Non-Local Representation based Mutual Affine-Transfer Network for Photorealistic Stylization

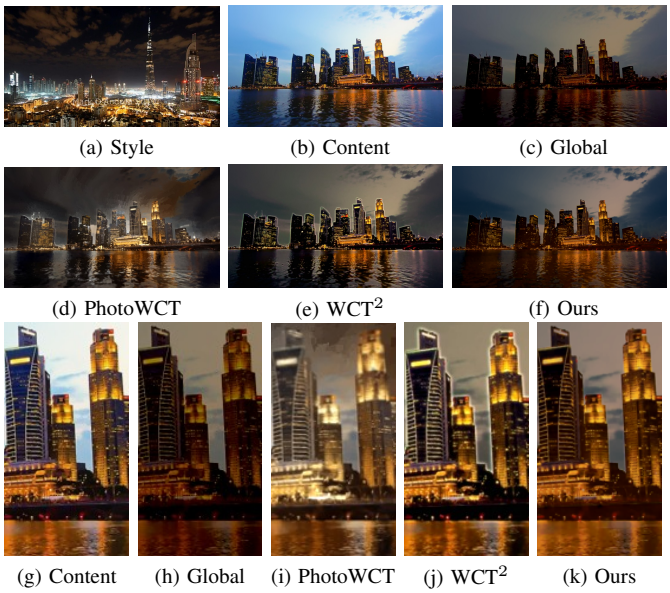
Ying Qu, *Member, IEEE,*, Zhenzhou Shao*, *Member, IEEE,* Hairong Qi, *Fellow, IEEE*

***Abstract*—Photorealistic stylization aims to transfer the style of a reference photo onto a content photo in a natural fashion, such that the stylized image looks like a real photo taken by a camera. State-of-the-art methods stylize the image locally within each matched semantic region and are prone to global color inconsistency across semantic objects/parts, making the stylized image less photorealistic. To tackle the challenging issues, we propose a non-local representation scheme, constrained with a mutual affine-transfer network (NL-MAT). Through a dictionary-based decomposition, NL-MAT is able to successfully decouple matched non-local representations and color information of the image pair, such that the context correspondence between the image pair is incorporated naturally, which largely facilitates local style transfer in a global-consistent fashion. To the best of our knowledge, this is the first attempt to address the photorealistic stylization problem with a non-local representation scheme, such that no additional models or steps for semantic matching are required during stylization. Experimental results demonstrate that, the proposed method is able to generate photorealistic results with local style transfer while preserving both the spatial structure and global color consistency of the content image. Please find the final version from IEEE Transactions on Pattern Analysis and Machine Intelligence on IEEE Xplore. The code will be released on https://github.com/yingutk/NL-MAT.**

***Index Terms*—Photorealistic Stylization, Non-local Representation, Mutual Information, Affine-Transfer**

I. Introduction

THE objective of photorealistic style transfer is to change the style of a content photo to that of a reference photo as shown in Fig. 1. By choosing different reference photos, one could make the content photo look as if, for example, it was taken under different illuminations, at different time of the day or season of the year [1]–[3]. It is worth mentioning that photorealistic style transfer is different from the general stylization approaches [4], [5], which tend to generate the painting-like photos with artistic textures and distorted structures. As emphasized in previous works [1]–[3], [6], a successful photorealistic stylization method should be able to transfer sophisticated *matched* styles with *local* color changes while at the same time *preserve the spatial* (or structural) information as well as *global color consistency* of the content photo naturally, such that the resulting image looks like a real photo taken by a camera.



(a) Style (b) Content (c) Global (d) PhotoWCT (e) WCT^2 (f) Ours (g) Content (h) Global (i) PhotoWCT (j) WCT^2 (k) Ours

Fig. 1: Given a reference style photo taken at night, the content image is stylized as if it was taken at night. (a) Style image. (b) Content image. Photorealistic-stylized images with (c) global-based method [7] without local color changes, (d) PhotoWCT [3] with structure distortion, (e) WCT^2 [6] with abrupt color changes between sky and buildings, and (f) our method (NL-MAT) with global consistency. (g-k) Sub-images of (b)-(f), respectively.

Existing approaches generally perform photorealistic stylization either in a global or a local way. Global-based methods [7], [8] transfer the style of a photo with spatially invariant functions. Although they perform well for global color shifting, they usually could not stylize images effectively within local areas. For example, to stylize the content photo in Fig. 1b according to the style photo in Fig. 1a, the bright day sky of Fig. 1b should become dark night sky, while the light of the buildings should be preserved. However, as shown in Fig. 1c, the global-based method turns both light and the day sky to dark. Local-based methods [1]–[3], [6], [9]–[11] are generally carried out in three major steps, including 1) extraction of high-level features from the content and style image pair with pre-trained models on large datasets, 2) semantic context matching of the image pair, and 3) local context-based stylization on the extracted features. For most approaches, post-processing has to be performed

Ying Qu, and Hairong Qi are with the Advanced Imaging and Collaborative Information Processing Group, Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996 USA (e-mail: yqu3@vols.utk.edu; hqi@utk.edu).

Zhenzhou Shao is with the Beijing Key Laboratory of Light-weight Industrial Robot and Safety Verification, College of Information Engineering, Capital Normal University, Beijing 100048 China. (e-mail: zshao@cnu.edu.cn).

Corresponding author: Zhenzhou Shao.

to preserve the structure of the stylized image, due to the information loss in high-level features [6]. However, as shown in Fig. 1d, post-processing may not be effective in some scenarios. By way of the *context correspondence* between the content and style images, the local-based approaches could transfer local styles successfully. *Nonetheless*, the extracted features with pre-trained models generally could not indicate the context matching information of the image pair directly. Thus, the stability of the style transfer relies heavily on the context correspondence estimated from additional supervised segmentation or classification models. Such supervised models may fail if the given image has complex or unknown objects, or has more than three channels. The failure of the models may lead to the failure of stylization. *More importantly*, these approaches perform stylization locally within each context area, thus tending to ignore the global consistency within or across objects. As shown in Fig. 1e, abrupt color changes across context regions can be easily introduced with the resulting image being less photorealistic.

There are, in general, three key challenging issues with the current stylization approaches, namely, 1) how to match the context correspondence of the image pairs without additional segmentation or classification models? 2) how to perform context-based local style transfer in a globally consistent fashion, i.e., without introducing abrupt changes/artifacts within or across semantic regions? and 3) how to preserve the structure information of the stylized image naturally?

To addressing the challenges, we exploit the potential of "non-local" features that would effectively indicate the context information of images and break the barrier between local style transfer and global consistency. We argue that "context" should not be kept as a local feature, as shown in Fig. 2, where the similar context regions, *e.g.*, the trees, usually share similar color but may scatter at disjoint locations across the image. Since such non-local similarities could not be readily captured by the current approaches, *a new representation scheme* should be developed to overcome the challenges above.

