to the generator in a precise and efficient manner.

As shown in Figure 2, we feed the feature maps in discriminator into \(1 \times 1\) convolution layers to generate new feature maps \(Query\) and \(Key\), respectively. We are inspired by Self-Attention GAN (SAGAN) [43] to reduce the channel number to \(C/8\) for memory efficiency. Then, we reshape them to \(C/8 \times (H \times W)\). After we perform a matrix multiplication between the transpose of \(Query\) and \(Key\), and we use a softmax function to calculate an attention map as

\[
w_{ij} = \frac{\exp(\langle Query^T, Key_j \rangle)}{\sum_{j=1}^{H \times W} \exp(\langle Query^T, Key_j \rangle)}, \tag{1}
\]

where \(w_{ij}\) measures the \(j^{th}\) \(Query\)'s impact on \(i^{th}\) \(Key\). If two pixels belong to the same class, the feature representation will be similar. \(H \times W\) is the total number of pixels.

Meanwhile, we feed the feature map in discriminator into \(1 \times 1\) convolution layer to generate a new feature map \(Value\) and reshape it to \(C/2 \times (H \times W)\). Then, we perform a matrix multiplication between the attention map and the transpose of \(Value\) and reshape the result to \(C/2 \times H \times W\), and then to \(C \times H \times W\) by \(1 \times 1\) convolution. Finally, we multiply it by a scale parameter \(\alpha\) and perform a element-wise sum operation with the feature maps in generator to obtain the final output as

\[
S_i = \alpha \sum_{j=1}^{H \times W} (w_{ij} Value_j^T)^T + F_i, \tag{2}
\]

where \(\alpha\) is initialized as 0 and gradually trains to assign more weight [43]. \(S_i\) indicates the output and \(F_i\) indicates the feature map of the generator.

Equation 2 means that the output \(S_i\) is the weighted sum of all positions in the discriminator. Therefore, the difference between the real and fake extracted by discriminator is aggregated by the similarity of each pixel. The segmentation accuracy is improved by transmitting the information of discriminator to the generator.

3.2. Top-Down PDA Module

The discriminator trains the difference between the generated image and ground truth to prevent the generator from deceiving the generated image as the real one. By leaking
Figure 3: Network structure of Top-Down PDA-Module. The predicted probability of true class at the decoder is subtracted from 1 to form an attention map. Pixels with low probability have high attention values. Attention map is used for both of generator and discriminator.

The information about the difficult pixels from the generator to the discriminator, the discriminator can use the information of the difficult pixels in the generator as a reason to judge the difference between the generated image and ground truth, thereby improving the performance of the discriminator. As the performance of the discriminator is improved, the performance of the generator is also improved by training the generator to fool the discriminator by adversarial loss. Therefore, we propose the Top-Down Pixel-wise Difficulty Attention Module (Top-Down PDA-Module), which creates an attention map that enhances the pixels with high difficulty in the generator and leaks the attention map to the discriminator. Top-Down PDA-Module is a top-down attention mechanism that refers to ground truth, so it works only during training. Since we cannot use to ground truth during inference, we do not use Top-Down PDA-Module in test phase. This is not a problem because the pixel-wise difficulty is close to 0 throughout training.

In semantic segmentation, the output layer compresses the number of channels of the feature map into the number of classes, and converts them into probability for each class by a softmax function. In this case, the pixels with low probability for true class can be considered difficult. This information is explicitly given to the discriminator as an attention map, and it can be used as a reason to judge the difference between the generated image and the ground truth, so that the discriminator can accurately classify.

As show in Figure 3, Top-Down PDA-Module performs segmentation at each resolution of the decoder in the generator in order to enhance the pixels with high difficulty in classification based on the probability. This makes it possible to calculate the difficulty level for each pixel at each resolution. For a feature map $A \in \mathbb{R}^{C \times H \times W}$ in the decoder, the number of channels is compressed to $B \in \mathbb{R}^{C/2 \times H/2 \times W/2}$ by $1 \times 1$ convolution, Batch Normalization and ReLU function. Then we compress the channel dimension to the number of classes by $1 \times 1$ convolution, and we perform upsampling to the original input image size and use the softmax function to obtain the probability values for all pixels. From this probability map, we pick up the probability of true class at each pixel to create a map $C \in \mathbb{R}^{H \times W}$. By subtracting the map from 1 (true probability) and downsampling to the original feature map $C' \in \mathbb{R}^{H \times W}$, we can create an attention map with high values for difficult pixels.

