2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA) | 978-1-6654-7278-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICPECA56706.2023.10075728

Image Style Transfer–A Critical Review

Yakun Zhang
School of Engineering and Applied Science
George Washington University
Washington, District of Columbia, USA
alexander_zhang@gwu.edu

***Abstract*—Style Transfer refers to the conversion of images into two different domains. Specifically, it provides a style image, converts any image into this style, and preserves the content of the original image as much as possible. Image style transfer is now often employed in image processing applications. It offers a number of evaluation techniques for contrasting numerous prevalent and significant applications of image style transfer algorithms on various level. The examination of numerous Neural Style Transfer applications and unresolved issues for further study comes to a close in this paper.**

Keywords*—*image transfer, style transfer, neural network

I. Introduction

Neural networks based on deep learning have become popular for decades, it has been wildly used in image processing. Image style transfer is one of the greatest research areas. It provides an easy way for people to generate an artistic image. More user-facing applications appear in various fields, such as selfies, video-log, as well as short videos.

In deep learning, neural networks have an excellent ability to extract features and generalize the corresponding photos, because deep learning makes use of multi-layer neural networks enabling extracting various features from target objects automatically. Gatys et al.[1] proposed a seminal idea that both feature image and artistic image can be represented statistically and use Convolution Neural Networks (CNN) to extract content and style separately. Then, it uses appropriate weight to match the image content, the artistic image would be generated. Fig.1 shows the example of image style transfer.

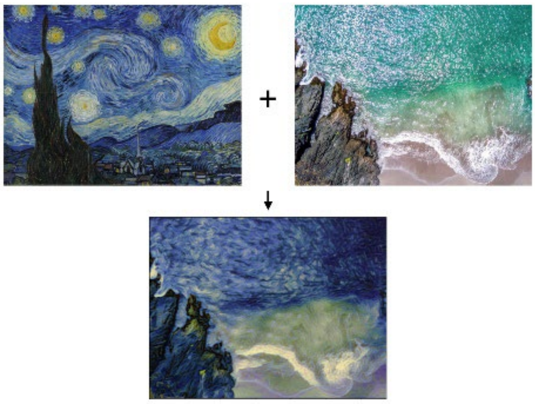

Fig. 1. Syncronize Artistic Image And Target Image Into One Style Image

This paper introduces several image style transfer methods. Including the conventional neural networks, feed-forward convolutional neural networks, instance normalization for fast stylization, and explicit representation. It also offers a chronological path to explore the development of image style transfer. We first introduce their algorithms, then, we analyze the influence of the algorithms. This paper is also a kick-off for junior researchers who are focusing on deep learning and convolutional neural networking.

II. Development of neural style transfer

A. A Neural Artistic Style Algorithm

A well known method proposed by Gatys et al. [1] for transferring image styles is utilising convolutional neural networks(CNN). This approach extracts content and style-related information instead of synthesizing physical models and texture in the conventional manner. then enter a white noise image at random. The network computes and stores the style representation on all layers as the style image travels through it. The content image is stored in the content representation in a single layer and travels via the network. After that, the white noise image is passed, allowed to travel over the network, and calculate the loss of content and style loss respectively. This model updates the bottom image iteratively using a gradient so that its content and style are similar to those of the artistic image. Squared-error loss below:

$$\mathcal{L}_{content}(\vec{p},\vec{x},l) = \frac{1}{2}\sum_{i,j}(F_{ij}^{l} - P_{ij}^{l})^2 \tag{1}$$

Gram matrices [2][3] can be generated and the mean squared distance between them, and the original image's Gram matrices can be reduced significantly by make full use of gradient descent from an intermediate image called white noise. Squared-error loss is:

$$E_l = \frac{1}{4N_l^2 M_l^2}\sum_{i,j}(G_{ij}^{l} - A_{ij}^{l})^2 \tag{2}$$

The features that describe semantic content and style are extracted in this approach using a VGG-16 [4] architecture that has been pre-trained using Image-Net. This algorithm performed well when balancing the accuracy in both content and style matching. However, since every iteration of the optimization needs both forward and backward passes across neural network, it always iterates around 200–300 times to obtain the optimum image. The style transfer model is shown in Fig.2.