Fig. 2: The non-local nature of context. For example, the similar context regions, the trees, although possess similar color, are scattered at disjoint regions across the image.

In this paper, we propose a non-local representation scheme through dictionary-based image decomposition to address the challenges of photorealistic style transfer. In this scheme, each pixel (as a 3-D vector) can be decoupled into a linear combination of a set of color bases with the corresponding coefficients serving as the "representation" of the pixel. The physical implication of this representation is how much each color basis contributes to the color formation of that pixel. This representation thus needs to satisfy two physical constraints, *i.e.*, the sum-to-one constraint and the non-negative constraint. Since the entire image shares the same set of color bases, the representations (*i.e.*, the coefficients of color bases) for similar context would be similar regardless of their spatial location in the image. Thus, the proposed representation scheme can successfully capture the non-local similarities over the entire image, and has the potential to distinguish context according to the coefficients/proportions of the color bases. From this perspective, we consider the extracted representation to be *context-correlated*. Since such decoupling procedure is done at the pixel level, it *preserves the structure information* of the images well. In addition, the color bases of the content and style images hold an affine relationship such that the colors of the content image can be transferred to those of the style image *without structure distortion*. By enforcing the extracted representations to be sparse, for each context region, only the dominant color bases with non-zero representations would take effect in the global-affine transfer. Since the dominant color bases are generally different for different contexts, it allows for *diverse local transfer in a global-consistent fashion*. Finally, to transfer the correct color for each context, mutual information is employed to *match the context correspondence* of the representations for the image pair.

The proposed method is referred to as the non-local representation based mutual affine-transfer network, or NL-MAT. The contribution of this work is three-fold, pertaining to the three major components of the proposed NL-MAT network:

- First, a non-local representation scheme is realized which projects the images from the three-dimensional RGB color space to a k-dimensional representation space with each of the k elements reflecting the proportion of a certain color basis in making up the RGB color. This way, the context information is embedded naturally without adopting any additional models. The representation is extracted with a stick-breaking encoder to enforce the two physical constraints of the proportions, without losing images' spatial information.
- Second, an affine-transfer decoder is constructed that embeds the shared color bases of the content and style images, such that the potential transfer relationship between the image pair can be learned without structure distortion. By enforcing the sparsity of the non-local representations, we are able to perform local style transfer using the decoder, while preserving global consistency.
- Third, in order to match the colors of similar objects (or parts) in the image pairs for affine style transfer, we design a mutual discriminative network to extract the context-correspondence representations from the image pairs by maximizing the mutual information (MI) between the representations and their own RGB inputs. The statistic transformation is adopted to enforce the matched representations to have similar statistical characteristics, which further improves the photorealistic-stylization capacity of the proposed method.

The proposed representation scheme can naturally match the context correspondence between the image pair and realize diverse local style transfer in a global-consistent fashion. This is fundamentally different from existing representation

schemes that require additional segmentation models or steps to match the correspondence. To the best of our knowledge, this work is the first effort that performs photorealistic style transfer with a non-local representation model.

The rest of the paper is organized as follows. Sec. II provides an overview of the state-of-the-art photorealistic style transfer approaches. Sec. III formulates the photorealistic style transfer problem. Sec. IV elaborates and analyzes the proposed NL-MAT method. Sec. V performs comprehensive evaluations of the proposed approach. Conclusions are drawn in Sec. VI.

II. Related Work

Classical style transfer methods stylize an image in a global fashion with spatial-invariant transfer functions [1], [7], [8], [13]–[15]. These methods can handle global color shifts well, but they are limited in matching sophisticated styles with color changes [1], [3], as shown in Fig. 1c.

Recent photorealistic approaches can be generally categorized into patch-based and context-based, which perform the stylization in a local fashion. Patch-based approaches stylize images according to the patch similarity on high-level features extracted with the supervised pre-trained CNN. Liao *et al*. [10] transferred images by finding dense correspondences between the high-level patch features of the content and style images, with nearest-neighbor field search. A weighted least squares filter (WLS) [16] was adopted as a post-processing step to refine the structures of the resulting images. He *et al*. [11] further improved the stylization results by generating a guidance image based on the strategy of [10], where local style transfer can be guided in the image domain according to the guidance image, to avoid structure distortion.

Context-based approaches perform the style transfer based on the high-level features extracted from the supervised pre-trained CNN model and semantic matching obtained from the supervised pre-trained segmentation model. Luan *et al*. [1] preserved the structure of the content image by adopting a color-affine-transfer constraint and color transfer is performed according to the semantic regions generated using the pre-trained DeepLab segmentation model [17]. Mechrez *et al*. [2] proposed to maintain the fidelity of the stylized image with a post-processing step based on the screened poisson equation (SPE). Li *et al*. [3] improved the spatial consistency of the output image by adopting the manifold ranking algorithm as the post-processing step. LST [9] concatenated a linear propagation module after the stylization network to preserve the structure information of the resulting image. To address the blurry artifact caused by post-processing, WCT^2 [6] was proposed, which introduces a wavelet module in the network to preserve the structure information of the stylized image without post-processing.

Although these methods could transfer the local styles well, the effectiveness of the stylization relies heavily on the semantic correspondence estimated from additional supervised segmentation models pre-trained on a different dataset. Segmentation itself is a non-trivial task, and the failure of which may lead to the failure of stylization. Moreover, since stylization is performed within each context region, the light and color changes of different parts and materials across the entire image may not be smooth or natural. See Figs. 10, 11 and 12 for a comparison later. Recently, PhotoNAS [12] was proposed to perform smooth stylization by applying WCT [5] on the stacked multi-level features extracted from pre-trained models and the normalized skip links. However, since the context information is not considered during the optimization, the stylization is not dramatic within local areas. See Fig. 13 for a comparison later.

Based on the discussions above, Table I summarizes, from five aspects, the pros and cons of some state-of-the-art photorealistic stylization approaches, including preserving the spatial structure of the content image, realizing local style transfer while maintaining global consistency, transferring styles based on context or semantic correspondence between the content and the style images, and needing no additional models from supervised classification or segmentation. In the following, we elaborate on how the proposed approach tackles the challenges brought from each aspect.

III. Problem Formulation

As discussed in Sec. I, the key issue in obtaining a context-correspondence and high-quality photorealistic style transfer is the realization of a new representation scheme that extracts the matched non-local representations from the image pair. Such scheme is designed according to the theory of dictionary-based image decomposition, where natural images can be represented by a set of color bases with its coefficient vectors (*i.e*., representations) [18]–[20]. In this paper, the decomposition is learned through neural network by minimizing the reconstruction error of the image pair.