Attention Map created from the correct answer probability map is multiplied by the feature map from the decoder to enhance the pixels with high difficulty. Furthermore, by performing a residual addition with the feature map from the decoder, the gradient does not disappear and the magnitude of the value does not change drastically between training and inference. In addition, by calculating the loss function using the probability map in the middle of Top-Down PDA-Module, the training process can be optimized by back propagation of error like auxiliary loss from the middle output. Since the size of the probability map is the same as that of the input image, the training process is close to the final output even with low-resolution feature maps. By using the segmentation output in the middle layer of the decoder, the Top-Down PDA-Module is able to create an attention map that focuses on the difficulty of pixels. Softmax cross-entropy is used as a loss function.

The decoder’s feature map $A$ in the generator can be multiplied by the attention map, and $A$ is enhanced based on the difficulty level by the residual addition. In addition, we also leak the created attention map to the discriminator. Thus, the attention map is used for both of generator and discriminator. In conventional adversarial training, the difference between the generated image and ground truth, which is the reason for the discriminator’s decision, is acquired bottom-up manner during the training process, so that unintended pixels with low importance may become the basis for the decision. If the discriminator clearly trains the difference between the generated image and ground truth based on the attention map and the judged results are used for training the generator, the accuracy can be improved from the case of the generator alone. Attention Map is multiplied with the feature map at the same resolution in the discriminator, and the pixels with high difficulty by the generator are enhanced. Furthermore, by performing a residual addition, the discriminator can leak the information that the generator is training to be important.

3.3. Adversarial Mutual Leakage Network

The performance of the generator is improved by using our proposed two modules; the ATA-Module which leaks from the discriminator to the generator and the Top-Down PDA-Module which leaks from the generator to the discriminator. The performance of the generator can be further im-
Figure 4: **Network structure of Adversarial Mutual Network.** First, the generator segments an image without attention mechanism. The segmentation result is fed into the discriminator. An attention map is created by ATA-Module from the discriminator’s feature map, and it is used in the second generator. Next, an attention map is created from the decoder in the generator by Top-Down PDA-Module, and it is leaked to the generator and the discriminator. Finally, the discriminator judges true or false. The first and second generators and discriminators share the same weights except for the attention mechanism.

Whole structure is shown in Figure 4. The generator is based on U-Net and the discriminator is a CNN with 6 convolutional layers. AML-Net performs segmentation twice using the generators with shared weights for ATA-Module, which uses the feature map of the discriminator. First, the segmentation is done by the generator without ATA-Module and Top-Down PDA-Module, and then the result is fed into the discriminator to classify real or fake. The feature map in the discriminator is obtained to use for ATA-Module. The attention map made from feature map in the discriminator at each resolution in discriminating the generated image at the first round is sent to the encoder at the same resolution of the generator at the second round through ATA-Module. This allows us to generate an image that is closer to ground truth than the generated image at the first round.

Next, the decoder part of each resolution at the second round generator uses the Top-Down PDA-Module to create an attention map based on the difficulty level in the generator. This allows the generator to accurately train the pixels of high difficulty that are difficult to predict using explicit information from ground truth. The created attention map is also leaked to the discriminator to improve the discrimination performance between the generated image and ground truth, and the generator trains to successfully deceive the discriminator by adversarial loss.

