Deep Translation Prior: Test-Time Training for Photorealistic Style Transfer

Sunwoo Kim* **Soohyun Kim*** **Seungryong Kim**†

Korea University, Seoul, Korea
{sw-kim, shkim1211, seungryong_kim}@korea.ac.kr

Abstract

Recent techniques to solve photorealistic style transfer within deep convolutional neural networks (CNNs) generally require intensive training from large-scale datasets, thus having limited applicability and poor generalization ability to unseen images or styles. To overcome this, we propose a novel framework, dubbed Deep Translation Prior (DTP), to accomplish photorealistic style transfer through test-time training on given input image pair with untrained networks, which learns an image pair-specific translation prior and thus yields better performance and generalization. Tailored for such test-time training for style transfer, we present novel network architectures, with two sub-modules of correspondence and generation modules, and loss functions consisting of contrastive content, style, and cycle consistency losses. Our framework does not require offline training phase for style transfer, which has been one of the main challenges in existing methods, but the networks are to be solely learned during test time. Experimental results prove that our framework has a better generalization ability to unseen image pairs and even outperforms the state-of-the-art methods.

Introduction

Photorealistic style transfer is one of appealing image manipulation and editing tasks, which aims to, given a pair of images, i.e., the content and style image, synthesize an image by transferring the style to the content. Recent approaches for this task leverage statistics of content and style features extracted by deep convolutional neural network (CNNs) (Gatys, Ecker, and Bethge 2016; Li et al. 2017; Ulyanov, Vedaldi, and Lempitsky 2017), which can be divided into *optimization*-based and *learning*-based methods. Optimization-based methods (Gatys, Ecker, and Bethge 2016; Li and Wand 2016; Luan et al. 2017) directly obtain a stylized image by optimizing an image itself with well-defined content and style loss functions. As the seminal work, Gatys et al. (Gatys, Ecker, and Bethge 2016) present the style loss function based on Gram matrix and optimize the stylized image with the loss function, of which many variants were also proposed (Li and Wand 2016; Gatys, Ecker, and Bethge 2016; Luan et al. 2017). Since the loss function for style transfer is often non-convex, most methods leverage an iterative solver to optimize the output image itself (Li and Wand 2016; Gatys, Ecker, and Bethge 2016; Luan et al. 2017), and thus they can benefit from error feedback for stylization. Moreover, they are limited to encode an image translation prior on synthesized images, and thus often generate artifacts and show limited photorealism.

In contrast, recent learning-based methods (Li et al. 2017, 2018; Gu et al. 2018; Yoo et al. 2019; Huang and Belongie 2017; Park and Lee 2019) attempt to address these limitations by learning such image translation prior within networks from large-scale datasets (Deng et al. 2009; Lin et al. 2014), often followed by pre- or post-processing (Li et al. 2018; Yoo et al. 2019; Huang and Belongie 2017). Since it is notoriously challenging to collect training pairs for photorealistic style transfer due to its subjectivity, most methods alternatively leverage an auto-encoder to learn a decoder that captures the translation prior (Huang and Belongie 2017; Chen and Schmidt 2016; Li et al. 2017). However, during the training, these methods (Li et al. 2018, 2019) do not leverage explicit content and style loss functions, and thus may have poor generalization ability on unseen images or styles. In addition, adopting fixed network parameters at test-time may not account for the fact that a pair of images may require their own prior, namely an image pair-specific translation prior.

In this paper, we explore an alternative, dubbed Deep Translation Prior (DTP), to overcome aforementioned limitations of both optimization- and learning-based methods. Our work accomplishes this without need of intensive training process using large-scale dataset or paired data, but through a test-time training on given input image pair. We argue that the translation prior does not necessarily need to be learned from intensive learning. Instead, an image pair-specific translation prior can be captured by solely minimizing explicit content and style loss functions on the image pair with untrained network for stylization. Tailored to this framework, we formulate novel network architectures consisting of two sub-modules, namely correspondence and generation modules, which are learned with well-designed content, style, and cycle consistency loss functions at test time.

Our experiments on standard benchmark for photorealistic style transfer (Luan et al. 2017; An et al. 2020), CelebA-HQ (Liu et al. 2015), and Flickr Faces HQ (FFHQ) (Karras, Laine, and Aila 2019) demonstrate that our framework con-

*Equal contribution.

†Corresponding author.

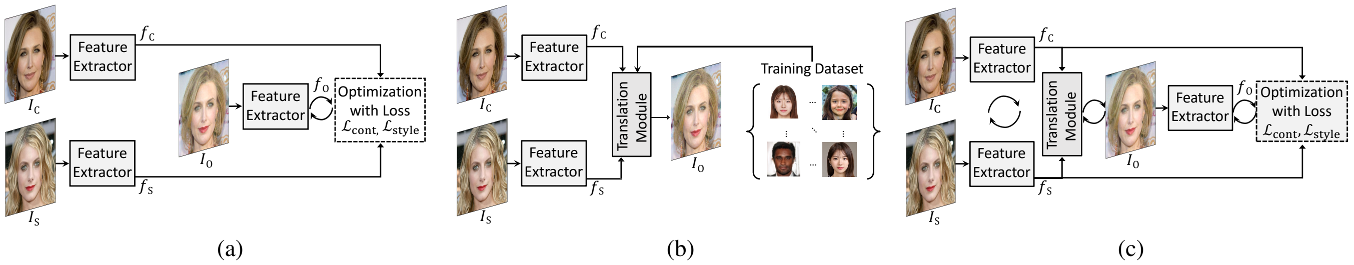

Figure 1: Intuition of DTP: (a) conventional optimization-based methods (Li and Wand 2016; Gatys, Ecker, and Bethge 2016; Luan et al. 2017) that optimize an output *image itself at test time* with explicit content loss and style loss, which often generate artifacts and produce limited photorealism due to the lack of translation prior, (b) recent learning-based methods that require *intensive offline training* from large-scale training data and use *pretrained and fixed* networks at test time (Huang and Belongie 2017; Li et al. 2017, 2018; Gu et al. 2018; Yoo et al. 2019), and (c) DTP that *learns the untrained networks at test time on an input image pair* to capture an image pair-specific translation prior and thus provides better generalization to unseen images or styles.

sistently outperforms the existing methods.