Given a content image, $I_c \in \mathbb{R}^{m\times n\times l}$, where m, n, and l denote its width, height, and number of channels, respectively, and a style image, $I_s \in \mathbb{R}^{M\times N\times l}$, where M, N, and l denote its width, height, and number of channels, respectively, the goal is to generate the image $I_{cs} \in \mathbb{R}^{m\times n\times l}$ with its content coming from I_c but using the style from I_s. For each image, to capture the non-local similarities, the entire image is enforced to share the same set of color bases. That is, a single pixel in the content and style images can be expressed as Eqs. (1) and (2), respectively.

$$\boldsymbol{i}_c = \mathbf{s}_c D_c \tag{1}$$

$$\boldsymbol{i}_s = \mathbf{s}_s D_s \tag{2}$$

where $\boldsymbol{i}_c \in \mathbb{R}^{1\times l}$ and $\boldsymbol{i}_s \in \mathbb{R}^{1\times l}$ denote a single pixel of the content image and the style image, respectively. $D_c \in \mathbb{R}^{k\times l}$ and $D_s \in \mathbb{R}^{k\times l}$ are two matrices with each row of which denoting the color basis that preserves the color information of the entire content and style images, respectively. $\mathbf{s}_c \in \mathbb{R}^{1\times k}$ and $\mathbf{s}_s \in \mathbb{R}^{1\times k}$ denote the corresponding coefficients for each of the k color bases of the content and style image, respectively. Note that, in our case, we have $k \gg l$. That is, the number of color bases is much larger than the dimension of the input pixels. Taking the content image as an example, the decomposition is illustrated in Fig. 3. For an individual pixel $\boldsymbol{i}_c$, Eq. (1) can be written as

$$\boldsymbol{i}_c = \mathbf{s}_c D_c = [s_1, \cdots, s_i, \cdots, s_k][\mathbf{d}_1, \cdots, \mathbf{d}_i, \cdots, \mathbf{d}_k]^T,$$

TABLE I: Capabilities of the State-of-the-art Photorealistic Stylization Approaches

	Global-based		Context-based				Patch-based		SOTA	Proposed
Capabilities	Reinhard [7]	Pitie [8]	Luan [1]	Li [3]	LST [9]	WCT^2 [6]	Liao [10]	He [11]	PhotoNAS [12]	NL-MAT
Structure preservation	✓	✓	✗	✓	✗	✓	✗	✓	✓	✓
Local style transfer	✗	✗	✓	✓	✓	✓	✓	✓	✗	✓
Context-correspondence transfer	✗	✗	✓	✓	✓	✓	✓	✓	✗	✓
Global consistency	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓
No additional models	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓

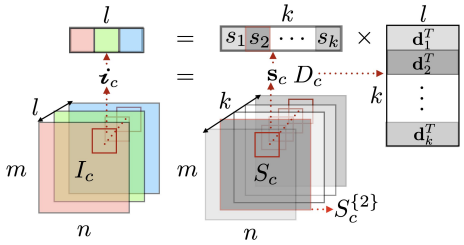


Fig. 3: Dictionary-based image decomposition. Note: The darker the shades, the larger the proportions.

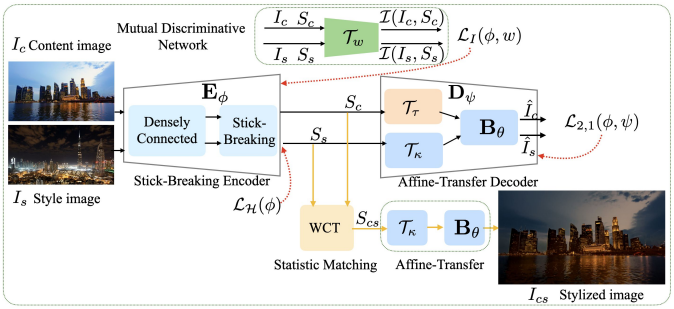


Fig. 4: Flowchart of the proposed NL-MAT.

where the transpose of $\mathbf{d}_i \in \mathbb{R}^{l\times 1}$ denotes a row vector of D_c carrying one color basis, and s_i, the ith component of $\mathbf{s}_c$, indicates the proportion (representation) of the color basis, $\mathbf{d}_i$, in making up the color of the given pixel, and $i = 1, \cdots, k$. With all the $\mathbf{s}_c$'s for each pixel, we obtain the representation $S_c \in \mathbb{R}^{m\times n\times k}$ for the entire image, where the ith slice (or plane) of the representation, $S_c^{\{i\}} \in \mathbb{R}^{m\times n}$, indicates the proportions of the ith color basis, $\mathbf{d}_i$, for all the pixels in the image. Based on the physical properties of the proportions, the non-negative and sum-to-one constraints are enforced on the representation S_c to extract the desired color bases. With such settings, similar contexts, although scatter across the image in disjoint locations, would have similar representations, thus the scheme is able to capture the non-local context-correlated information by nature.

The goal of the photorealistic style transfer is to transfer the colors of the content image to that of the style image. To find the potential transfer without structure distortion, we enforce the color bases of the image pair D_c and D_s to have an affine relationship. As analyzed in Sec. I, by further enforcing the representations to be sparse, only the dominant color bases (*e.g.*, $\mathbf{d}_2$ in Fig. 3) with larger representations (*e.g.*, s_2 in Fig. 3) would be transferred effectively in the global affine transfer. On the other hand, since different objects or parts consist of different dominant color bases, the extracted representations are *context-sensitive* with discriminative capacity, *i.e.*, the objects or parts with different colors can be easily identified by their representations. To transfer the correct colors for different contexts, the representations between the content and style images are matched through mutual information. In the next section, we elaborate on how these constraints are enforced to realize the proposed photorealistic style transfer.

IV. Proposed Method

We propose a non-local representation based mutual affine-transfer network (NL-MAT) architecture that mainly consists of three unique structures: 1) a shared stick-breaking encoder for the decoupling of non-local representations and color information of both the content and style images, 2) a sparse entropy function with affine-transfer decoder for the local style transfer in a global-consistent fashion, and 3) a mutual discriminative network to enforce the correspondence of context-sensitive representations with statistical matching between the content-style image pair. The stylized image is generated by performing both the color transfer with bases, as well as statistic matching on the corresponding representations with WCT. The unique architecture is shown in Fig. 4.

A. Overview of Network Architecture

As shown in Fig. 4, the inputs of the network are the content and style images $\mathcal{G} = \{I_c, I_s\}$ with l channels ($l = 3$ for RGB color images), and the outputs of the network are their reconstructed images $\hat{\mathcal{G}} = \{\hat{I}_c, \hat{I}_s\}$. The network decomposes both the content image I_c and the style image I_s by learning a shared encoder structure, $\mathbf{E}_\phi$, and an affine-transfer decoder structure, $\mathbf{D}_\psi$. The representation domain in the hidden layer is denoted as $\mathcal{S} = \{S_c, S_s\}$. The encoder of the network, $\mathbf{E}_\phi : \mathcal{G} \to \mathcal{S}$, maps the input data to high-dimensional representations (latent variables on the hidden layer), *i.e.*, $p_\phi(\mathcal{S}|\mathcal{G})$, and the affine transfer decoder, $\mathbf{D}_\psi : \mathcal{S} \to \hat{\mathcal{G}}$, reconstructs the images from the representations, *i.e.*, $p_\psi(\hat{\mathcal{G}}|\mathcal{S})$. The representation $\mathcal{S}$ contains the coefficients that reflect the local contribution of different color bases, and the weights of the decoders $\mathcal{T}_\tau(\mathbf{B}_\theta)$ and $\mathcal{T}_\kappa(\mathbf{B}_\theta)$ [to be explained in Eqs. (4) and (5)] serve as color bases D_c and D_s in Eqs. (1) and (2), respectively. The representation layer is built with the stick-breaking structure to naturally enforce the non-negative and sum-to-one physical properties of the proportions. This will be further elaborated in Sec. IV-B.