Discriminator in AML-Net has an encoder structure and extracts features from the input image paired with the generated image or ground truth by $3 \times 3$ convolution. Next, we perform downsampling by convolution with a stride number of 2 using a kernel of $4 \times 4$ at each resolution. In this process, we use Spectral Normalization [22], which is a GAN training stabilization method. We apply Leaky ReLU to the obtained feature map, and extract features by multiplication and residual addition with the attention map.

proved by using both of our proposed modules. Finally, we propose Adversarial Mutual Leakage Network (AML-Net), which combines ATA-Module and Top-Down PDA-Module to leak the information in both directions.
leaked from the generator. In the final layer, a $1 \times 1$ convolution is used to perform binary classification so that the output is true for ground truth and fake for the generated image. In addition, by referring to the Patch GAN proposed by pix2pix \cite{isola2017image}, the discriminator uses $16 \times 16$ pixels of the input image as a patch to determine whether the image is real or fake. This improves the generator based on local judgments in the cropped patch instead of global judgments for a single image. AML-Net trains the generator with the loss function used for segmentation and the adversarial loss function. In final output of the U-Net used in the generator, the loss ($L_{CE}$) is calculated using softmax cross-entropy. For Top-Down PDA-Module at each of the three resolutions included in the generator, softmax cross-entropy is also used to calculate the respective losses ($L_{PDA1}$, $L_{PDA2}$, and $L_{PDA3}$). In addition, by using adversarial loss ($L_{adv}$) to determine whether the generator deceives the discriminator, the generator is trained by considering the reality that cannot be determined only by the loss of the generator ($L_{CE}$). Adversarial loss $L_{adv}$ is given as

$$L_{adv}(G, D) = \mathbb{E}_{x,y \sim p_{data}(x,y)} \left[ \log D(x, y) \right] + \mathbb{E}_{x \sim p_{data}(x)} \left[ \log (1 - D(x, G(x))) \right].$$

(3)

There are two inputs in Equation(3) $x$ represents the input image and $y$ represents ground truth. In addition, $p_{data}$ represents the distribution of training data, and $x, y$ $p_{data}(x,y)$ is the process of sampling data from $p_{data}$. Let $G(x)$ denote the segmentation output by the generator, and $D(\cdot)$ denote the result output by the discriminator, which determines whether the data is real or fake. Based on this equation, adversarial training is optimized with the following Min-Max objective function. $G(x)$ denotes the segmentation result output by the generator, and $D(\cdot)$ denotes the discriminant result output by the discriminator. From Equation(3) adversarial training is optimized with the Min-Max objective function as

$$\text{arg}\min_G \text{max}_DL_{adv}(G, D).$$

(4)

The generator $G$ minimizes $L_{adv}$ so that the segmentation image $G(x)$ deceives the discriminator, and the discriminator $D$ maximizes $L_{adv}$ so that the output $G(x)$ of the generator is fake and the ground truth $y$ is true. Especially, the adversarial loss for the generator is as follows:

$$L_{adv}(G) = \mathbb{E}_{x \sim p_{data}(x)} \left[ \log (1 - D(x, G(x))) \right].$$

(5)

The final loss function of the generator is as

$$L_{total} = L_{CE} + L_{PDA1} + L_{PDA2} + L_{PDA3} + \lambda_{adv} L_{adv}(G),$$

(6)

where $\lambda_{adv}$ is a hyperparameter, which is set weakly to prevent the discriminator from becoming too strong. In this paper, we set $\lambda_{adv} = 0.01$ as same as pix2pix.

4. Experiments

In this section, we first introduce the implementation details and the dataset used in our experiments. Next, we evaluated our method on two cell image datasets.

4.1. Implementation Details

We adopted Adam as the Optimizer in our experiments. The training coefficients were set to $\alpha = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.9$, and $\varepsilon = 10^{-8}$. A single GeForce RTX 2080 Ti GPU was used for training in this experiment. 5-fold cross-validation were conducted to properly evaluate the generalization performance. The model given the highest mean Intersection of Union (mIoU) for the validation images was used for the evaluation. Since the experimental results are changed according to the initial seed of a random function, the same experiment was conducted three times with different initial values, and the results were evaluated based on the average of 15 times in total.

4.2. Dataset

4.2.1 White Blood Cell

White Blood Cell(WBC) \cite{45} consists of 100 images of $300 \times 300$ pixels in three classes: cell nucleus of white blood cells, cytoplasm, and background including red blood cells. For the experiment, the images were resized to $320 \times 320$ pixels by bilinear interpolation. 64 images were training images, 16 images were validation images, and 20 images were test images. The batch size is set to 2 for training. In the segmentation images, white color indicates cell nucleus, gray color indicates cytoplasm, and black color indicates the background including red blood cells.

