ADVERSARIAL CONTRASTIVE FOURIER DOMAIN ADAPTATION FOR POLYP SEGMENTATION

Ta Duc Huy[†] *Hoang Cao Huyen*[†] *Chanh D. T. Nguyen*[†] *Soan T. M. Duong*[†‡]
Trung Bui *Steven Q. H. Truong*[†]

[†] VinBrain JSC. [‡] Le Quy Don Technical University

ABSTRACT

Due to the shortage of experienced endoscopists, Computer-Aided Diagnosis (CAD) systems for colonoscopy have recently attracted many research interests. There exist several public polyp segmentation datasets, giving way to the adoptions of domain adaptation methods to address the shift in distributions. Current domain adaptation frameworks often comprise (i) a domain discriminator trained with an adversarial loss and (ii) an image-translation network. Due to the complexity of image-translation networks, such methods are generally hard to train to achieve satisfactory results. Hence, we propose a domain adaptation method that leverages Fourier transform as a simple alternative to the image-translation network. We introduce an adversarial contrastive training strategy to jointly learn an embedding space that considers both *style* and *content* of the sample. Our method demonstrated consistent gains over state-of-the-arts on polyp semantic segmentation task for four public datasets. The code is available at: `https://github.com/tadeephuy/CoFo`

***Index Terms*—** Domain adaptation, Fourier transform, contrastive loss, adversarial learning, polyp segmentation.

1. INTRODUCTION

Up to 3.7% of colorectal carcinoma are diagnosed in three years after a normal colonoscopy due to the lack of senior endoscopists [1]. This gives rise to the development of CAD systems for colonoscopy. Alba et al. [2] conducted an extensive survey on deep learning approaches for polyp localization and polyp classification with very potential results. There exist several public datasets [3, 4, 5, 6] to support the progress in this area. This calls for the development of domain adaptation methods to address the distribution shift problem across different datasets. Distribution shift could stem from the disparity between the hospital sites, the image acquisition protocols, the configurations of imaging machines, which renders the model fail to operate. Domain adaptation refers to adapting a trained model using annotations from a source dataset to perform reliably on a target distribution. Existing domain adaptation methods usually transfer the style between datasets using an image translation model and try to fool the domain discriminator with an adversarial loss [7, 8]. However, stylization models are very brittle, which requires many trials for hyperparameters tuning. Instead, a recent approach transfers the style of the image in the Fourier domain and outperforms those methods that use the extra stylization network [9].

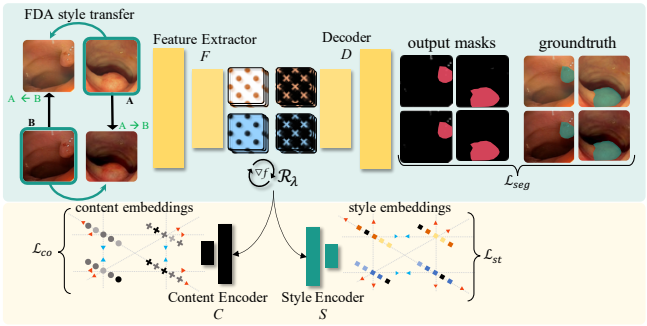

Fig. 1. Overview of the proposed adverarial contrastive Fourier domain adaptation method.

In this paper, we propose an adversarial contrastive Fourier (CoFo) method for unsupervised domain adaptation in the context of semantic segmentation. We follow [9] and leverage Fourier transform to transfer the style between the source and target datasets, resulting in training samples with its original semantic *content* but in either source or target *style*. We append a style encoder and a semantic content encoder to the feature extractor of the model. The two encoders are learned using contrastive loss with supervisions from the *content* and the *style* label of the sample. A gradient reversal layer [10] is also placed between the feature extractor and the two encoders to ensure that the feature extractor learns label-discriminative but domain-invariant features. We illustrate our method in Fig. 1. In summary, our main contributions include:

1. We propose CoFo, a domain adaptation method that transfers the style between the domains using Fourier transform and adversarial training. CoFo disentangles the *style* and *content* concept to learn a compact embedding space.
2. We extensively benchmark our method on polyp segmentation on four public datasets and demonstrate a favorable performance compared to several baselines, proving its effectiveness in this task.