Related Work

Style Transfer. Traditional optimization-based methods for style transfer (Gatys, Ecker, and Bethge 2016; Li and Wand 2016; Luan et al. 2017) using pre-trained feature extractors, such as VGG networks (Simonyan and Zisserman 2015), can be divided into parametric and non-parametric methods. Categorized as parametric methods, some methods (Gatys, Ecker, and Bethge 2016; Berger and Memisevic 2016) designed Gram matrix to capture global statistics of features. However, the loss function based on Gram matrix leads to poor results as it captures the per-pixel feature correlations and does not constrain the spatial layout. To address this issue, non-parametric approaches (Li and Wand 2016; Luan et al. 2017; Aberman et al. 2018) match a style on patch-level. Categorized as non-parametric methods, some works have realized style transfer inspired by an image analogy (Hertzmann et al. 2001), which is based on dense correspondence (Shih et al. 2014; Liao et al. 2017). STROTSS (Kolkin, Salavon, and Shakhnarovich 2019) uses optimal transport algorithm. The aforementioned optimization-based methods are limited to encode an image translation prior, thus often generating artifacts and showing limited photorealism.

On the other hand, recent learning-based methods (Chen and Schmidt 2016; Li et al. 2017; Huang and Belongie 2017; Gu et al. 2018; Sanakoyeu et al. 2018; Li et al. 2018; Yoo et al. 2019; Park and Lee 2019; Liu et al. 2021) tried to solve style transfer by data-driven ways. They mostly focused on designing loss functions, and often included pre- or post-processing (Li et al. 2018) to produce spatially smooth output. For instance, contextual loss is proposed (Mechrez, Talmi, and Zelnik-Manor 2018), which trains CNNs solely using the content images without need of large-scale paired dataset. Several works (Huang and Belongie 2017; Gu et al. 2018; Yoo et al. 2019; Qu, Shao, and Qi 2019; Park and Lee 2019) trained a decoder network with MS-COCO dataset (Lin et al. 2014) or ImageNet dataset (Deng et al. 2009), or needed training the whole network per style (Sanakoyeu et al. 2018) before optimization process. However, they may be biased to the training images or styles and may not generalize well to unseen data.

Image Prior. Deep Image Prior (DIP) (Ulyanov, Vedaldi, and Lempitsky 2018) proves the structure of generator network itself can serve as a prior for image restoration, against the assumption that learning from large-scale data is necessary to capture realistic image prior (Zhang et al. 2017), of which many variants were proposed, tailored to solve an inverse problem (Burger et al. 2005; Dabov et al. 2007; Burger, Schuler, and Harmeling 2012). SinGAN (Shaham, Dekel, and Michaeli 2019) and SinIR (Yoo and Chen 2021) fine-tune GAN or AE on a single input and can be applied to image manipulation and restoration. GAN inversion (Jahanian, Chai, and Isola 2020; Menon et al. 2020; Gu, Shen, and Zhou 2020) aims at generating an image by solely optimizing a latent code of pre-trained GAN given a target image. Different from the aforementioned methods that attempts to learn an image prior, our framework is the first attempt to learn the image translation prior.

Methodology

Motivation

Photorealistic style transfer aims at transferring the style of image $I_{\mathcal{S}}$ to the content of image $I_{\mathcal{C}}$ to synthesize a stylized image $I_{\mathcal{C}\leftarrow\mathcal{S}}$. To achieve this, traditional methods (Li and Wand 2016; Gatys, Ecker, and Bethge 2016; Luan et al. 2017) focused on an image *optimization* technique, from which deep convolutional features were extracted from content and style images, denoted by $F_{\mathcal{C}} = \Phi(I_{\mathcal{C}})$ and $F_{\mathcal{S}} = \Phi(I_{\mathcal{S}})$ with feature extractor $\Phi(\cdot)$, and used to define an objective function, consisting of content loss $\mathcal{L}_{\text{cont}}$ and style loss $\mathcal{L}_{\text{style}}$ functions, as in Figure. 1(a):

$$I_{\mathcal{C}\leftarrow\mathcal{S}} = \underset{I}{\operatorname{argmin}}\{\mathcal{L}_{\text{cont}}(\Phi(I), F_{\mathcal{C}}) + \mathcal{L}_{\text{style}}(\Phi(I), F_{\mathcal{S}})\}. \tag{1}$$

Since it is often a non-convex optimization, most methods leverage an iterative solver, e.g., gradient descent (Li and Wand 2016; Gatys, Ecker, and Bethge 2016; Luan et al. 2017), and thus they benefit from an error feedback to find better

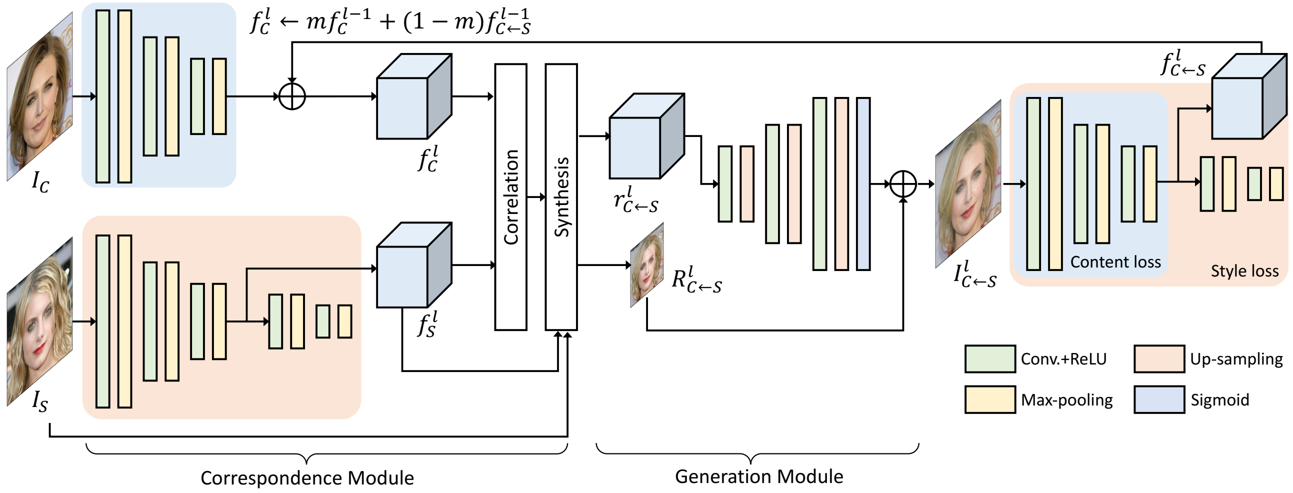
$f_{\mathcal{C}}^{l} \leftarrow m f_{\mathcal{C}}^{l-1} + (1-m) f_{\mathcal{C}\leftarrow\mathcal{S}}^{l-1}$
$I_{\mathcal{C}}$
$f_{\mathcal{C}}^{l}$
Correlation
Synthesis
$r_{\mathcal{C}\leftarrow\mathcal{S}}^{l}$
$R_{\mathcal{C}\leftarrow\mathcal{S}}^{l}$
$I_{\mathcal{C}\leftarrow\mathcal{S}}^{l}$
$f_{\mathcal{C}\leftarrow\mathcal{S}}^{l}$
Content loss
Style loss
$I_{\mathcal{S}}$
$f_{\mathcal{S}}^{l}$
Conv.+ReLU
Up-sampling
Max-pooling
Sigmoid
Correspondence Module
Generation Module