4.2.2 Drosophila Cell Image

Drosophila cell image \cite{6} consists of 20 images including 4 classes of cell membrane, cytoplasm, mitochondria, and synapses with an image size of $1024 \times 1024$ pixels. In experiment, we divided the images into 12 training images, 3 validation images, and 5 test images, and divided the original image size into 16 regions of $256 \times 256$ pixels without overlap. The final number of images was 192 for training, 48 for validation, and 60 for evaluation. The batch size is set to 4 for training. In the segmentation images, red indicates cell membrane, black indicates cell membrane, green indicates mitochondria, and blue indicates synapses.

4.3. Experiments Results

4.3.1 Results on WBC dataset

We use U-Net \cite{28} as a baseline, which has been widely applied in medical and biological images. Following methods are the comparison methods. two adversarial training.
methods; pix2pix \[15\] based on U-Net with a discriminator and Self-Attention GAN (SAGAN) \[43\] which adds Self-Attention \[36\] to the discriminator of pix2pix. We also evaluate attention U-Net \[23\] introduced an attention mechanism called sSE-Block \[29\] to Skip Connection of U-Net. PDA-Module \[21\] is a method that applies a bottom-up attention mechanism based on difficulty to U-Net.

Table 1 shows the IoU accuracy on the WBC dataset, and Table 2 shows the experimental results of evaluating WBC dataset with Precision and Recall. Bold red letters in Table represent the best accuracy, and blue and green bold letters are the second and third best. We see that the accuracy of AML-Net is better than that of conventional methods based on U-Net. In particular, we confirm that the accuracies of Top-Down PDA-Module and AML-Net were improved in almost classes compared to pix2pix and SAGAN, which are adversarial training methods without leakage. Especially, in cytoplasm which is the most difficult, large accuracy improvement was observed. In addition, the accuracy of AML-Net was better than other conventional methods in terms of the mean and standard deviation. This indicates that the accuracy of AML-Net is stable and does not depend on the combination of datasets or initial values. In addition, we confirm that the mean IoU of AML-Net based on U-Net is higher than that of Deeplabv3+ \[3\] and FastFCN \[40\] based on ResNet-50 \[8\]. By avoiding the use of very deep networks such as ResNet-50, the training and inference speed can be accelerated, which is an advantage for cell biologists in practical use. From Table 2, we can see that AML-Net has the best Recall for cytoplasm and cell nucleus, and Precision also shows high accuracy. Conventional methods only have high precision or high recall, but
AML-Net has an advantage in the balance of precision and recall. Figure 5 shows segmentation results. We see that the most of conventional methods in the upper image group over-detect the cytoplasm because the color of red blood cells and cytoplasm are similar. In addition, in the lower image group, we can see that the cytoplasm staining is darker and therefore closer to the color of the cell nucleus, and that segmentation is not working well with many conventional methods. Our AML-Net can accurately segment the regions even for images with very similar colors.

### 4.3.2 Results on Drosophila Cell Images

As shown in Table 3, the proposed method improved the accuracy compared to the conventional methods in many classes. In particular, our proposed AML-Net and Top-Down PDA-Module improved the accuracy of synapses, which is the most difficult class. Top-Down PDA-Module is a top-down attention mechanism using ground-truth, and it is considered that the accuracy is improved due to its ability to create attention maps that explicitly strengthen particularly difficult synapses. In addition, the accuracy of A-ATA-Module was improved in many classes compared to conventional methods. From the feature maps obtained from the discriminator, it can be concluded that the A-ATA-Module contributes to the accuracy improvement because the similarity between pixels is appropriately enhanced by the A-ATA-Module. Therefore, AML-Net with the appropriate combination of Top-Down PDA-Module and A-ATA-Module can improve the accuracy of IoU for many classes. DeepLabv3+ with ResNet-50 as its backbone does not train well on Drosophila cell images, indicating that even successful models for scene segmentation are not effective for cell image segmentation. In addition, AML-Net has better IoU accuracy than FastFCN using ResNet-50.