2. METHOD

Here we give an overview of each component in CoFo and describe the loss function to train the model.

Unsupervised domain adaptation (UDA). Given an RGB image $x^s \in \mathbb{R}^{H\times W\times 3}$ and its respective semantic mask $y^s \in \mathbb{R}^{H\times W}$ from the source dataset $D^s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, we have similarly an RGB image $x^t \in \mathbb{R}^{H\times W\times 3}$ from the target dataset $D^t = \{x_i^t\}_{i=1}^{N_t}$. In UDA settings, given a trained model ϕ^s using the source dataset D^s to map $x^s \rightarrow y^s$, we try to adapt ϕ^s to D^t such that it performs well on the target dataset without the availability of the annotated semantic mask $y^t \in \mathbb{R}^{H\times W}$ despite the distribution shift between the two datasets.

Fourier domain adaptation (FDA). Yang et al. propose a simple method to align the low-level statistics between the source and target distribution [9]. This stems from the observation that low-level amplitude spectrum can vary significantly without altering the visual of high-level semantics. Therefore, manipulating the low-level components in the frequency domain is the simplest method to transfer the statistics across different datasets. Given a sampled image pair $x^s \sim D^s$, $x^t \sim D^t$, it first computes the Fourier transform $\mathcal{F}$ of the two images, with $\mathcal{F}^A, \mathcal{F}^P$ being the amplitude and phase components, respectively, and $\mathcal{F}^{-1}$ being the inverse Fourier transform. Let $M_\beta(h,w) = \mathbb{1}_{(h,w)\in[-\beta H:\beta H, -\beta W:\beta W]}$ be the mask of the region where $\beta \in (0,1)$ surrounding the center $(0,0)$. FDA swaps the low frequency components bounded by $[-\beta H{:}\ \beta H, -\beta W{:}\ \beta W]$ of the two amplitude spectral signals and map it back to image space to create $x^{s\rightarrow t}$:

$$\begin{aligned} x^{s\rightarrow t} = \mathcal{F}^{-1}\big([M_\beta \odot \mathcal{F}^A(x^t) \\ + (1 - M_\beta) \odot \mathcal{F}^{\mathcal{A}}(x^s),\ \mathcal{F}^P(x^s)]\big), \end{aligned} \tag{1}$$

where $\odot$ denotes the element-wise multiplication. Such transformation creates a sample having the style of x^t from the target dataset while preserving the source semantic content of x^s, making y^s the corresponding semantic mask of $x^{s\rightarrow t}$. The same operation could be applied to create its reciprocal sample $x^{t\rightarrow s}$ with the style of x^s and the semantic content of x^t, having the semantic mask y^t.

Contrastive learning. Given two inputs $z_a, z_b \in \mathcal{Z}$ and their respective labels y_a, y_b, to learn the function $f : \mathcal{Z} \rightarrow \mathbb{R}^d$ that maps z to the d-dimensional vector in the embedding space in which similar samples ($y_a = y_b$) are close while dissimilar ones ($y_a \neq y_b$) are far apart, Chopra et al. [11] proposes a contrastive loss $\mathcal{L}_c$:

$$\begin{aligned} \mathcal{L}_c\big(f(z_a), f(z_b)\big) = \mathbb{1}_{[y_a=y_b]}\mathcal{D}\big(f(z_a), f(z_b)\big) \\ + \mathbb{1}_{[y_a\neq y_b]}\max\{0, m - \mathcal{D}\big(f(z_a), f(z_b)\big)\}, \end{aligned} \tag{2}$$

where m is the margin specifying the minimum distance between two different classes and $\mathcal{D}(a,b) : \mathbb{R}^d \rightarrow \mathbb{R}$ is a distance function between a pair of d-dimensional vector a, b.