Figure 2: Network configuration of DTP. Our network consists of two sub-modules, *correspondence* module and *generation* module. At the first, we predict a translation hypothesis by first computing the similarity between each source point and all target points and then warping the style image and feature in a probabilistic manner. At the second, the warped feature goes through a decoding network that generates the residual of the final stylized image.

stylized images. However, they are limited to encode an image *translation* prior on synthesized images, and thus often generate artifacts and limited photorealism.

To overcome these limitations, recent *learning*-based methods (Li et al. 2017; Huang and Belongie 2017; Li et al. 2018; Gu et al. 2018; Yoo et al. 2019) attempt to learn such translation prior within the networks during training. Starting with feature extraction module, they designed feature fusion module, e.g., AdaIN (Huang and Belongie 2017) or WCT (Li et al. 2017), and trained decoder module on large-scale image data, e.g., ImageNet (Deng et al. 2009) or MS-COCO (Lin et al. 2014), as in Figure. 1(b), which can be formulated as

$$\omega^{\dagger} = \underset{\omega}{\operatorname{argmin}} \sum\nolimits_{n} \mathcal{L}_{\text{recon}}(\mathcal{F}(F_{\mathcal{C},n}, F_{\mathcal{S},n}; \omega), I_{\mathcal{C}\leftarrow\mathcal{S},n}), \tag{2}$$

where $F_{\mathcal{C},n}$ and $F_{\mathcal{S},n}$ are features from n-th image pair and $I_{\mathcal{C}\leftarrow\mathcal{S},n}$ is n-th stylized image sampled from massive training data. $\mathcal{F}(\cdot;\omega)$ is a feed-forward process with decoder parameters ω. $\mathcal{L}_{\text{recon}}$ is an image reconstruction loss function (Gatys, Ecker, and Bethge 2016; Liu, Breuel, and Kautz 2017; Huang et al. 2018; Park et al. 2019). In practice, since it is notoriously challenging to collect training pairs for style transfer, $\{(F_{\mathcal{C},n}, F_{\mathcal{S},n}, I_{\mathcal{C}\leftarrow\mathcal{S},n})\}_{n\in\{1,\ldots,N\}}$, due to its subjectivity, most methods (Li et al. 2017; Huang and Belongie 2017; Li et al. 2018; Gu et al. 2018; Yoo et al. 2019) alternatively leverage an auto-encoding setting to learn the decoder with parameters ω to reconstruct an input image itself learned by $\mathcal{L}_{\text{recon}}(\mathcal{F}(\Phi(I);\omega), I)$. At test time, given $F_{\mathcal{C}}$ and $F_{\mathcal{S}}$, a stylization process can be formulated as follows:

$$I_{\mathcal{C}\leftarrow\mathcal{S}} = \mathcal{F}(F_{\mathcal{C}}, F_{\mathcal{S}}; \omega^{\dagger}). \tag{3}$$

These methods are based on the assumption that the image translation prior can be learned within the model itself from massive training data. However, during the training phase, these methods do not leverage *explicit* content and style loss functions, as done in optimization methods (Li and Wand 2016; Gatys, Ecker, and Bethge 2016; Luan et al. 2017), thus providing limited stylization performance when their assumptions are violated, e.g., under unseen images or styles. In addition, adopting fixed network parameters at test time may not capture an image pair-specific translation prior.

Overview

To overcome aforementioned limitations and take the best of both approaches, we present Deep Translation Prior (DTP) framework. We argue that the translation prior does not necessarily need to be learned from intensive learning or datasets. Instead, an *image pair-specific translation prior* can be captured by solely *minimizing explicit content and style loss functions on the image pair*, like what is done by conventional optimization-based methods (Li and Wand 2016; Gatys, Ecker, and Bethge 2016; Luan et al. 2017), with an *untrained network for stylization*, which takes benefits of large capacity and robustness of networks as in recent learning-based methods (Li et al. 2017; Huang and Belongie 2017; Li et al. 2018; Gu et al. 2018; Yoo et al. 2019), as in Figure. 1(c), formulated as

$$\begin{aligned} \omega^{*} = \underset{\omega}{\operatorname{argmin}} \{&\mathcal{L}_{\text{cont}}(\Phi(\mathcal{F}(F_{\mathcal{C}}, F_{\mathcal{S}}; \omega)), F_{\mathcal{C}}) \\ &+ \mathcal{L}_{\text{style}}(\Phi(\mathcal{F}(F_{\mathcal{C}}, F_{\mathcal{S}}; \omega)), F_{\mathcal{S}})\}, \\ I_{\mathcal{C}\leftarrow\mathcal{S}} = \mathcal{F}(&F_{\mathcal{C}}, F_{\mathcal{S}}; \omega^{*}), \end{aligned} \tag{4}$$

where ω^{*} is overfitted to the input image pair, which encodes the image pair-specific translation prior. Unlike conventional optimization-based methods (Gatys, Ecker, and Bethge 2016; Li and Wand 2016), our framework generates better stylization results while eliminating the artifacts thanks to the *structure* of networks that can encode the image pair-specific prior during test-time training. In addition, unlike recent learning-based methods (Li et al. 2017; Huang and Belongie 2017;

Li et al. 2018; Gu et al. 2018; Yoo et al. 2019), our framework does not require an intensive training for decoder, but only requires an off-the-shelf feature extractor and untrained generator, thus having better generalization ability to unseen images or styles.