B. Feed-Forward Convolutional Neural Network

A feed-forward technique that transforms networks for the transfer of image style is put out by Johnson et al. [5]. Using perceptual loss functions, this method trains the networks. High layer characteristics from a loss network that already pre-trained are required for the perceptual loss functions. In training processing, perceptual losses are used to measure image resemblance stronger than each pixel loss, which can be run in real-time [6][7][8]. Fig.3 shows the image transformation network.

Authorized licensed use limited to: VIT University. Downloaded on September 13,2023 at 15:06:28 UTC from IEEE Xplore. Restrictions apply.

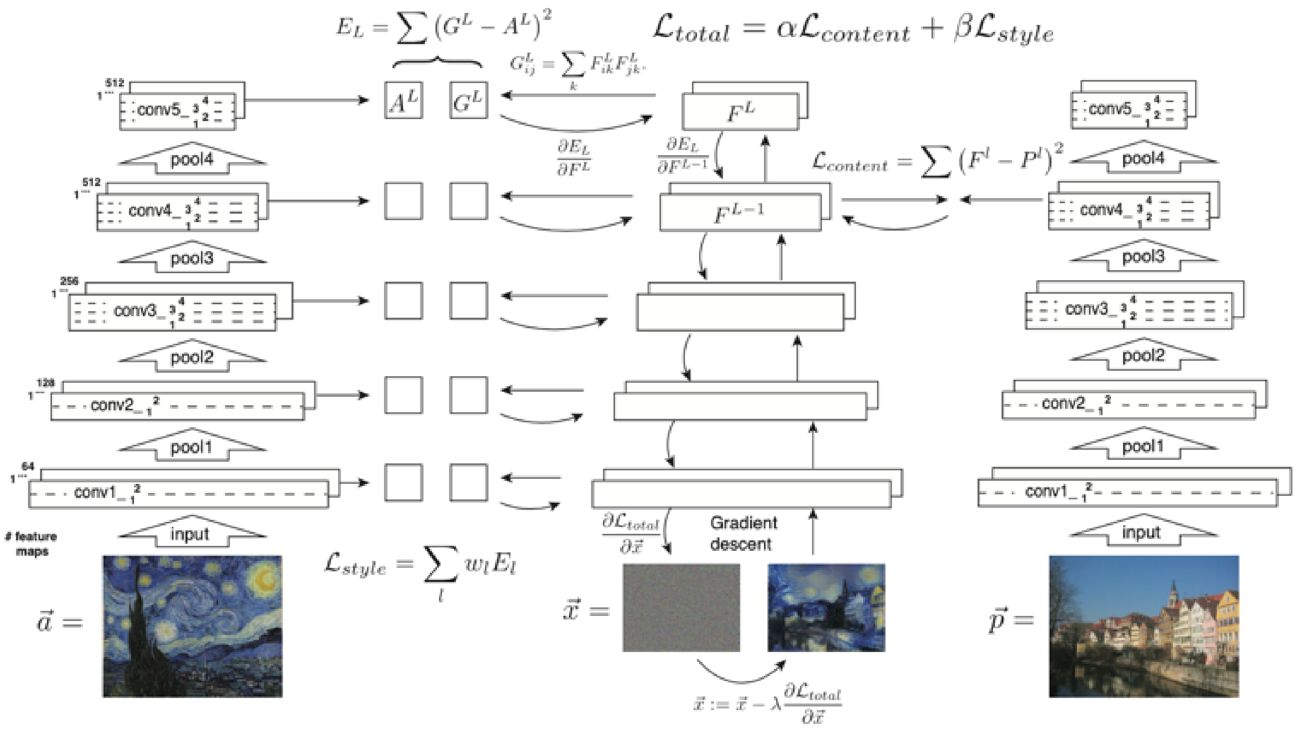
$E_L = \sum (G^L - A^L)^2$
$\mathcal{L}_{total} = \alpha\mathcal{L}_{content} + \beta\mathcal{L}_{style}$
$G_{ij}^L = \sum_k F_{ik}^L F_{jk}^L$
A^L
G^L
F^L
F^{L-1}
$\frac{\partial E_L}{\partial F^L}$
$\frac{\partial E_L}{\partial F^{L-1}}$
$\mathcal{L}_{content} = \sum (F^l - P^l)^2$
conv5_3
conv4_3
conv3_3
conv2_2
conv1_2
pool4
pool3
pool2
pool1
input
512
256
128
64
# feature maps
$\frac{\partial \mathcal{L}_{total}}{\partial \vec{x}}$
Gradient descent
$\mathcal{L}_{style} = \sum_l w_l E_l$
$\vec{a} =$
$\vec{x} =$
$\vec{p} =$
$\vec{x} := \vec{x} - \lambda \frac{\partial \mathcal{L}_{total}}{\partial \vec{x}}$