To encourage local style transfer, the sparsity constraint defined by the entropy function $\mathscr{L}_\mathscr{H}(\phi)$ is applied to the representation domain. Both the inputs, I_c, I_s, and the representations, S_c, S_s, are fed into the mutual discriminator $\mathcal{T}_w$ with weights w to enforce the context correspondence between S_c and S_s for a context-correspondence style transfer. The network is constructed with only fully connected layers, which are optimized according to the reconstruction error and regularized by the physical constraints incorporated. This will be further elaborated in Sec. IV-C.

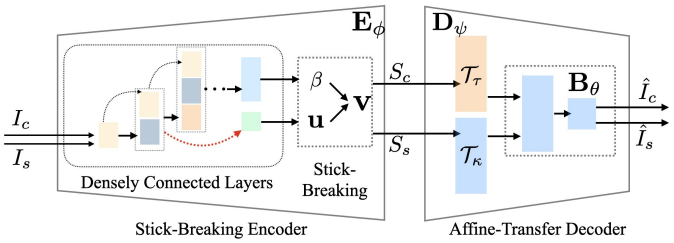


Fig. 5: Network structure of the shared stick-breaking encoder and the affine-transfer decoder.

In the stylization procedure, as shown in the lower-right part of Fig. 4, the distribution of S_c is matched with that of S_s using the whitening and coloring transformation (WCT) [5]. The transferred S_{sc} is then fed into the style's affine-transfer decoder $\mathcal{T}_\kappa(\mathbf{B}_\theta)$, to generate the stylized image I_{cs}. Note that the dashed lines in Fig. 4 show the path of back-propagation which will be further elaborated in Sec. IV-D.

B. Non-local Representation and Color Information Decoupling

As elaborated in Sec. III, both the content and style images can be decoupled to the representations (indicating proportion coefficients) and color bases (indicating color composition). However, for different images, such representations or color bases have diverse statistic distributions. To facilitate the style transfer, we construct a network with a shared stick-breaking encoder for the extraction of representations, and an affine-transfer decoder carrying transferred color information. Since the arrangement of adjacent pixels are untouched, this decoupling mechanism effectively preserves spatial/structural distribution of the content image while performing color transfer based on the style image. The detailed structure is shown in Fig. 5.

1) Representation Extraction with Shared Stick-Breaking Encoder: As discussed in Sec. III, pixels in natural images can be represented as a linear combination of a set of color bases with the coefficients satisfying two physical constraints, *i.e.*, sum-to-one and non-negativity. In order to guarantee the constraints are met, in the network design, we adopt a shared stick-breaking encoder to naturally incorporate the physical constraints. The detailed network structure of the shared stick-breaking encoder is shown in the left part of Fig. 5, where each rectangle block denotes a fully-connected layer with neurons.

The stick-breaking process can be illustrated as breaking a unit-length stick into k pieces, where the length of each piece follows the Dirichlet distribution [21]. Samples collected from the Dirichlet distribution naturally satisfy the sum-to-one and non-negativity constraints. Here, we follow the work of [22], [23], which draw the samples of representation $\mathcal{S}$ from the Kumaraswamy distribution [24]. Assuming that the row vector of representations for a single pixel is denoted as $\mathbf{s}_\rightarrow = \{s_i\}_{1\leq i\leq k}$, we have $0 \leq s_i \leq 1$, and $\sum_{i=1}^{k} s_i = 1$, where k is the number of color bases. Each variable s_i can be defined as

$$s_i = \begin{cases} v_1 & \text{for} \quad i = 1 \\ v_i \prod_{t<i}(1-v_t) & \text{for} \quad i > 1, \end{cases} \tag{3}$$

where $v_i \sim 1-(1-u_i^{\frac{1}{\beta_i}})$ is drawn from the inverse transform of the Kumaraswamy distribution. Both parameters u_i and β_i are learned through the network for each row vector, as illustrated in Fig. 5. Since $\beta_i > 0$, a softplus is adopted as the activation function [25] at the $\boldsymbol{\beta}$ layer. Similarly, a sigmoid [26] is used to map u into the $(0, 1)$ range at the $\mathbf{u}$ layer. The input of the encoder has three neurons carrying the color information of the RGB channels of each pixel in the images, and it is densely connected to all the subsequent layers by stacking the layers together to increase the representation power of the network, as shown in the left block of Fig. 5. More details are described in Sec.1 of the supplementary file.

2) Color Information Extraction with Affine-Transfer Decoder: As analyzed in Secs. I and III, to transfer the colors of the content image to those of the style image without structure distortion, we enforce the color bases of the content image, D_c, and the style image, D_s, to hold an affine relationship with the proposed affine transfer decoder. Such decoder not only carries the color information of both images but also their transfer information. The network structure of the affine-transfer decoder is shown in the right part of Fig. 5.

The transfer between the color bases could be modeled as $D_s = \mathbf{a}D_c + \mathbf{b}$. To improve the flexibility and the representative power of the network so as to facilitate decoupling, instead of relating D_c and D_s with an affine transformation modeled by $\mathbf{a}$ and $\mathbf{b}$, we express D_c and D_s as affine transformation of a shared basis weights $\mathbf{B}_\theta$, with $\mathcal{T}_\tau(\mathbf{B}_\theta)$ and $\mathcal{T}_\kappa(\mathbf{B}_\theta)$, respectively, as

$$\mathcal{T}_\tau(\mathbf{B}_\theta) = \mathbf{a}_\tau \mathbf{B}_\theta + \mathbf{b}_\tau \tag{4}$$

$$\mathcal{T}_\kappa(\mathbf{B}_\theta) = \mathbf{a}_\kappa \mathbf{B}_\theta + \mathbf{b}_\kappa, \tag{5}$$