Tailored for such test-time training for style transfer, we design our stylization networks in a two-stage fashion, as illustrated in Figure. 2; on one hand, the model predicts a translation hypothesis by first computing the similarity between each content point and all style points by means of the feature vectors $F_{\mathcal{C}}$ and $F_{\mathcal{S}}$, called *correspondence* module, and on the other, the model refines the hypothesis through the decoder for more plausible stylization, called *generation* module. Since the generated output is desired to preserve the structure of the content image $I_{\mathcal{C}}$ while faithfully stylizing from semantically similar parts in the style image $I_{\mathcal{S}}$, we present contrastive content loss function and style loss function to boost the convergence of our test-time training framework.

It should be noted that there exist similar literature for image restoration tasks, e.g., Deep Image Prior (DIP) (Ulyanov, Vedaldi, and Lempitsky 2018), that have shown that the structure of a generator network can capture a low-level *image* prior by optimizing a randomly-initialized network with a task-dependent fidelity term on a single image. To the best of our knowledge, our framework is the first attempt to learn the *translation* prior at test time for photorealistic style transfer.

Network Architecture

Correspondence Module. We first present a correspondence module to measure the similarities between each point in content feature $f_{\mathcal{C}}$ and all other points in style feature $f_{\mathcal{S}}$, enabling generating a translation hypothesis. It is inspired by the classical matching pipeline (Rocco, Arandjelovic, and Sivic 2017) in that we first extract the feature vectors and then compute the similarity between them.

Following the previous approaches for style transfer, we first extract the deep convolution features, e.g., VGGNet (Simonyan and Zisserman 2015) pretrained on ImageNet (Deng et al. 2009), as follows:

$$\begin{aligned} f_{\mathcal{C}} &= \Phi(I_{\mathcal{C}};\omega_f) \in \mathbb{R}^{H\times W\times C}, \\ f_{\mathcal{S}} &= \Phi(I_{\mathcal{S}};\omega_f) \in \mathbb{R}^{H\times W\times C}, \end{aligned} \tag{5}$$

where H and W are spatial size, with C channels of f. ω_f are feature extraction parameters. Unlike most existing methods that use *fixed* feature extraction parameters, we adaptively *fine-tune* the parameters to the input image pair. We then compute a correlation matrix $M \in \mathbb{R}^{HW\times HW}$, of which each term is a pairwise feature correlation such that

$$M(u,v) = \frac{\hat{f}_{\mathcal{C}}(u)^T \hat{f}_{\mathcal{S}}(v)}{\|\hat{f}_{\mathcal{C}}(u)\|\|\hat{f}_{\mathcal{S}}(v)\|}, \tag{6}$$

where u and v represent all the points in the content and style images, respectively. $\hat{f}_{\mathcal{C}}(u)$ and $\hat{f}_{\mathcal{S}}(v)$ are channel-wise centralized features of $f_{\mathcal{C}}(u)$ and $f_{\mathcal{S}}(v)$ as

$$\hat{f}_{\mathcal{C}}(u) = f_{\mathcal{C}}(u) - \bar{f}_{\mathcal{C}}, \quad \hat{f}_{\mathcal{S}}(v) = f_{\mathcal{S}}(v) - \bar{f}_{\mathcal{S}}, \tag{7}$$

where $\bar{f}_{\mathcal{C}}$ is an average of $f_{\mathcal{C}}(u)$ across all the points in the content. $\bar{f}_{\mathcal{S}}$ is similarly defined. Since $M(u,v)$ represents a similarity between u and v, the higher, the more similar.

By using the correlation matrix M, we synthesize an warped style feature $r_{\mathcal{C}\leftarrow\mathcal{S}}$, i.e., the style feature spatially-aligned to the content image. The warping function can be formulated in many possible ways, but we borrow the technique in (Zhang et al. 2020) that uses a reconstruction:

$$r_{\mathcal{C}\leftarrow\mathcal{S}}(u) = \sum\nolimits_v \Omega(M(u,v)/\tau) f_{\mathcal{S}}(v), \tag{8}$$

where Ω means the softmax operator across v, and τ is a temperature parameter.

generation module. Our generation module aims at reconstructing an image from warped feature $r_{\mathcal{C}\leftarrow\mathcal{S}}$. We present the decoder that has a symmetric structure of feature extractor architecture, similar to (Li et al. 2017; Huang and Belongie 2017). This decoding process can be formulated as follows:

$$I_{\mathcal{C}\leftarrow\mathcal{S}} = \mathcal{F}(r_{\mathcal{C}\leftarrow\mathcal{S}};\omega_g), \tag{9}$$

where ω_g is decoding parameters. As described above, the parameters are first *randomly*-initialized and then learned with *explicit* loss functions for style transfer at test time.

However, due to non-convexity of the loss functions for style transfer, generating the image $I_{\mathcal{C}\leftarrow\mathcal{S}}$ through the decoder directly is extremely hard to converge. To elevate the stability and boost the convergence, we exploit not only warped style feature $r_{\mathcal{C}\leftarrow\mathcal{S}}$, but also warped style image $R_{\mathcal{C}\leftarrow\mathcal{S}}$, extracted such that $R_{\mathcal{C}\leftarrow\mathcal{S}}(u) = \sum_v \Omega(M(u,v)/\tau) I_{\mathcal{S}}(v)$, as a guidance for style transfer, where the networks only learn the residual for the final result as follows:

$$I_{\mathcal{C}\leftarrow\mathcal{S}} = \lambda_w \mathcal{F}(r_{\mathcal{C}\leftarrow\mathcal{S}};\omega_g) + (1-\lambda_w) R_{\mathcal{C}\leftarrow\mathcal{S}}, \tag{10}$$

where λ_w is a weight parameter. By leveraging such a residual prediction, convergence of our test-time training could be greatly improved. Moreover, it enables directly flowing the loss gradients to both feature extractor with ω_f and image generator with ω_g, which helps to boost the performance.