The top image group in Figure 6 shows that Top-Down PDA-Module and AML-Net can accurately detect synapses that are easily over-detected by conventional methods. However, our method also caused excessive false positives for cell membrane and mitochondria on the right side. The false positives were probably caused by the fact that the input image shows something very similar to cell membrane and mitochondria. In the lower group of images, U-Net and Top-Down PDA-Module fail to detect mitochondria, and misidentify them as mitochondria in some cell membranes, while AML-Net correctly identifies mitochondria, reducing the number of undetected or false positives. Thus, we can see that AML-Net is able to recognize mitochondria by utilizing the advantages of both A-ATA-Module and Top-Down PDA-Module.

### 4.3.3 Visualization of ATA-Module

The attention maps of ATA-Module in AML-Net are visualized in Figure 7. When we select a cytoplasmic pixel as a reference, the cytoplasmic regions with high similarity are highlighted in red as shown in Figure 7B. As shown in Figure 7C the attention map will respond only to cell nucleus and the other classes are blue with low similarity when cell nucleus is selected as a reference. In addition, as shown in Figure 7D when a pixel of an red blood cell which is a part of the background pixel is selected, only the similar red blood cells are reacted to not the entire background. Therefore, in Figure 7E it can be confirmed that red blood cells are not reacted when the pixels of the background other than...
red blood cells are selected. These results show that ATA-Module can enhance the similarity not only among the three classes; cytoplasm, cell nucleus and background, but also between red blood cells and other parts of the background. Based on these attention maps created from the discriminator, efficient leakage to the generator is performed.

4.3.4 Visualization of Top-Down PDA-Module

Visualization results of the attention map in Top-Down PDA-Module are shown in Figure 8. In the first epoch, the probability of the correct class is low at almost pixels. In the second epoch, the probability of the correct class is low for the cell nucleus and red blood cells in the background. As training advances, we can see that the probability of the correct class is low at the edges and in small regions. This indicates that these pixels are more difficult to recognize than the other pixels. Since the pixels with relatively low probability for the correct class contain a lot of information for identifying the object’s outline, they are considered to be of relatively high importance during training.

4.4. Ablation studies

To show the effectiveness of the proposed method, we perform various ablation studies on WBC dataset [45].

4.4.1 Ablation studies on ATA-Module

At first, we compare different connection methods in ATA-Module, which leaks the information from the discriminator to the generator. Comparison methods are as follows. "Add" is the method which adds feature maps in the discriminator to the generator. "1 × 1 Conv" is the method which uses 1 × 1 convolution and then add it to the feature map of generator. In addition, we also compare the method using some attention mechanisms; SE block[12](SE block), Light Attention[9] which enhances in spatial and channel by convolution, and Top-Down PDA-Module which is applied to the discriminator and leaks to the generator (Top-Down PDA-Module). ATA-Module is also compared with a method that creates an Attention Map from the feature maps of the generator and discriminator(Source-Target-Attention).

Table 4 shows the results while changing the connection in ATA-Module. We see that the leakage from the discrim-
inator to the generator by our ATA-Module achieved the highest accuracy. This indicates that our proposed method is the most effective for enhancing the important information from the discriminator and transmitting it to the generator. Other attention methods enhance the range of the channel or receptive field, so they cannot take into account the detailed relationship between pixels as in A-Module. In addition, Sawada’s method using multiple generators cannot be compared fairly because it requires at least two generators. Therefore, we compare our A-Module with Sawada’s method (Concat) under the same condition using two generators and one discriminator. A-Module using Generator2 and Discriminator1 in Table 4 shows better accuracy. The effectiveness of the A-Module is also confirmed in the case of multiple generators.

Table 5 shows the results of changing the connection of A-Module included in AML-Net. We can see that the accuracy of our AML-Net is the highest. This indicates that the A-Module is also highly effective in the connection mechanism from the discriminator to the generator in adversarial mutual leakage. In addition, not only the average IoU accuracy of each class is superior, but also the standard deviation is lower than that of other methods, indicating that the accuracy is more stable and higher than that of conventional methods.