where $\mathbf{B}_\theta$, $\mathbf{a}_\tau$, $\mathbf{b}_\tau$ and $\mathbf{a}_\kappa$, $\mathbf{b}_\kappa$ are the network structure consisting of weights $\{\theta, \tau, \kappa\}$. $\mathcal{T}_\tau(\mathbf{B}_\theta)$ and $\mathcal{T}_\kappa(\mathbf{B}_\theta)$ correspond to D_c and D_s in Eq. (1) and Eq. (2), respectively, and share the same basis weights $\mathbf{B}_\theta$. The bases of the content and style images still hold an affine relationship $\mathcal{T}_\tau(\mathbf{B}_\theta) = \mathbf{a}\mathcal{T}_\kappa(\mathbf{B}_\theta) + \mathbf{b}$, where

$$\mathbf{a} = \mathbf{a}_\kappa \mathbf{a}_\tau^{-1} \tag{6}$$

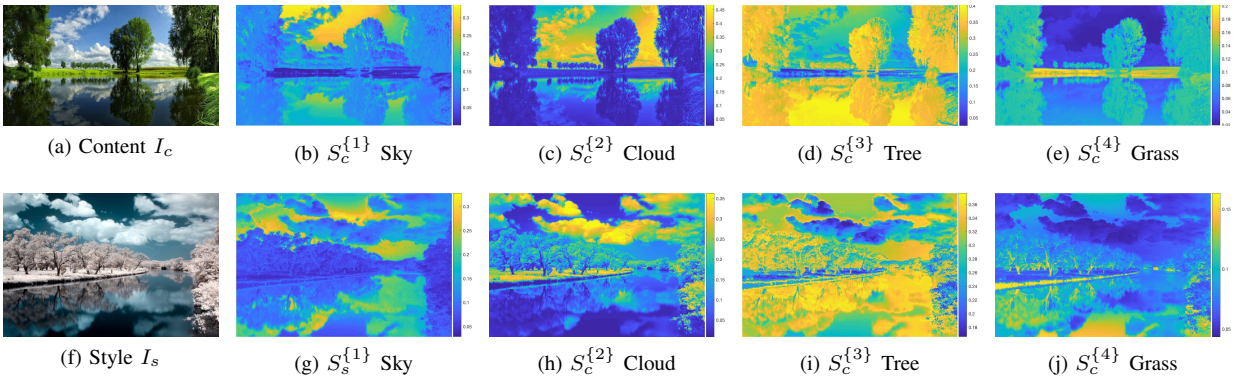
$$\mathbf{b} = \mathbf{b}_\kappa - \mathbf{a}_\kappa \mathbf{a}_\tau^{-1} \mathbf{b}_\tau, \tag{7}$$

In this way, both the color information of the content D_c and style images D_s and their transfer information $\{\mathbf{a}, \mathbf{b}\}$ are encoded into the network.

With the stick-breaking encoder and the affine-transfer decoder, the proposed scheme is able to decouple the non-local representations and their color bases of both the content and style images successfully. Since the color bases are shared for the entire image, the representations are context-correlated, enabling the capture of the non-local similarities in the image.

C. Context-Correspondence Local Style Transfer

1) Local Style Transfer with Entropy Function: For each context, its representation, extracted by the encoder of the network, indicates the proportion of each color basis in making up the given context. To enforce the local style transfer, we



(a) Content I_c (b) $S_c^{\{1\}}$ Sky (c) $S_c^{\{2\}}$ Cloud (d) $S_c^{\{3\}}$ Tree (e) $S_c^{\{4\}}$ Grass

(f) Style I_s (g) $S_s^{\{1\}}$ Sky (h) $S_c^{\{2\}}$ Cloud (i) $S_c^{\{3\}}$ Tree (j) $S_c^{\{4\}}$ Grass

Fig. 6: The representations extracted with NL-MAT from the content image (top) and style image (bottom). Ten color bases are assumed and the representations of the top four most contributing color bases are shown. Brighter color indicates higher proportion (or coefficient) value. Columns 2–5 show the representation slices of the content (top) and style (bottom) images corresponding to the proportions of their corresponding color bases. With the sparsity constraint, different objects have different dominant color bases, thus the representation is context-sensitive. With the mutual discriminative network, the extracted representations of the content and style images are encouraged to be matched with each other, *i.e.*, tree-to-tree, sky-to-sky.

encourage the representations to be sparse, such that each context is mainly constructed by a few dominant color bases determined by the corresponding representations. From color transformation perspective, the color transfer for each context is mainly related to a few dominant color bases with large representations. In this way, we are able to perform local style transfer smoothly in a global consistent fashion due to the non-local characteristic of the representation scheme. With sparse constraint, the representation itself is more discriminative. That is, the contexts with different colors are easier to be distinguished. From this perspective, we say that the proposed non-local representation scheme is context-sensitive that would enable local style transfer.

The traditional widely used l_1 regularization or Kullback-Leibler divergence [27] regularizes the sparsity of the network by reducing the summation of the representations. However, they cannot be used here to measure the sparsity as the representations are equal to one almost surely, due to the stick-breaking structure. Instead, we adopt the normalized entropy function [28], defined in Eq. (8), which decreases monotonically when the data become sparse. For example, if the representation for the ith pixel has two dimensions (*i.e.*, two color bases) with $s_1 + s_2 = 1$, the local minimum only occurs at the boundaries of the quadrants, *i.e.*, either s_1 or s_2 is zero. This nice property guarantees the sparsity of arbitrary data even under the condition that the data need to sum-to-one.

$$\mathscr{H}_p(\mathbf{s}_{\rightarrow}) = -\sum_{i=1}^{num} \frac{\|\mathbf{s}_{i\rightarrow}\|^p}{\|\mathbf{s}_{i\rightarrow}\|_p^p} \log \frac{\|\mathbf{s}_{i\rightarrow}\|^p}{\|\mathbf{s}_{i\rightarrow}\|_p^p}, \tag{8}$$

In Eq. (8), num denotes the total number of pixels of the image. For the content image $num = m \times n$, and for the style image, $num = M \times N$. We choose $p = 1$ for efficiency. The objective function for sparse loss can then be defined as

$$\mathscr{L}_{\mathscr{H}}(\phi) = \mathscr{H}_1(\mathbf{E}_\phi(I_c)) + \mathscr{H}_1(\mathbf{E}_\phi(I_s)). \tag{9}$$

Let's examine a toy example for an intuitive illustration of the effectiveness of the proposed representation scheme. Fig. 6 shows a pair of content-style images and the representation slices of the content and the style images, respectively. Note that each representation slice corresponds to the coefficients of a certain color basis in constructing the original content/style image at each spatial location; therefore, the representation slice related to a color basis is an image itself with pixel values ranging from 0 to 1. Take the content image as an example, Figs. 6b, 6c, 6d and 6e show the representation slices of the first four color bases, respectively. We can observe that, the dominant color basis of the sky is the first color basis because the sky is mostly highlighted in the first representation slice $S_c^{\{1\}}$, *i.e.*, the proportion value of the first color basis is higher for the "sky" object than for other objects in the image, as shown in Fig. 6b. Similarly, the dominant color basis of the tree is the third color basis, as shown in Fig. 6d. This simple example clearly illustrates that the proposed scheme is able to capture the non-local and context-sensitive representations, where objects (or parts) with different colors are able to be differentiated by such representations. When we transfer the representations of the entire image, different components of the transfer function would be activated according to the representations of the context, which means the network is able to perform diverse local style transfer in a global consistency fashion. That is, contexts that share the same color will be transferred in the same way even though they are not spatially adjacent.