Iterative Formulation. Since the loss function for style transfer, which will be discussed later, is non-convex, we formulate our test-time training as an iterative framework As evolving the iteration, the image $I^l_{\mathcal{C}\leftarrow\mathcal{S}}$ at l-th iteration converges to better stylization results, since it is generated from the updated feature extractor parameters ω_f and decoder parameters ω_g. Since the image $I^l_{\mathcal{C}\leftarrow\mathcal{S}}$ is getting close to the optimal, if the content image $I_{\mathcal{C}}$ can be substituted by $I^l_{\mathcal{C}\leftarrow\mathcal{S}}$ in a recurrent fashion, the iterative solver can converge faster and boost performance. However, at early stages during optimization, $I^l_{\mathcal{C}\leftarrow\mathcal{S}}$ contains blurry regions and noises, which prohibit using such an explicit recurrent formulation. To overcome this limitation, we adopt a moving averaging technique similar to (Kim et al. 2019; Schmidt et al. 2020) in a manner that we smoothly substitute the content feature $f_{\mathcal{C}}$ by the output feature $f_{\mathcal{C}\leftarrow\mathcal{S}} = \Phi(I_{\mathcal{C}\leftarrow\mathcal{S}};\omega_f)$ with a momentum parameter m, such that

$$f^l_{\mathcal{C}} \leftarrow m f^l_{\mathcal{C}} + (1-m) f^{l-1}_{\mathcal{C}\leftarrow\mathcal{S}}, \tag{11}$$

which is used to the current content feature.

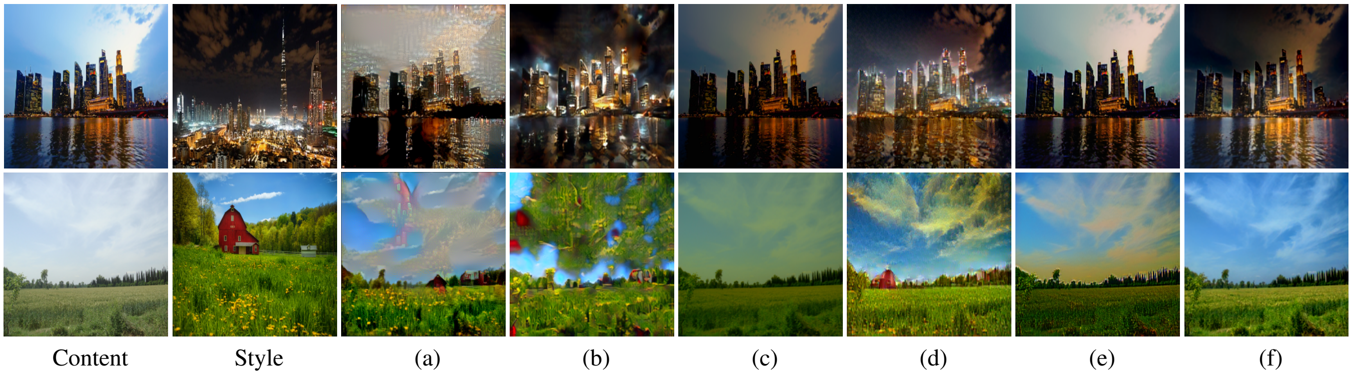

Figure 3: Comparison of DTP with other methods on standard benchmark (Luan et al. 2017; An et al. 2020): Given content and style images, stylized results are achieved by (a) Gatys et al. (Gatys, Ecker, and Bethge 2016), (b) WCT (Li et al. 2017), (c) WCT2 (Yoo et al. 2019), (d) STROTSS (Kolkin, Salavon, and Shakhnarovich 2019), (e) SinIR (Yoo and Chen 2021), and (f) Ours. Compared to others, our model generates realistic results while successfully transferring both global and local style information and preserving structure. Our model does not require any mask, pre- or post-processing.

Loss functions

Content Loss. In photorealistic style transfer, the structure of content image should be preserved on the output image, and thus the content loss is generally defined as the feature difference between the content image $I_{\mathcal{C}}$ and the generated image $I_{\mathcal{C}\leftarrow\mathcal{S}}$, e.g., $\|f_{\mathcal{C}} - f_{\mathcal{C}\leftarrow\mathcal{S}}\|^2$. However, this content loss does not consider other pixels, thus making the result blurry and inducing the trivial solution when the features are simultaneously trained as in our test-time training. To overcome this, we revisit infoNCE loss (Oord, Li, and Vinyals 2018) to set the *pseudo* positive samples between content image $I_{\mathcal{C}}$ and generated image $I_{\mathcal{C}\leftarrow\mathcal{S}}$. We first encode the feature stacks as used in (Chen et al. 2020; Park et al. 2020a), compiling stacks of features $\{f^l_{\mathcal{C}}\}$ and $\{f^l_{\mathcal{C}\leftarrow\mathcal{S}}\}$, where $l \in \{1, ..., L_{\mathcal{C}}\}$. We define the exponential of inner product function of two vectors to express the equation more comfortably.

$$s(f,g) = \exp\left(\left(f^T g/\|f\|\|g\|\right)/\tau\right), \qquad (12)$$

where τ is a temperature parameter. Our final content loss is then defined as follows:

$$\mathcal{L}_{\text{cont}} = -\sum_l \sum_u \log\left(\frac{s(f^l_{\mathcal{C}}(u), f^l_{\mathcal{C}\leftarrow\mathcal{S}}(u))}{\sum_v s(f^l_{\mathcal{C}}(u), f^l_{\mathcal{C}\leftarrow\mathcal{S}}(u))}\right). \qquad (13)$$

Style Loss. We additionally adopt style loss functions. Unlike parametric methods (Gatys, Ecker, and Bethge 2016; Li et al. 2017; Chen and Koltun 2017; Li et al. 2018; Yoo et al. 2019) that enforce the style loss globally, we adopt the style loss similar to non-parametric methods (Li and Wand 2016; Kim et al. 2019) for getting detailed results which are more suitable for photo realistic style transfer. Similar to the content loss, the style loss is defined in a multi-scale manner. Our style loss is defined as below:

$$\mathcal{L}_{\text{style}} = \sum\nolimits_l \sum\nolimits_v \|\Psi(f^l_{\mathcal{C}\leftarrow\mathcal{S}}(v)) - \Psi(f^l_{\mathcal{S}}(NN(v)))\|^2_F, \qquad (14)$$

for $l \in \{1, ..., L_{\mathcal{S}}\}$. $\|\cdot\|^2_F$ denotes a Frobenius norm. $NN(v)$ is the index of the patch in $\Psi(f_{\mathcal{S}})$ that is the nearest patch of $\Psi(f_{\mathcal{C}\leftarrow\mathcal{S}}(v))$.