### 4.4.2 Ablation studies on Top-Down PDA-Module

We evaluate different connection in Top-Down PDA-module. Table 6 shows that our Top-Down PDA-Module has the best leakage from the generator to the discriminator. Top-Down PDA-Module and PDA-Module have only one Attention Map. Therefore, the information to be transmitted can be compressed based on the difficulty level, and only important pixels can be enhanced. In other methods, attention maps are created between channels and pixels to enhance the feature map in the generator before leakage. The accuracy of Top-Down PDA-Module and the PDA-Module differs greatly, and pixels that is recognized incorrectly with high confidence by the PDA-Module are judged to be low difficulty by the PDA-Module and remains incorrect. The Top-Down PDA-Module can accurately judge the difficulty as high even if the pixel is recognized incorrectly with a high probability, resulting in higher accuracy.

In addition, the accuracy of changing the Top-Down PDA-Module included in AML-Net is shown in Table 7. "Concat" uses two discriminators because the number of feature maps changes between the first and the second time when the generator leaks to the discriminator. From Table 7, we can see that our AML-Net has the highest accuracy. This indicates that our AML-Net can transfer more effectively than other leakage methods in the connection mechanism from the generator to the discriminator in adversarial mutual leakage. In particular, we can confirm that AML-Net outperforms PDA-Module in both the mean and standard deviation of IoU accuracy. This is an advantage of the top-down attention mechanism that uses correct images over the bottom-up attention map.

### 4.4.3 w/o Top-Down PDA-Module

We confirm the effect on accuracy by not using the Attention Map in the generator or discriminator, which is cre-
Table 8: Ablation study with some deletions from Top-Down PDA-Module of AML-Net

| Method | Cytoplasm [%] | Nucleus [%] | Background [%] | Mean IoU [%] 
|--------|---------------|-------------|----------------|----------------|
| w/o 2Gen Top-Down PDA-Module | 78.76±5.93 | 90.94±1.11 | 96.92±1.59 | 88.87±2.77 |
| w/o 2DIns Top-Down PDA-Module | 76.66±7.36 | 90.78±1.12 | 96.20±2.45 | 87.88±3.46 |
| AML-Net(ours) | 81.12±5.02 | 90.81±1.56 | 97.59±1.26 | 89.84±2.41 |

Table 9: Hyperparameter sensitivity.

| Method | Cytoplasm [%] | Nucleus [%] | Background [%] | Mean IoU [%] 
|--------|---------------|-------------|----------------|----------------|
| \(\lambda_{adv} = 0\) | 89.98±2.03 | 77.42±6.23 | 90.32±1.46 | 96.60±2.03 |
| \(\lambda_{adv} = 0.1\) | 93.92±2.05 | 78.19±5.13 | 90.62±1.85 | 96.80±1.44 |
| \(\lambda_{adv} = 1\) | 93.16±1.91 | 78.19±3.79 | 90.74±1.73 | 96.80±1.44 |
| \(\lambda_{adv} = 0.01(\text{mm})\) | 94.10±1.81 | 81.12±5.02 | 90.81±1.56 | 97.59±1.26 |

4.4.4 Hyperparameter

We performed an experiment to investigate the effectiveness of the discriminator. Table 8 shows the experimental results when the \(\lambda_{adv}\) in loss function is changed. In both the validation and test images, a decrease in accuracy can be observed when the hyperparameter \(\lambda_{adv}\) is increased. The accuracy degradation is also observed when \(\lambda_{adv} = 0\) is used in adversarial mutual leakage. In this case, the discriminator is trained but the adversarial loss of the discriminator is not used to train the generator. The result shows that the adversarial loss from the discriminator to the generator is necessary even in the case of adversarial mutual leakage. From this experiment, the hyperparameter in this paper is set to \(\lambda_{adv} = 0.01\).

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 801–818, 2018.

[4] Jaehoon Choi, Taekyang Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision—Proceedings of the IEEE/CVF International Conference on Computer Vision*Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6830–6840, 2019.

[5] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[6] Stephan Gerhard, Jan Funke, Julien Martel, Albert Coradona, and Richard Fetter. Segmented Anisotropic sSTEM Dataset of Neural Tissue. *https://figshare.com/articles/Segmented_anisotropic_sSTEM_dataset_of_neural_tissue/856713*.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[9] Yuki Hiramatsu and Kazuhiro Hotta. Semantic segmentation using light attention mechanism. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 622–625, 2020.