Domain adversarial neural networks (DANN). When the source and target datasets come from different domains, a robust feature space for the classifier to yield predictions must be disentangled from the domain-discriminative properties. Ganin et al. incorporate into the baseline model $G_y \circ F$ a *gradient reversal layer (GRL)* $\mathcal{R}_\lambda$ followed by a domain classifier G_d [10]. This setups encourages the feature extractor F to map a given sample x to an embedding z such that: (i) the classifier G_y can accurately predict the label y of x, (ii) the domain classifier G_d fails to discriminate its domain. $\mathcal{R}_\lambda$ is a parameterless layer which acts as an identity transformation during the forward pass $\mathcal{R}_\lambda(x) = x$, but scales the gradients from G_d by a negative constant λ before passing it to the preceding layers during backpropagation:

$$\frac{d\mathcal{R}_\lambda}{dx} = -\lambda I, \tag{3}$$

where I is an identity matrix. As the training progresses by minimizing the label prediction loss $\mathcal{L}_y$ and domain classification loss $\mathcal{L}_d$, $\mathcal{R}_\lambda$ reverts $\partial\mathcal{L}_d/\partial\theta_F$ to $-\lambda\partial\mathcal{L}_d/\partial\theta_F$, thus updating F in the opposite of the optimal direction for domain classification. Such training process encourages F to learn label-discriminative features for the prediction task and diminish domain-discriminative features that could introduce bias due to the shift between domains.

Network architecture. A decoder $D : \mathcal{Z} \rightarrow \mathbb{R}^{H\times W}$ is appended to the feature extractor $F : \mathbb{R}^{H\times W\times 3} \rightarrow \mathcal{Z}$ to decode the features $z \in \mathcal{Z}$ of an RGB image $x \in \mathbb{R}^{H\times W\times 3}$ to its semantic mask $y \in \mathbb{R}^{H\times W}$:

$$\tilde{y} = D(z), \tag{4}$$

where z is the feature maps encoded by F from x:

$$z = F(x), \tag{5}$$

$D \circ F : \mathbb{R}^{H\times W\times 3} \rightarrow \mathbb{R}^{H\times W}$ is a standard framework of a semantic segmentation model. To approach the UDA problem, we place a GRL $\mathcal{R}_\lambda$ after branching out from F, followed by a style encoder and a content encoder $S, C : \mathcal{Z} \rightarrow \mathbb{R}^d$ that map the encoded features z to two d-dimensional vectors $\tilde{s}, \tilde{c} \in \mathbb{R}^d$:

$$\tilde{s} = S(\mathcal{R}_\lambda(z)), \tag{6}$$

$$\tilde{c} = C(\mathcal{R}_\lambda(z)), \tag{7}$$

where $\tilde{s}$ is the domain *style* vector and $\tilde{c}$ is the domain semantic *content* vector. Fig. 1 illustrates the placement of the components and highlights the learning process.

Combining FDA and DANN. Given a pair of images from source and target dataset $x^s \sim D^s$, $x^t \sim D^t$, we applied equation Eq. (1) to generate $x^{s\rightarrow t}$ and $x^{t\rightarrow s}$, forming a set of input $(x^s, x^t, x^{s\rightarrow t}, x^{t\rightarrow s})$. For each element in the input set, we denote its original domain to be the semantic *content* label c and the domain of its low-frequency components to be its *style* label s. Specifically, the label for *content* and *style* of each input are as follows:

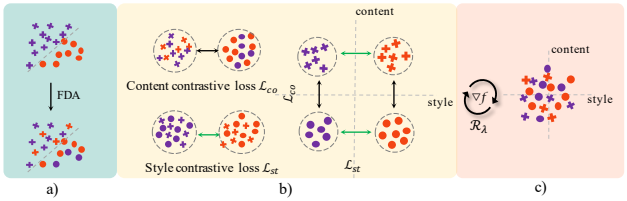

Fig. 2. Combining FDA with DANN. Shape (circle/cross) indicates the *content*, color (red/purple) indicates the *style* of the sample. a) Apply FDA to transfer the *style* between the two datasets, creating red crosses and purple circles. b) Jointly learn the embedding space where *style* and *content* are discrimintative. c) With GRL updating the feature extractor in the opposite of the optimal direction, the learnt embedding space is *content* and *style* invariant. Best viewed in color.

- x^s : source *content*, source *style*,
- x^t : target *content*, target *style*,
- $x^{s\to t}$: source *content*, target *style*,
- $x^{t\to s}$: target *content*, source *style*.