Cycle Consistency Loss. To improve the stability during training, we further present cycle consistency loss as a regularization which enforces the stylized feature r should be back-warped to the original feature f well, defined such that

$$r_{\mathcal{S}} = \sum_u \Omega\left(\frac{M(u,v)}{\tau}\right) \sum_v \Omega\left(\frac{M(u,v)}{\tau}\right) f_{\mathcal{S}}(v). \qquad (15)$$

$r_{\mathcal{C}}$ is similarly defined. Then the cycle consistency loss $\mathcal{L}_{\text{cyc}}$ is bidirectionally defined as

$$\mathcal{L}_{\text{cyc}} = \sum_u \{\|f_{\mathcal{C}}(u) - r_{\mathcal{C}}(u)\|^2_F + \|f_{\mathcal{S}}(u) - r_{\mathcal{S}}(u)\|^2_F\}. \qquad (16)$$

Total Loss. Finally, the total loss can be summarized such that $\mathcal{L} = \lambda_{\text{c}}\mathcal{L}_{\text{cont}} + (1-\lambda_{\text{c}})\mathcal{L}_{\text{style}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}}$, where λ_c, λ_w and λ_{cyc} represent loss adjusting hyperparameters.

Experiments

Implementation Details

We first summarize implementation details in our framework. For feature extractor, we used the ImageNet (Deng et al. 2009) pre-trained VGG-19 (Simonyan and Zisserman 2015) network. We set the temperature parameter τ as 0.07 and weight parameter λ_w as $1/9$. We also empirically set momentum parameter m as 0.4. $\lambda_{\text{c}} = 1/5$ and $\lambda_{\text{cyc}} = 1$ were used to adjust loss functions. We set the learning rates as $1e^{-4}$. We conduct experiments using a single 24GB RTX 3090 GPU. The network occupies memories about 6GB. We optimize our network over 1000 iterations, which takes about 150 seconds. The pair of content and style images are bilinearly resized to the size of 256×256 in our experiment.

Experimental Setup

We used three kinds of datasets to evaluate our method, including standard datasets for photorealistic style transfer (Luan et al. 2017; An et al. 2020), CelebA-HQ (Liu et al. 2015), and Flickr Faces HQ (FFHQ) (Karras, Laine, and

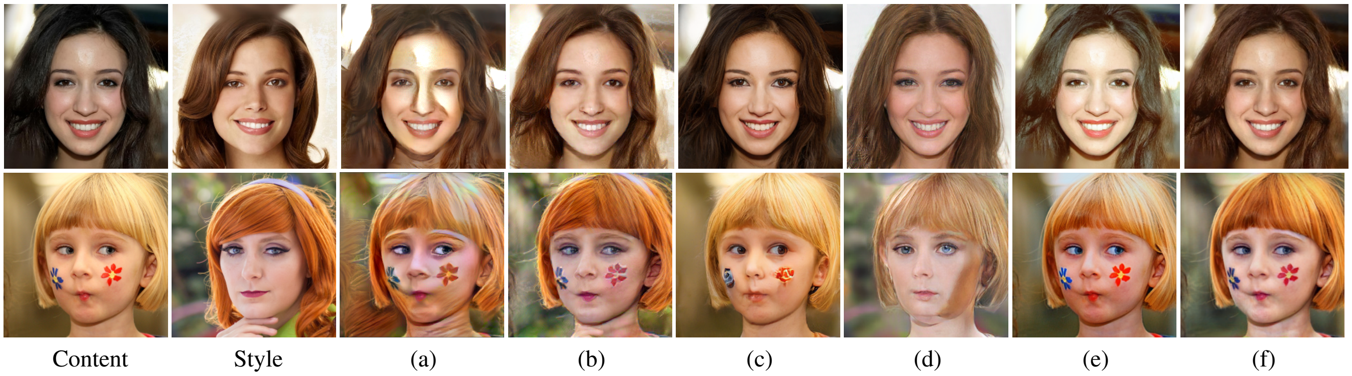

Figure 4: Comparison of DTP with other methods on CelebA-HQ dataset (Liu et al. 2015) and FFHQ dataset (Karras, Laine, and Aila 2019): Given content and style images, translation results are achieved by (a) Gatys et al. (Gatys, Ecker, and Bethge 2016), (b) STROTSS (Kolkin, Salavon, and Shakhnarovich 2019), (c) Swapping Autoencoder (Park et al. 2020b), (d) CoCosNet (Zhang et al. 2020), (e) SinIR (Yoo and Chen 2021), and (f) Ours. The results show that our networks can translate local features as well as global features from both structure and style. Note that CoCosNet (Zhang et al. 2020) was trained on CelebA-HQ, including segmentation masks, and Swapping Autoencoder (Park et al. 2020b) was trained on FFHQ.

	Photo-R.		CelebA-HQ		FFHQ	
Methods	PE↓	SSIM↑	PE↓	SSIM↑	PE↓	SSIM↑
Gatys et al.	3.09	0.44	4.37	0.66	1.97	0.69
WCT	2.48	0.22	-	-	-	-
WCT2	1.99	0.70	-	-	-	-
STROTSS	3.23	0.37	1.81	0.60	1.65	0.66
Swap. AE	2.92	0.22	0.89	0.60	**1.13**	0.58
CoCosNet	-	-	1.56	0.43	-	-
SinIR	**1.00**	**0.84**	1.56	0.84	1.28	0.93
DTP, λ_c=4/5	1.36	0.75	0.79	0.86	1.4	0.93
DTP, λ_c=1/5	1.09	0.82	**0.47**	**0.96**	1.21	**0.98**

Table 1: Quantitative evaluation on standard benchmark for photorealistic style transfer (Luan et al. 2017; An et al. 2020), CelebA-HQ (Liu et al. 2015), and FFHQ (Karras, Laine, and Aila 2019).

Aila 2019). We compared our method with recent state-of-the-art style transfer methods, such as Gatys et al. (Gatys, Ecker, and Bethge 2016), WCT (Li et al. 2017), WCT2 (Yoo et al. 2019), STROTSS (Kolkin, Salavon, and Shakhnarovich 2019), SinIR (Yoo and Chen 2021). We also compared with image-to-image translation tasks, such as CoCosNet (Zhang et al. 2020), and Swapping Autoencoder (Park et al. 2020b). It should be emphasized that learning-based style transfer methods (Li et al. 2017; Yoo et al. 2019) and image-to-image translation methods (Zhang et al. 2020; Park et al. 2020b) are trained on tremendous training data, while our method just trains the networks at test time with a pair of images.

Experimental Results

Qualitative Evaluation. In this section, we evaluated photorealistic style transfer results of our method compared with state-of-the-art methods, with respect to two aspects, including synthesized image quality and semantic consistency. Qualitative results are shown in Figure. 3 and Figure. 4. Our generated results are realistic and contain both global-local style features while successfully preserving the structure from the contents. Traditional optimization-based method (Gatys, Ecker, and Bethge 2016) shows poor synthesis quality. Learning-based methods (Li et al. 2017; Yoo et al. 2019) that use fixed decoder parameters at test time. Since both kinds of methods do not consider translation prior, they show limitations in preserving fine details and make artifacts.