2) Context-Sensitive Transfer through Mutual Discriminative Network: In addition to the ability of extracting context-sensitive representations and performing local transfer in a global-consistent fashion, for a context-correspondence (or semantically accurate) color transfer, the proposed scheme also needs to find the correct color matching according to the context correspondence of the objects. This will be achieved by the representation matching.

With the affine-transfer decoder, the affine relationship is learnt between the color bases of the content image, D_c, and the color bases of the style image, D_s, as shown in Fig. 7. As analyzed in the previous section, with the sparse constraint, each context is mainly constructed by a few dominant color

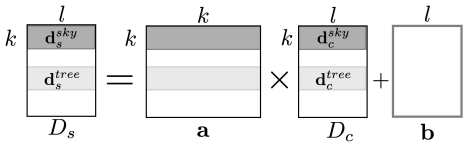


Fig. 7: The affine relationship between the bases of the content image and the style image with representation matching, *i.e.*, $D_s = \mathbf{a}D_c + \mathbf{b}$.

bases with large representation values. Let us go back to the same toy example as shown in Figs. 6 and 7. For the content image, the representation indicates the third color basis is the dominant basis of the tree, denoted as $\mathbf{d}_c^{tree}$. $\mathbf{d}_c^{tree}$ is transferred to the third color basis of the style image as shown in Fig. 7. For a context-correspondence transfer, the third color basis of the style image should also be the dominant basis of the tree, $\mathbf{d}_s^{tree}$, *i.e.*, the proportion of $\mathbf{d}_s^{tree}$ is higher for the "tree" object than for other objects in the third representation slice $S_s^{\{3\}}$ as shown in Fig. 6i. This would then imply that the representations of the content and style image should be matched. Let's take another more intuitive example: Say we take two pictures of the same object under two different lighting conditions, the color of the object would then look different, so do the corresponding color bases. However, how the color bases are combined to form the color of the object, *i.e.*, the proportions, would remain the same for the object.

In essence, to perform semantically-accurate color transfer, the distributions of the representations extracted from the same context should be similar in both the content and style images. Such correspondence can be encouraged by maximizing the dependency between S_c and S_s. Since our encoder is non-linear, traditional constraints like correlation may not catch such dependency. Instead, we maximize the dependency by maximizing their mutual information.

We propose a mutual discriminative network based on mutual information to enforce the correspondence of the extracted representations. The network structure is shown in Fig. 8.

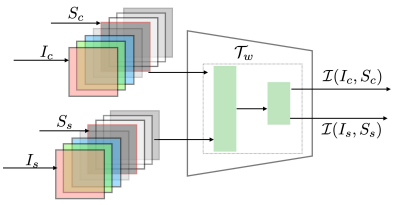


Fig. 8: Structure of the mutual discriminative network.

Mutual information (MI) has been widely used for multi-modality registrations [29], [30]. It is a Shannon-entropy based measurement of mutual independence between two random variables, *e.g.*, S_c and S_s. The mutual information $\mathcal{I}(S_c; S_s)$ measures how much uncertainty of one variable (S_c or S_s) is reduced given the other variable (S_s or S_c). Mathematically, it is defined as

$$\begin{aligned}\mathcal{I}(S_c; S_s) &= H(S_c) - H(S_c|S_s) \\ &= \int_{\mathcal{S}_c \times \mathcal{S}_s} \log \frac{\mathbb{P}_{S_c S_s}}{\mathbb{P}_{S_c} \otimes \mathbb{P}_{S_s}} d\mathbb{P}_{S_c S_s},\end{aligned} \tag{10}$$

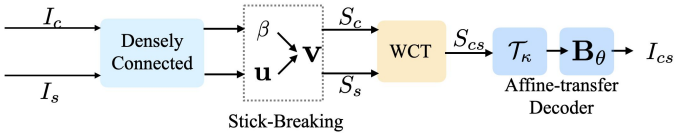


Fig. 9: Flowchart of the style transfer procedure.

where H indicates the Shannon entropy, $H(S_c|S_s)$ is the conditional entropy of S_c given S_s. $\mathbb{P}_{S_c S_s}$ is the joint probability distribution, and $\mathbb{P}_{S_c} \otimes \mathbb{P}_{S_s}$ denotes the product of marginals.

In our problem, since $S_c = \mathbf{E}_\phi(I_c)$ and $S_s = \mathbf{E}_\phi(I_s)$, their MI can also be expressed as $\mathcal{I}(\mathbf{E}_\phi(I_c); \mathbf{E}_\phi(I_s))$. However, it is difficult to maximize their dependency by maximizing their MI through MI estimator directly, because the resolution of the content and style images might be different. Instead, we maximize the average MI between the representations and their own inputs, *i.e.*, $\mathcal{I}(I_c, \mathbf{E}_\phi(I_c))$ and $\mathcal{I}(I_s, \mathbf{E}_\phi(I_s))$, simultaneously, through the same discriminative network $\mathcal{T}_w$. Note that, in the representation space, S_c and S_s are context-sensitive, so their distributions are related to the distributions of objects in their images, not their color information. When we maximize the MI with the same $\mathcal{T}_w$, the dependency between I_c and $\mathbf{E}_\phi(I_c)$ would be similar to that of I_s and $\mathbf{E}_\phi(I_s)$, *i.e.*, the slices of $\mathbf{E}_\phi(I_c)$ and $\mathbf{E}_\phi(I_s)$, which carry the context distribution information, are encouraged to be similar if they possess similar objects.

Let's take $\mathcal{I}(I_c, \mathbf{E}_\phi(I_c))$ as an example. It is equivalent to Kullback-Leibler (KL) divergence [31] between the joint distribution $\mathbb{P}_{I_c \mathbf{E}_\phi(I_c)}$ and the product of the marginals $\mathbb{P}_{I_c} \otimes \mathbb{P}_{\mathbf{E}_\phi(I_c)}$. Such MI can be maximized by maximizing the KL-divergence's lower bound based on the Donsker-Varadhan (DV) representation [32]. Since we do not need to calculate the exact MI, we introduce an alternative lower bound based on Jensen-Shannon which works more stable than DV-based objective function [33].