Fig. 2. Style transfer model

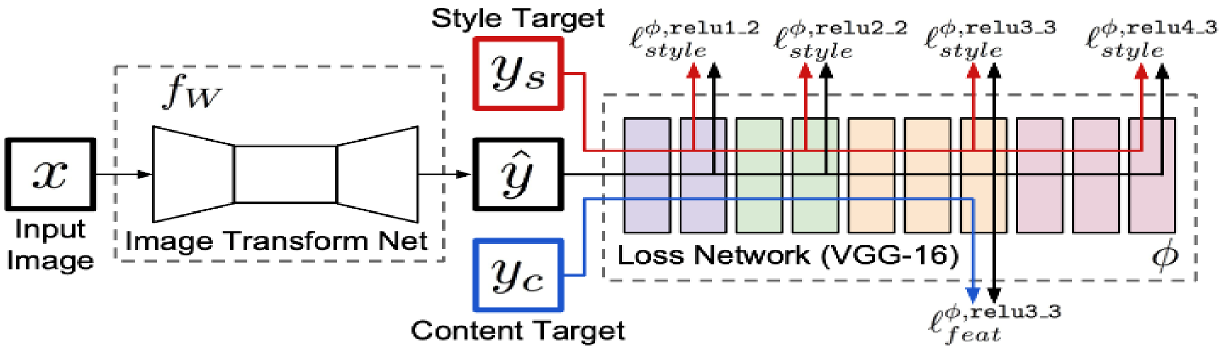
Style Target
y_s
f_W
x
Input Image
Image Transform Net
$\hat{y}$
y_c
Content Target
$\ell_{style}^{\phi,relu1\_2}$
$\ell_{style}^{\phi,relu2\_2}$
$\ell_{style}^{\phi,relu3\_3}$
$\ell_{style}^{\phi,relu4\_3}$
Loss Network (VGG-16)
ϕ
$\ell_{feat}^{\phi,relu3\_3}$

Fig. 3. Utilize an image transformation processing network architecture to convert raw phtoes to target images. measuring perceptual distinction in content and between phtoes using a loss network architecture that pre-trained to establish the perceptual loss function.

Style transfer process and used for single-image applying super-resolution method were the two image alteration tasks that Johnson et al. evaluated. An approximate qualitative outcome is provided by feed-forward networks used in earlier style transfer studies, which generated images through optimization and reduced processing time by up to three orders of magnitude. A per-pixel loss was employed in earlier work on single-image super-resolution using CNN; however, utilizing a perceptual loss instead yielded positive qualitative results.

C. Instance Normalization for Fast Stylization

The contrast of the content image shouldn't dominate the stylization outcome, according to Ulyanov et al. [9]. The issue is whether contrast normalization is best performed directly in the architecture or whether it may be efficiently implemented by mixing common convolutional neural network building parts. The generators were then retrained using the same hyperparameters and batch normalization was replaced with instance normalization. Instance normalization significantly improved both architectures, they found. Although both generators have equivalent quality, they chose Johnson's residuals for the findings because they thought its architecture was a little more effective and user-friendly.