In UDA settings, only the semantic masks y^s of the images x^s from source datasets are available. However, the samples with source *content* in target *style* $x^{s\to t}$ still effectively have y^s as their semantic masks as aforementioned. We learn the model for the semantic segmentation task by minimizing the binary cross-entropy loss $\mathcal{L}_{\text{seg}}$:

$$\mathcal{L}_{\text{seg}} = -\sum (y \log(D(F(x))) + (1-y)\log(1-D(F(x)))), \tag{8}$$

where $x \in \{x^s, x^{s\to t}\}$, and y is its respective semantic mask.

For the content encoder C, we apply the contrastive loss Eq. (2) to learn an embedding space in which inputs with similar *content* label c are grouped together and vice versa:

$$\begin{aligned}\mathcal{L}_{co}(\tilde{c}_a,\tilde{c}_b) &= \mathbb{1}_{[c_a=c_b]}\mathcal{D}(\tilde{c}_a,\tilde{c}_b) \\ &+ \mathbb{1}_{[c_a\neq c_b]}\max(0, m-\mathcal{D}(\tilde{c}_a,\tilde{c}_b).\end{aligned} \tag{9}$$

Similarly for the style encoder S, we learn an embedding space that clusters the inputs by their *style* label s:

$$\begin{aligned}\mathcal{L}_{st}(\tilde{s}_a,\tilde{s}_b) &= \mathbb{1}_{[s_a=s_b]}\mathcal{D}(\tilde{s}_a,\tilde{s}_b) \\ &+ \mathbb{1}_{[s_a\neq s_b]}\max(0, m-\mathcal{D}(\tilde{s}_a,\tilde{s}_b).\end{aligned} \tag{10}$$

Jointly learn S and C results in an embedding space where inputs with both similar style and similar content are close. On the other hand, backpropagation through $\mathcal{R}_\lambda$ reverts $\partial\mathcal{L}_{st}/\partial\theta_F$ to $-\lambda\partial\mathcal{L}_{st}/\partial\theta_F$ and $\partial\mathcal{L}_{co}/\partial\theta_F$ to $-\lambda\partial\mathcal{L}_{co}/\partial\theta_F$, rendering the features z ineffective for *style* and *content* discrimination. In other words, F are forced to learn domain-invariant and label-discriminative features z for D to decode into the semantic mask. The procedure is illustrated in Fig. 2. The final loss function $\mathcal{L}$ of CoFo is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \alpha_{co}\mathcal{L}_{co} + \alpha_{st}\mathcal{L}_{st}, \tag{11}$$

where α_{co} and α_{st} are the weights of $\mathcal{L}_{co}$ and $\mathcal{L}_{st}$.

	CVC.	**Kvasir.**	**ETIS.**	**EndoTect.**
N_s	547	700	0	0
N_{t*}	183	100	50	50
N_t	182	200	146	150

Table 1. Distributions of the datasets. N_s is the number of samples in D^s, N_{t*} is the number of samples in D^{t*}, N_t is the number of samples in D^t.

3. EXPERIMENTS

This section describes the experiments of applying our method on different colonoscopy datasets for the polyp segmentation task.

Datasets. We evaluated CoFo on four public colonoscopy datasets including CVC-EndoScene Still [3], ETIS-Larib [5], Kvasir-SEG [6], and Endotect [4]. Following UDA settings, each dataset was divided into three parts:

- source training set $D^s = \{(x_i^s, y_i^s)\}_i^{N_s}$,
- target adaptation set without annotations $D^{t*} = \{x_i^{t*}\}_i^{N_{t*}}$,
- target test set $D^t = \{(x_i^t, y_i^t)\}_i^{N_t}$.

In each experiment, we trained the model using the source dataset D_a^s and the target adaptation dataset D_b^{t*}. We then tested the trained model on D_b^t, where a, b are the dataset names and $a \neq b$. Due to the small number of samples in Endotect and EITS-Larib, we only used them as the target dataset, each having only D^{t*} and D^t. We appreciate the original train/test split of the dataset if available, and randomly split otherwise. The distributions of the datasets are summarized in Table 1.