The mutual discriminative network, $\mathcal{I}_w : \mathcal{G} \times \mathcal{S} \to \mathbb{R}$, is constructed by fully-connected layers with weights w. The raw image and its extracted representations are stacked and fed into the network as shown in Fig. 8. The MI estimator can be defined as

$$\begin{aligned}\mathcal{I}_{\phi,w}(I_c, \mathbf{E}_\phi(I_c)) \quad &= \mathbb{E}_{\mathbb{P}}[-sp(-\mathcal{T}_{\phi,w}(I_c, \mathbf{E}_\phi(I_c)))] \\ &-\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[sp(\mathcal{T}_{\phi,w}(I'_c, \mathbf{E}_\phi(I_c)))],\end{aligned} \tag{11}$$

where $sp(x) = \log(1 + e^x)$ and I'_c is an input sampled from $\tilde{\mathbb{P}} = \mathbb{P}$ by randomly shuffling the input data. The term carrying the shuffling data is called the negative sample. Combined with the MI of I_s, our objective function is defined as

$$\mathcal{L}_\mathcal{I}(\phi, w) = \mathcal{I}_{\phi,w}(I_c, \mathbf{E}_\phi(I_c)) + \mathcal{I}_{\phi,w}(I_s, \mathbf{E}_\phi(I_s)) \tag{12}$$

By maximizing $\mathcal{L}_\mathcal{I}(\phi, w)$, we could extract optimized representations S_c and S_s that can best represent I_c and I_s, and the slices of S_c and S_s are ordered in a similar way as shown in Fig. 6. For example, the third slice of S_c carries the spatial distribution information of the tree, and the third slice of S_s also carries the distribution of the tree object in the style image. Hence S_c and S_s have been encouraged to be matched semantically, achieving context-sensitive representation.

D. Style Transfer and Implementation Details

1) Style Transfer with WCT and Affine-Transfer Decoder: As analyzed in Sec. IV-C, the learned color bases embed the transfer information between the content and style images, thus we can achieve preliminary style transfer results by feeding the representation of the content image to the affine-transfer decoder of the style image.

In the ideal case, the distributions of the matched representations S_c and S_s would be similar for the same type of context. Depending on the similarity between the content and the style images, the matching, however, might not be perfect due to content differences between the content and the style images. In order to further match the statistic characteristics of S_c to that of S_s, we adopt the classical signal whitening and coloring transforms (WCTs) approach [5], which changes the covariance of S_c to that of S_s. The toy example in Fig. 15 demonstrates that WCT could move the representations in S_c to the matched representations in S_s effectively. Then the transferred representations S_{cs} is fed into the style decoder to generate the stylization image I_{cs}. The style transfer procedure is illustrated in Fig. 9, and it will be further demonstrated using a proposed visualization mechanism in Sec. V-C.

2) Implementation Details: In order to extract better color bases, we adopt the $l_{2,1}$ norm [34] instead of the traditional l_2 norm for reconstruction loss. The objective function for $l_{2,1}$ loss is defined as

$$\begin{aligned}\mathscr{L}_{2,1}(\phi,\psi) &= \|D_\psi(E_\phi(I_c)) - I_c\|_{2,1} \\ &+\|D_\psi(E_\phi(I_s)) - I_s\|_{2,1},\end{aligned} \tag{13}$$

where $\|X\|_{2,1} = \sum_{i=1}^{m}\sqrt{\sum_{j=1}^{n} X_{i,j}^2}$. $l_{2,1}$ can be treated as first applying l_2 norm on each pixel, then applying l_1 norm to enforce the reconstruction errors on the entire image to be sparse. In this way, it will enforce most of the reconstruction errors of individual pixels to be zero. That is, the network is designed to learn individual pixels as accurate as possible, which would extract better color bases to further facilitate the style transfer.

The objective function of the proposed network architecture can then be expressed as:

$$\mathscr{L}(\phi,\psi,w) = \mathscr{L}_{2,1}(\phi,\psi) + \alpha\mathscr{L}_{\mathscr{H}}(\phi) - \lambda\mathscr{L}_{\mathscr{I}}(\phi,w), \tag{14}$$

where α and λ are the parameters that balance the trade-off among the reconstruction error, the sparse loss, and the negative of mutual information. The network is optimized with back-propagation as illustrated in Fig. 4 with red-dashed lines. More details of the WCT transfer and network structure are described in Sec.1 of the supplementary file.

V. Experimental Results

The stylization results of the proposed NL-MAT on various types of photos in two datasets from [1] and [12] are compared with those from the state-of-the-art methods, including two patch-similarity based methods [10], [11], four context-based methods [1], [3], [6], [9], and PhotoNAS [12]. For the methods we compare to, we either use the published results provided by the authors or generate the results from published pre-trained models. Both visual comparisons (Sec. V-A) and user study results (Sec. V-B) are provided to evaluate the effectiveness of the proposed method both qualitatively and quantitatively. Experiments are also conducted to show the important role of the non-local representation scheme (Sec. V-C) as well as its contributing components, *i.e.*, the affine-transfer decoder (Sec. V-D) and the mutual-discriminative network (Sec. V-E). The effect of the number of color bases is discussed in Sec. V-F. Computational efficiency and some failure cases are discussed and analyzed in Secs.2 and 3 of the supplementary file, respectively.

A. Visual Comparison

Fig. 10 shows visual results of the proposed method as compared to the patch-based photorealistic stylization methods, *i.e.*, Liao *et al*. [10] and He *et al*. [11]. We can observe that patch-based methods are able to perform local style transfer on the content image successfully, because they could find the correspondence between the image pairs according to the patch similarity measured with the help of the pre-trained VGG-net. However, although post-processing is applied to smooth the reconstructed result, Liao *et al*. [10] still suffers from abrupt color changes within or across objects, as shown in Fig. 10. For example, the mountain in the 2nd row, the building in the 3rd row, the background in the 4th four, the floor in the 6th row, and the tree in the 7th row do not have smooth color transitions. This is because patch-based methods mainly focus on style transfer in the local area while neglecting the global consistency within or across objects. He *et al*. [11] transfers style better than Liao *et al*. [10] because it optimizes a local linear model for color transfer satisfying both local and global constraints. However, due to patch-similarity, the style of one patch from one object may be matched to similar patches belonging to other objects. For example, in the 4th row, the background of the car is transferred with the same color as the car, and its windshield has abrupt color changes. In the 6th row, the floor is transferred with the same color of the furniture. In the 8th row, the sky is transferred with the color of the tree. As a comparison, the proposed method is able to generate photorealistic results without color inconsistency caused by patch mismatch. This is largely due to the realization of the proposed representations scheme, which can capture the non-local representations with matched context information that facilitates the local style transfer with global consistency.