After analyzing the impact of the change, Ulyanov et al. switched batch normalization for instance normalization. employing the same hyperparameters, normalization was applied before retraining the generators. They discovered that the application of instance normalization considerably enhanced both systems.

D. An Explicit Representation

Although Johnson et al. improved the processing speed, it has limitations. Each model can deal with only one style image transformation, which means every time increases a style image, we need to train a new model. Network architecture is shown in Fig.4 and Fig.5.

Authorized licensed use limited to: VIT University. Downloaded on September 13,2023 at 15:06:28 UTC from IEEE Xplore. Restrictions apply.

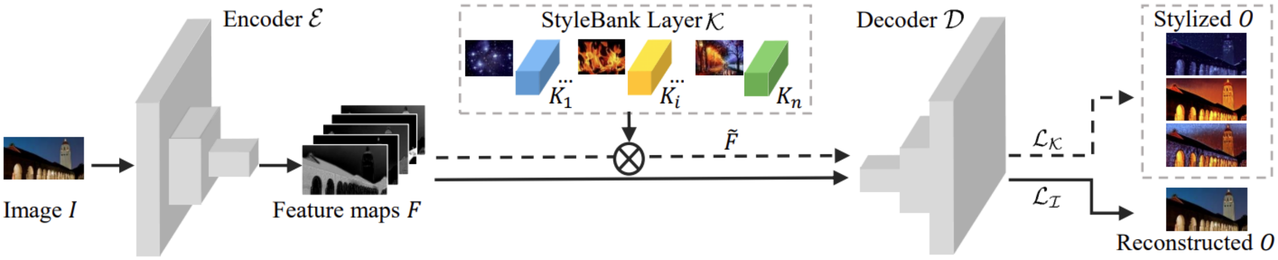

Fig. 4. Three modules make up the network architecture: the image encoder E, the style bank layer K, and the image decoder D.