Unlike these, our approach has shown high generalization ability to any unseen input images. On the other hand, while Swapping Autoencoder (Park et al. 2020b) was trained with a large-scale dataset, but limited to generating plausible results, our results show competitive, even better, results. Our success in fine details of both style and content can be found in Figure. 4, where our method produces the most visually appealing images with more vivid details. For example, the top right shows the closest skin color from style and the exact same texture of hair from content as well as overall clarity outperform the state-of-the-art methods.

Quantitative Evaluation. We further evaluate our method with the quantitative results as in Table. 1 on photorealistic style transfer examples, CelebA-HQ dataset, and Flicker-Faces dataset (FFHQ) with metrics of Pieapp (Prashnani et al. 2018) (PE) and Structural Similarity (SSIM) (Wang et al. 2004). Pieapp is a reference-based quality assessment used for semantic consistency. SSIM index is an error measurement which is computed between the original content images and stylized images. Here, we do not measure WCT (Li et al. 2017) and WCT2 (Yoo et al. 2019) on CelebA-HQ and FFHQ since WCT requires pre-training on each task. Similarly, since CoCosNet (Zhang et al. 2020) has been trained on example-based image-to-image translation task, we do not evaluate it on photorealistic style transfer examples. We also show the effect of changing weight λ_c of content and style loss. In the results, our model significantly outperforms most of the methods on photorealistic examples under the both evalua-

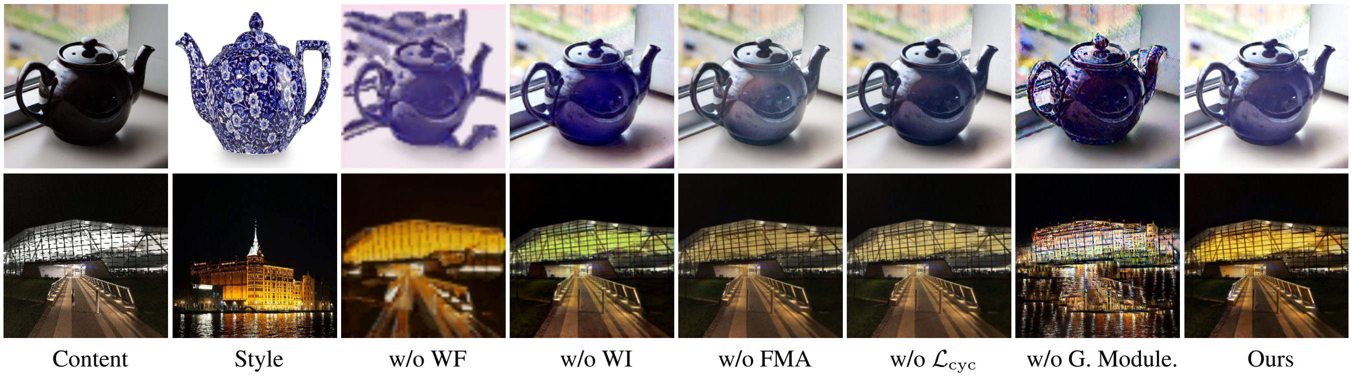

Figure 5: Ablation study on priors, cycle consistency loss, generation module and feature moving average: Warped features (WF) and images (WI) help generation module to contain translation prior in the feature and image level. Feature moving average (FMA) helps to converge to better solution and $\mathcal{L}_{\text{cyc}}$ stabilizes test-time optimization. Generation module (G. Module) helps to learn translation prior better.

tion metrics. In particular, our results with $\lambda_{\text{c}} = 1/5$ tend to show better quantitative results, since smaller λ_{c} produces results with more similar structure to content image, which we analyze in more detail in the supplementary material. Also, results on CelebA-HQ are very close to best metrics. The results indicate significant performance gains with our method in all metrics. Interestingly, though our method is designed for style transfer, our architecture also works on image-to-image translation task.

Ablation Study

In order to validate the effectiveness of each component in our method, we conduct a comprehensive ablation study. In particular, we analyze the effectiveness of warped feature and image, cycle consistency loss $\mathcal{L}_{\text{cyc}}$, feature moving average, and generation module in Figure. 5. To validate the effect of the warped feature and image in our model, we conduct ablation experiments by replacing the warped feature with random Gaussian noise and eliminating the residual connection with warped image. Without the warped feature, it fails to preserve edges while the result without the warped image fails to capture any style information. We also validate the influence of cycle consistency loss, which makes the optimization process more stable. Feature moving average makes the synthesized image more vivid because the previous synthesized feature can give guidance for content and style feature correlation. Without the generation module, the output cannot converge to the optimal result. With all these components, our work is more effective in resulting style relevant outputs while preserving content clearly.

User Study

We also conducted a user study on 80 participants to evaluate the quality of synthesized images in the experiments with the following questions: *"Which do you think has better image quality / similar content to content image / style relavance to style image?"*. On photorealistic style transfer examples, CelebA-HQ (Liu et al. 2015) and FFHQ dataset (Karras, Laine, and Aila 2019), our method ranks the first in every cases, which can be found in Figure. 6.

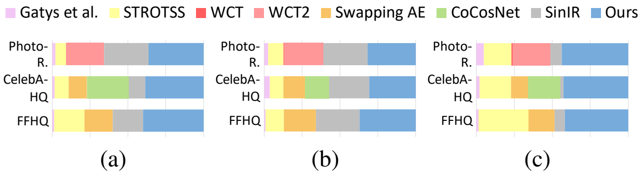

Figure 6: User study results: (a) image quality, (b) content relevance, and (c) style relevance.

Conclusion

In this paper, we proposed, for the first time, a novel framework to learn the style transfer network on a given input image pair at test time, without need of any large-scale dataset, hand-labeling, and task-specific training process, called Deep Translation Prior (DTP). Tailored for such test-time training for style transfer, we formulate overall networks as two submodules, including correspondence module and generation module. By training the untrained networks with explicit loss functions for style transfer at test time, our approach achieves better generalization ability to unseen image pairs or style, which has been one of the major bottlenecks of previous methods. Experimental results on a variety of benchmarks and in comparison to state-of-the-art methods proved that our framework outperforms the existing optimization- and learning-based solutions.