Network architecture. We used the standard U-net [12] architecture with ResNet18 [13] backbone for F. The content encoder C has three *ConvBlocks* of a Convolution layer - Batch Normalization [14] - Leaky ReLU activation function, followed by an Average Pooling layer and a Linear layer. The style encoder S is similar to the content encoder C with a slight different in the *ConvBlock* where we use Instance Normalization [15] instead of Batch Normalization. This tweak helps remove instance-specific contrast information as suggested in [16] and yields significant improvements in image stylization, which is widely employed in several stylization applications [17, 18, 19]. We empirically show modest improvement with this choice of design in Table 3. The two encoders output a vector of size $d = 256$.

Training hyperparameters. We used SGD optimizer with a learning rate of 0.03 and nesterov momentum of 0.95 [20]. We trained each experiment for 1000 epochs with a cosine annealing learning rate, reaching 0 after the last epoch. For memory efficiency, the batch is shuffled and matched with it original order for contrastive learning task. We set $\alpha_{co} = \alpha_{st} = 1$, the size β of the FDA swapping mask M_β to 0.01 and λ in the GRL $\mathcal{R}_\lambda$ to 1. We chose cosine distance for the distance function $\mathcal{D}$ with the margin $m = 0.5$. To suppress initial noisy signals from the

contrastive encoders, we scaled λ from 0 to 1 following the formula $\lambda_p = {}^{2}/_{(1+e^{-10p})} - 1$ where p is in range $[0, 1]$, indicating the training progress.

4. RESULTS AND DISCUSSION

Quantitative results. We report the dice score of polyps class of the experiments in Table 2. Our method demonstrates a consistent improvement over other methods by an average dice score of 2%. While DANN [10] is originally developed for image classification, it shows poor results when being applied for semantic segmentation. In comparison to BDL[7] and PCEDA[8], which rely on an image translation network to transfer the style between the two datasets, CoFo follows the simple procedure of FDA [9] but yields solid results. CoFo can be seen as an extension of FDA with an adversarial training task. The result shows that this extension comes with positive gains. We observe that GAN-based methods [7, 8] often comprise a clunky set of several networks that are inherently hard to train and implement, while CoFo is relatively easy to implement. Unlike [8], CoFo is trained end-to-end without the need to pre-generate the stylized images for training. ASN [21] aligns the multi-level feature space of the two distributions using a discriminator network and adversarial loss at each level. Compared to our method, ASN requires more memory footprint during training. CoFo is an adversarial method like [7, 8, 10, 21] but it is noticeably easier to train using the GRL like DANN instead of adversarial losses. It should be noted that the hyperparameter settings used for CoFo is very standard; thus, we believe that CoFo could be further tuned for better results and analysis in future works.

Qualitative results. We compare the qualitative results of CoFo and other methods in Fig. 3, using Kvasir as the source dataset and adapt on ETIS. CoFo outperforms BDL and PCEDA considerably for the small polyp in the first row by producing smoother masks which are also closer to the ground truth. Compared to FDA, which also performs well on small polyps, CoFo is still slightly better.

Ablation study. We conducted ablation studies on CoFo with different configurations and reported the results in Table 3. We used Kvasir as the source dataset and used the rest as the target datasets. We removed the GRL to force the model to learn domain-discriminative features of the two datasets, which resulted in a slight drop in performance compared to the baseline, suggesting that such features do hamper the performance of the model in the context of UDA. It is also noticeable that using Instance Normalization in the style encoder marginally improves the performance when using Batch Normalization in CoFo BN. Furthermore, we measured the effectiveness of the content encoder and the style encoder independently. Results show that the style encoder is the critical component of our approach. Using the style encoder alone leads to significant gains as compared to the content encoder. On the target test set EndoTect, the CoFo Style configuration even outperforms the combined configuration.

Source	CVC.			Kvsr.		
Target	En.T.	ETIS.	Kvsr.	En.T.	ETIS.	CVC.
w/o DA	0.722	0.659	0.769	0.836	0.668	0.732
DANN [10]	0.734	0.654	0.759	0.828	0.506	0.660
ASN [21]	0.601	0.607	0.801	0.833	0.673	**0.837**
BDL [7]	0.750	0.518	0.778	0.854	0.415	0.817
PCEDA [8]	0.765	0.293	0.736	0.822	0.504	0.701
FDA [9]	0.798	0.665	0.804	0.868	0.663	0.751
CoFo	**0.826**	**0.681**	**0.828**	**0.872**	**0.685**	0.811

Table 2. Dice score of the experiments with other SOTA methods. Each cell shows the result when train on the source dataset and test on the target dataset of a given method.w/o DA: without domain adaptation.