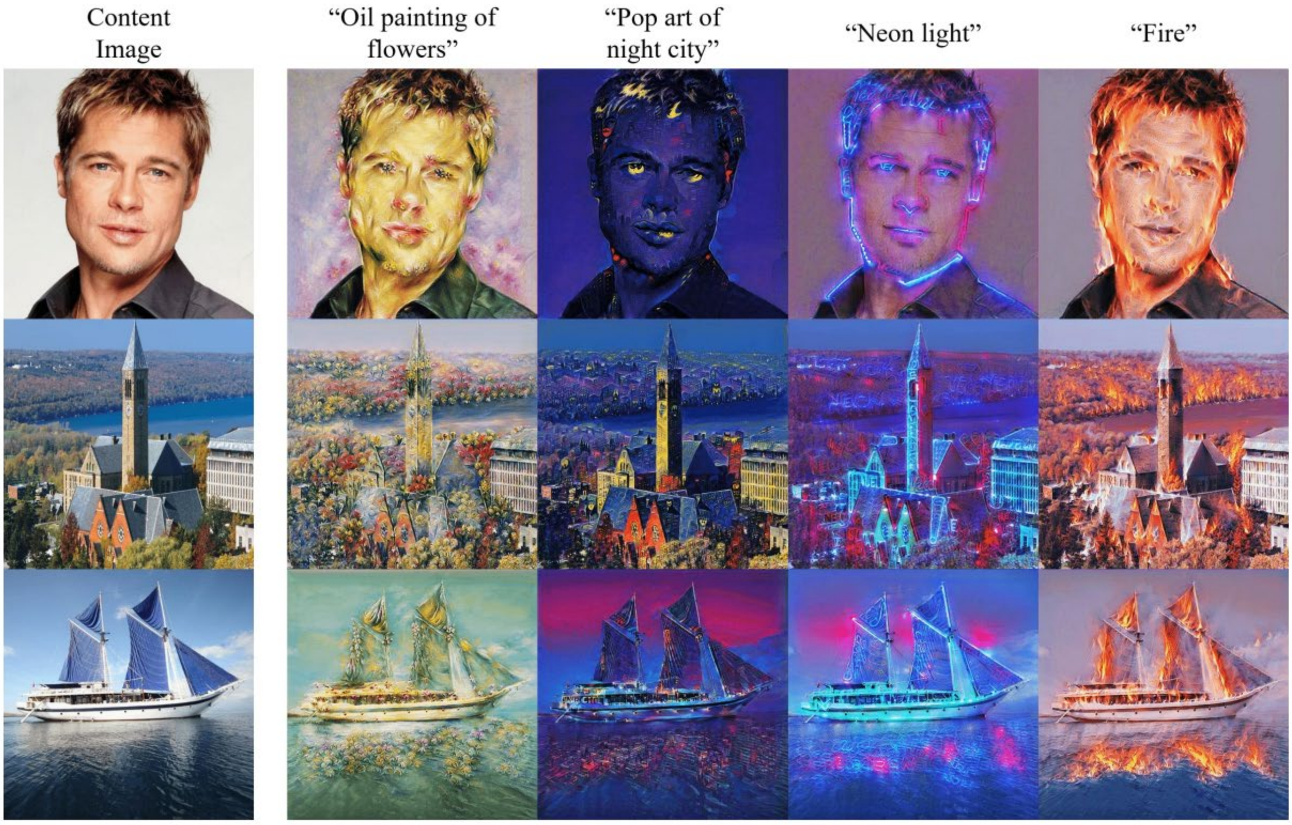

Fig. 5. This picture above shows the results of varied text circumstances on style transmission. Images that have been translated have a realistic spatial structure..

Dongdong et al. [10] proposed an explicit representation called StyleBank. They hoped to separate the process of generation in content and style. That means E and D focus on rebuilding content, and different K control various styles. This optimization allows 50 more different K. In training, Dongdong used (T +1)-step alternative training strategy. Each iteration trains T times (E + K + D), trains once (E + D).

This method has the following two advantages. First, it uses incremental training when adding a new style. Lock E and D at the same time can speed up the training when initializing a new K. Then, it transfers different areas of a picture into different styles is easier when we do Region-specific style fusion.

$$\tilde{F} = (\textstyle\sum_{i=1}^{m} w_i * K_i) \otimes F \qquad (3)$$

Dongdong et al. [10] presented an apparent indication for style and content that enables exciting new style synthesis influence, such as linear and region-specific style transformation, and facilitates faster training for numerous styles. Additionally, they put out a novel interpretation of neural style transmission that stimulates other theories for picture restoration and reconstruction. In this study, they offer a brand-new apparent indication of style and content, which our network can successfully decouple. More importantly, they offer a fresh perspective on neural style transfer, which could lead to new ideas for picture reconstruction and restoration.

E. Transfer of Image Style Using a Single Text Condition

Transfer techniques call for style pictures to transmit texture information to content images over and beyond image style. Users might, however, simply have their imaginations to transfer and not precise style photos. Kwon et al. [11] suggested an answer to this issue. They made use of CLIP's already-trained text-image embedding model. They specifically suggest a patch-wise text-image matching loss for texture content with multi-view augmentations. The CLIP text-image embedding model [12] uses instance normalization layer manipulation or pixel optimization. They suggest developing a thin CNN network that can express text condition texture information and provide output. By comparing the degree to which the transferred image's CLIP

Authorized licensed use limited to: VIT University. Downloaded on September 13,2023 at 15:06:28 UTC from IEEE Xplore. Restrictions apply.

model output and the text condition are comparable, CNN transforms the content image to adhere to the text condition. With this method, text-driven style transfer is possible.

This approach generates images without requiring any style images by just simply changing the text conditions.