Acknowledgements

This research was supported by the MSIT, Korea (IITP-2021-2020-0-01819, ICT Creative Consilience program, IITP-2021-0-00155, Context and Activity Analysis-based Solution for Safe Childcare, and IITP-2019-0-01343, Regional strategic Industry convergence security core talent training business) supervised by the IITP and National Research Foundation of Korea (NRF-2021R1C1C1006897).

References

Aberman, K.; Liao, J.; Shi, M.; Lischinski, D.; Chen, B.; and Cohen-Or, D. 2018. Neural best-buddies: Sparse cross-domain correspondence. *ACM Transactions on Graphics*, 37(4): 1–14.

An, J.; Xiong, H.; Huan, J.; and Luo, J. 2020. Ultrafast Photorealistic Style Transfer via Neural Architecture Search. In *AAAI*.

Berger, G.; and Memisevic, R. 2016. Incorporating long-range consistency in cnn-based texture generation. *arXiv preprint arXiv:1606.01286*.

Burger, H. C.; Schuler, C. J.; and Harmeling, S. 2012. Image denoising: Can plain neural networks compete with BM3D? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2392–2399. IEEE.

Burger, M.; Osher, S.; Xu, J.; and Gilboa, G. 2005. Nonlinear inverse scale space methods for image restoration. In *International Workshop on Variational, Geometric, and Level Set Methods in Computer Vision*, 25–36. Springer.

Chen, Q.; and Koltun, V. 2017. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, 1511–1520.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, T. Q.; and Schmidt, M. 2016. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337*.

Dabov, K.; Foi, A.; Katkovnik, V.; and Egiazarian, K. 2007. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8): 2080–2095.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. Ieee.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2414–2423.

Gu, J.; Shen, Y.; and Zhou, B. 2020. Image processing using multi-code gan prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3012–3021.

Gu, S.; Chen, C.; Liao, J.; and Yuan, L. 2018. Arbitrary style transfer with deep feature reshuffle. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8222–8231.

Hertzmann, A.; Jacobs, C. E.; Oliver, N.; Curless, B.; and Salesin, D. H. 2001. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 327–340.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1501–1510.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision*, 172–189.

Jahanian, A.; Chai, L.; and Isola, P. 2020. On the "steerability" of generative adversarial networks. In *International Conference on Learning Representations*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Kim, S.; Min, D.; Jeong, S.; Kim, S.; Jeon, S.; and Sohn, K. 2019. Semantic attribute matching networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12339–12348.

Kolkin, N.; Salavon, J.; and Shakhnarovich, G. 2019. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10051–10060.

Li, C.; and Wand, M. 2016. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2479–2486.

Li, P.; Zhao, L.; Xu, D.; and Lu, D. 2019. Optimal transport of deep feature for image style transfer. In *Proceedings of the 2019 4th International Conference on Multimedia Systems and Signal Processing*, 167–171.

Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; and Yang, M.-H. 2017. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, 386–396.

Li, Y.; Liu, M.-Y.; Li, X.; Yang, M.-H.; and Kautz, J. 2018. A closed-form solution to photorealistic image stylization. In *Proceedings of the European Conference on Computer Vision*, 453–468.

Liao, J.; Yao, Y.; Yuan, L.; Hua, G.; and Kang, S. B. 2017. Visual Attribute Transfer Through Deep Image Analogy. *ACM Transactions on Graphics*, 36(4): 120:1–120:15.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 740–755. Springer.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, 700–708.

Liu, S.; Lin, T.; He, D.; Li, F.; Wang, M.; Li, X.; Sun, Z.; Li, Q.; and Ding, E. 2021. AdaAttN: Revisit Attention Mechanism in Arbitrary Neural Style Transfer. In *Proceedings of the IEEE International Conference on Computer Vision*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Luan, F.; Paris, S.; Shechtman, E.; and Bala, K. 2017. Deep photo style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4990–4998.

Mechrez, R.; Talmi, I.; and Zelnik-Manor, L. 2018. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision*, 768–783.

Menon, S.; Damian, A.; Hu, S.; Ravi, N.; and Rudin, C. 2020. PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2437–2445.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR*.

Park, D. Y.; and Lee, K. H. 2019. Arbitrary style transfer with style-attentional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5880–5888.

Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020a. Contrastive Learning for Unpaired Image-to-Image Translation. In *Proceedings of the European conference on Computer Vision*.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2337–2346.

Park, T.; Zhu, J.-Y.; Wang, O.; Lu, J.; Shechtman, E.; Efros, A. A.; and Zhang, R. 2020b. Swapping Autoencoder for Deep Image Manipulation. In *Advances in Neural Information Processing Systems*.

Prashnani, E.; Cai, H.; Mostofi, Y.; and Sen, P. 2018. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1808–1817.

Qu, Y.; Shao, Z.; and Qi, H. 2019. One-Shot Mutual Affine-Transfer for Photorealistic Stylization. *arXiv preprint arXiv:1907.10274*.

Rocco, I.; Arandjelovic, R.; and Sivic, J. 2017. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6148–6157.

Sanakoyeu, A.; Kotovenko, D.; Lang, S.; and Ommer, B. 2018. A style-aware content loss for real-time hd style transfer. In *Proceedings of the European Conference on Computer Vision*, 698–714.

Schmidt, V.; Sreedhar, M. N.; ElAraby, M.; and Rish, I. 2020. Towards Lifelong Self-Supervision For Unpaired Image-to-Image Translation. *arXiv preprint arXiv:2004.00161*.

Shaham, T. R.; Dekel, T.; and Michaeli, T. 2019. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE International Conference on Computer Vision*, 4570–4580.

Shih, Y.; Paris, S.; Barnes, C.; Freeman, W. T.; and Durand, F. 2014. Style transfer for headshot portraits. *ACM Transactions on Graphics*.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2017. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6924–6932.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2018. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9446–9454.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Yoo, J.; and Chen, Q. 2021. SinIR: Efficient General Image Manipulation with Single Image Reconstruction. In *ICML*.

Yoo, J.; Uh, Y.; Chun, S.; Kang, B.; and Ha, J.-W. 2019. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE International Conference on Computer Vision*, 9036–9045.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2017. Understanding deep learning requires rethinking generalization. In *CVPR*.

Zhang, P.; Zhang, B.; Chen, D.; Yuan, L.; and Wen, F. 2020. Cross-domain Correspondence Learning for Exemplar-based Image Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5143–5153.