Source	Config.			Kvsr.		
Target	GRL	Style	Cont.	EnT.	ETIS.	CVC.
w/o DA				0.836	0.668	0.732
CoFo - GRL		✓	✓	0.823	0.665	0.722
CoFo Style	✓	✓		**0.872**	0.678	0.798
CoFo Cont.	✓		✓	0.859	0.666	0.789
CoFo BN	✓	✓	✓	0.861	0.679	0.792
CoFo	✓	✓	✓	0.864	**0.685**	**0.811**

Table 3. Dice score of different CoFo configurations. CoFo - GRL: not use GRL; CoFo Style: only use the style encoder S; CoFo Cont.: only use the content encoder C; CoFo BN: use Batch Normalization in S instead of Instance Normalization.

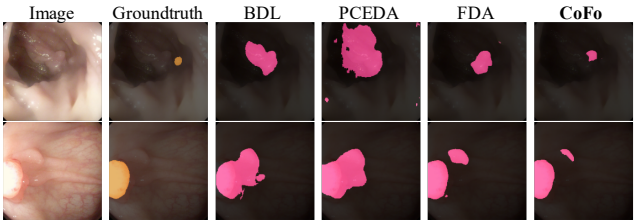

Fig. 3. Qualitative results on Kvasir → ETIS.

5. CONCLUSION

We proposed a UDA method for polyps segmentation. We conducted extensive experiments on four different polyp segmentation datasets to show favorable performance of CoFo compared to several SOTA while being very straightforward to implement. We plan to benchmark our method on other datasets to justify its applicability in broader domains. Further analysis of the hyperparameter setting and backbone architectures is also left for future work.

6. REFERENCES

[1] S. Singh, P. P. Singh, M. H. Murad, H. Singh, and J. N. Samadder, "Prevalence, risk factors, and outcomes of interval colorectal cancers: a systematic review and meta-analysis," *Official J. American Col. Gastroen.*, vol. 109, no. 9, 2014.

[2] A. Nogueira-Rodríguez, R. Domínguez-Carbajales, H. López-Fernández, A. Iglesias, J. Cubiella, f. Fdez-Riverola, M. Reboiro-Jato, and D. Glez-Peña, "Deep neural networks approaches for detecting and classifying colorectal polyps," *Neurocomput.*, vol. 423, pp. 721–734, 2021.

[3] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. Courville, "A benchmark for endoluminal scene segmentation of colonoscopy images," *J. Healthc. Eng.*, vol. 2017, pp. 4037190, 2017.

[4] "The EndoTect 2020 challenge: Evaluation and comparison of classification, segmentation and inference time for endoscopy," .

[5] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 9, no. 2, pp. 283–293, 2014.

[6] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "KVASIR: a multi-class image dataset for computer aided gastrointestinal disease detection," in *Proc. Multimedia Sys. Conf.*, 2017, p. 164–169.

[7] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *2019 Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6929–6938.

[8] Y. Yang, D. Lao, G. Sundaramoorthi, and S. Soatto, "Phase consistent ecological domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9008–9017.

[9] Y. Yang and S. Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2020.

[10] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Machine Learn.*, 2015, vol. 37, pp. 1180–1189.

[11] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 539–546 vol. 1.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Proc. Med. Image Computing and Computer-Assist. Interv.* 2015, pp. 234–241, Springer International Publishing.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[14] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Machine Learn.*, 2015, vol. 37, pp. 448–456.

[15] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: the missing ingredient for fast stylization," *CoRR*, vol. abs/1607.08022, 2016.

[16] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky, "Texture networks: feed-forward synthesis of textures and stylized images," *CoRR*, vol. abs/1603.03417, 2016.

[17] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *2019 Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4396–4405.

[18] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.

[19] J-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017.

[20] A. Botev, G. Lever, and D. Barber, "Nesterov's accelerated gradient and momentum as approximations to regularised update descent," in *Proc. Int. Joint Conf. on Neur. Net.*, 2017, pp. 1899–1903.

[21] Y-H. Tsai, W-C. Hung, S. Schulter, K. Sohn, M-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7472–7481.