[10] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1994–2003, 2017.

[11] AD Hoover, Valentina Kouznetsova, and Michael Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical imaging*, 19(3):203–210, 2000.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[13] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Cnet: Criss-cross attention for semantic segmentation. In *Proceedings of the
IEEE/CVF International Conference on Computer Vision-Proceedings of the IEEE/CVF International Conference on Computer Vision-Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 603–612, 2019.

[14] Ayako Imanishi, Tomokazu Murata, Masaya Sato, Kauhiero Hotta, Itaru Imayoshi, Michiyuki Matsuda, and Kenta Terai. A novel morphological marker for the analysis of molecular activities at the single-cell level. Cell Structure and function, pages 129–140, 2018.

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1125–1134, 2017.

[16] Fangzhou Liao, Ming Liang, Zhe Li, Xiaolin Hu, and Sen Song. Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network. IEEE transactions on neural networks and learning systems, 30(11):3484–3495, 2019.

[17] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1925–1934, 2017.

[18] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In Advances in Neural Information Processing Systems, pages 469–477, 2016.

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3431–3440, 2015.

[20] Pauline Luc, Camille Couprie, Soumichint Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. In Advances in Neural Information Processing Systems Workshops, 2016.

[21] Daisuke Matsuzuki and Kauhiero Hotta. Cell image segmentation using attention module at each layer. In Proceedings of The 20th International Conference on Systems Biology, 2019.

[22] Takeru Miyato, Toshihiko Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In In International Conference on Learning Representations, 2018.

[23] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. In Medical Imaging with Deep Learning, 2018.

[24] Christopher G Owen, Alicia R Rudnicka, Robert Mullen, Sarah A Barman, Dorothy Monekosso, Peter H Whincup, Jeffrey Ng, and Carl Paterson. Measuring retinal vessel tortuosity in 10-year-old children: Validation of the computer-assisted image analysis of the retina (caiair) program. Investigative ophthalmology & visual science, 50(5):2004–2010, 2009.

[25] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2337–2346, 2019.

[26] Pranjit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. In Advances in Neural Information Processing Systems, pages 68–80, 2019.

[27] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In International Conference on Machine Learning, 2016.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 234–241, 2015.

[29] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. Concurrent spatial and channel ‘squeeze & excitation’in fully convolutional networks. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 421–429, 2018.

[30] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3723–3732, 2018.

[31] Kyoya Sawada, Kauhiero Hotta, Ayako Imanishi, Michiyuki Matsuda, and Kenta Terai. Segmentation of cell images by leaking the information of discriminator to generator. In International Conference of the IEEE Engineering in Medicine and Biology Society, 2018.

[32] Pravendra Singh, PRATIK MAZUMDER, and Vinay Nambawardi. Accuracy booster: Performance boosting using feature map re-calibration. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 884–893, 2020.

[33] Joes Staal, Michael DAbramooff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. IEEE transactions on medical imaging, 23(4):501–509, 2004.

[34] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In In International Conference on Learning Representations, 2017.

[35] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7167–7176, 2017.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, pages 5998–6008, 2017.

[37] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaou Tang. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3156–3164, 2017.

[38] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wengang Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. arXiv preprint arXiv:1910.03151, 2019.
[39] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7794–7803, 2018.

[40] Huikai Wu, Junge Zhang, Kaiqi Huang, Kongming Liang, and Yu Yizhou. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. In arXiv preprint arXiv:1903.11816, 2019.

[41] Yuan Xue, Tao Xu, Han Zhang, L Rodney Long, and Xiaolei Huang. Segan: Adversarial network with multi-scale l1 loss for medical image segmentation. Neuroinformatics, 16(3-4):383–392, 2018.

[42] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In In International Conference on Learning Representations, 2016.

[43] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In International Conference on Machine Learning, pages 7354–7363, 2019.

[44] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stacgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, pages 5907–5915, 2017.

[45] Xin Zheng, Yong Wang, Guoyou Wang, and Jianguo Liu. Fast and robust segmentation of white blood cell images by self-supervised learning. Micron, 107:55–71, 2018.

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, pages 2223–2232, 2017.