F. Pix2Pix

Pix2Pix [14] learns the mapping using a conditional GAN model and adversarial training; as a result, the output cannot be distinguished from the input by a discriminator. The first automatic I2I approach is this one. Pix2Pix's architecture (as depicted in Fig. 6) is made up of a creator and a dis-creator. It employs U-net as its fundamental creator design in order to get over the basic encoder-decoder bottleneck and provide I2I results with high resolution. Because PatchGAN captures local style information and model high-frequencies, Pix2Pix employs it as a discriminator.

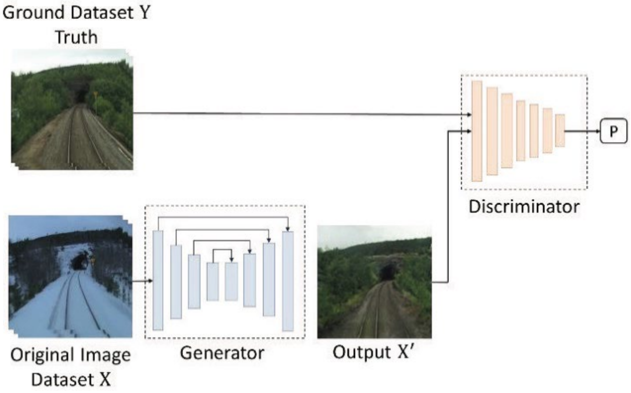

Fig. 6. Architecture of Pix2Pix

One L1 distance loss and one adversarial loss make up Pix2pix's optimization function. The adversarial loss condition, taken from cGAN, is as follows:

$$\mathcal{L}_{cGAN}(G,D) = min_G max_D \mathbb{E}_{x\sim\mathbb{P}_r}[logD(x,y)] + \mathbb{E}_{z\sim\mathbb{P}_g}[\log(1-D(x,G(z)))] \quad (4)$$

III. Conclusion

Image style transfer has significantly increased. The IST is explained in full in this work, along with its research ideas and research methodologies. In this work, we compare a variety of picture style transfer techniques that use deep learning and others that don't. Deep learning techniques have advanced quickly in recent years as a result of being inspired by these algorithms. The goal of research has increasingly been to increase transmission speed and quality. A generic evaluation of picture transmission quality will be necessary in the future because the majority of the studies in use today evaluate experimental outcomes using subjective evaluations.

In addition to being able to create beautiful artwork, style transfer has the ability to resolve increasing challenging computer vision and image processing issues. Through social media, cellphone cameras, photo-editing apps, and other means, it is helpful in day-to-day living. It can imitate any artist's painting style thanks to the integration of Generative Adversarial Networks into the style transfer process. This sort of visual intelligence exists. Since style transfer demands powerful computing and graphics hardware, the majority of study has been conducted on photos alone.

References

[1] Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2414-2423).

[2] Gatys, L., Ecker, A. S., & Bethge, M. (2015). Texture synthesis using convolutional neural networks. *Advances in neural information processing systems, 28.*

[3] Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision, 40*(1), 49-70.

[4] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.*

[5] Johnson, J., Alahi, A., & Fei-Fei, L. (2016, October). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision* (pp. 694-711). Springer, Cham.

[6] Eigen, D., & Fergus, R. (2015). Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision* (pp. 2650-2658).

[7] Mahendran, A., & Vedaldi, A. (2015). Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5188-5196).

[8] Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312*.6034.

[9] Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

[10] Chen, D., Yuan, L., Liao, J., Yu, N., & Hua, G. (2017). Stylebank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1897-1906).

[11] Kwon, G., & Ye, J. C. (2022). Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18062-18071).

[12] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748-8763). PMLR.

[13] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

[14] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125-1134).

[15] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).

[16] Kim, T., Cha, M., Kim, H., Lee, J. K., & Kim, J. (2017, July). Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning* (pp. 1857-1865). PMLR.

[17] I. P. Labs, "Prisma: Turn memories into art using artificial," 2016. [Online]. Available: http://prisma-ai.com.

Authorized licensed use limited to: VIT University. Downloaded on September 13,2023 at 15:06:28 UTC from IEEE Xplore. Restrictions apply.