Figs. 11 and 12 show visual results of the proposed method as compared to the context-based photorealistic stylization methods, *i.e.*, Luan *et al*. [1], Li *et al*. [3], LST [9], and WCT^2 [6]. The generated results from Luan *et al*. [1] and LST [9] can preserve the spatial structure well with the local color affine transfer constraint/filter. However, they tend to cause color inconsistency, especially in homogeneous areas, as shown in the 3rd column of Figs. 11 and 12. With the post-processing smoothing step, Li *et al*. [3] generates better results. However, the results have some blurry artifacts introduced by the post-processing. WCT^2 [6] produces smoother results with less artifacts due to the adopted wavelet module, as shown in the 4th column of Fig. 12. These methods can successfully transfer the color style to the content image in

Fig. 10: Visual comparison with patch-based photorealistic methods. 1st column: reference style image. 2nd column: content image. 3rd column: Liao *et al*. [10]. 4th column: He *et al*. [11]. 5th column: proposed NL-MAT.

most scenarios, based on the pre-trained segmentation model. However, they tend to fail when the contexts are mismatched or not recognized by the pre-trained segmentation model. For example, observe the style-content image pair in the 3rd to the last row of Fig. 11, where the red umbrella is in the semantic label of the style image but not in that of the content image, the red color is transferred, by both Luan *et al*. [1] and Li *et al*. [3] methods, to the content image at around the same spatial location as it appears in the style image even though semantically, there is no additional object at that location in the content image. Another example is shown in the 2nd row of Fig. 12, where the shadow of the flowers cannot be recognized by the pre-trained model, thus the methods fail to transfer the correct style. In addition to the problem caused by pre-trained segmentation model, all these methods still exhibit abrupt color changes between different semantic regions. This is because they perform region-based transfer within the semantic regions of the content and style images without adequately considering the global consistency.

On the contrary, the images generated by the proposed method can not only transfer the color style correctly but also preserve the natural color transitions among neighborhood pixels, especially the transitions between different contexts. The key contribution to the performance gain is that, instead of relying on additional segmentation models, the proposed representation scheme is able to extract matched context-sensitive non-local representations based on the characteristics of the context with the sparse constraint and mutual-discriminative network. With this scheme, we are able to transfer the style locally with WCT and affine-transfer decoder in a globally consistent fashion to produce more photorealistic photos with desired styles.

To further evaluate the capability of the proposed scheme in conducting local transfer with global consistency, we conduct experiments on high-resolution images and compare the results with that of the state-of-the-art method PhotoNAS [12]. From Fig. 13, we can observe that PhotoNAS is able to preserve the

Fig. 11: Visual comparison with context-based photorealistic methods. 1st column: reference style image. 2nd column: content image. 3rd column: Luan *et al*. [1]. 4th column: Li *et al*. [3]. 5th column: proposed NL-MAT.

Fig. 12: Visual comparison with context-based photorealistic methods. 1st column: reference style image. 2nd column: content image. 3rd column: LST *et al*. [9]. 4th column: WCT^2 *et al*. [6]. 5th column: proposed NL-MAT.

Fig. 13: Visual comparison with PhotoNAS. 1st column: reference style image. 2nd column: content image. 3rd column: PhotoNAS [12]. 4th column: proposed NL-MAT.

consistency very well in most cases, because it performs the stylization globally on stacked multi-level features extracted from pre-trained models. However, it tends to ignore the context information thus could not transfer dramatic color changes within local areas, as shown in the 3rd column of Fig. 13. On the other hand, the proposed scheme can not only perform local style transfer according to their contexts, but also preserve the color consistency of the images. Thus it is able to generate more photorealistic images with very fine details.

B. User Study

Since the evaluation of photorealistic style transfer tends to be subjective, we conduct two user studies to further validate the proposed method quantitatively. One study asks users to select the result that better carries the style of the reference style image. The other one asks users to select the result that looks more like a real photo without artifacts. We choose 30 images of different scenes from the benchmark dataset offered by Luan *et al*. [1] and PhotoNAS [12] and collect responses from Amazon Mechanical Turk (AMT) platform for both studies. The proposed method is compared with photorealistic stylization methods including patch-based (Liao *et al*. [10] and He *et al*. [11]), context-based (Luan *et al*. [1], Li *et al*. [3], LST [9], WCT^2 [6]), and PhotoNAS [12]. For each study, there are totally 210 questions. For each question, we show the AMT workers a pair of content and style images and the result of our method and one other method. Each question is answered by 30 different workers. Thus the evaluation is based on 6,300 responses for each study. The feedback is summarized in Table II. We can observe that, compared to the other photorealistic transfer methods, our method can not only stylize the image well but also generate more photorealistic images. Note that our method only need one pair of data, *i.e*., the content and style image without any additional segmentation or classification models, to generate such results.

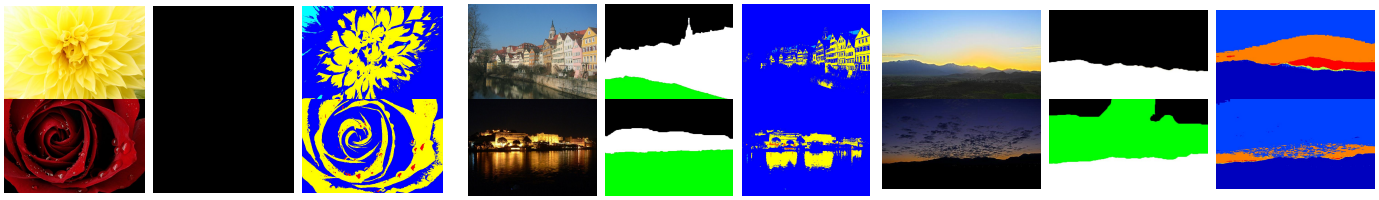
TABLE II: User study. ($x\%/y\%$ indicates that for each evaluation, $x\%$ users think the other method is better and $y\%$ users think the proposed NL-MAT is better.)

Methods	Better Stylization	Photorealistic
Liao *et al*. [10]/ours	43.76%/**56.24%**	39.44%/**60.56%**
He *et al*. [11]/ours	37.67%/**62.33%**	31.89%/**68.11%**
Luan *et al*. [1]/ours	30.33%/**69.67%**	21.17%/**78.83%**
Li [3]/ours	32.83%/**67.17%**	24.0%/**76.0%**
LST [9]/ours	38.56%/**61.44%**	31.22%/**68.78%**
WCT^2 [6]/ours	25.74%/**74.26%**	22.22%/**77.78%**
PhotoNAS [12]/ours	26.11%/**73.89%**	21.56%/**78.44%**

C. Effect of Non-local Representation Scheme

The key to the success of NL-MAT lies in the capability of decoupling the color bases from the matched context-sensitive non-local representations, where the discriminative capacity of the representations reflects the effectiveness of the local style transfer and the correct representation matching indicates the correct color transfer. To better demonstrate the capability of the proposed scheme, we develop a representation visualization mechanism to show the reasoning of the scheme and how the representations are matched after decoupling.

1) Visualization Mechanism for Representation: As described in Sec. IV, the representation vector of a single pixel, $\mathbf{s}_{\rightarrow} = \{s_i\}_{1\leq i\leq k}$ indicates the proportion of each of the k color basis in making up the given pixel. By introducing the



(a) (b) (c)

Fig. 14: Major color index map (MCIM) of the decoupled representations from the proposed NL-MAT. The first column of each group shows the content (top) and style (bottom) images, respectively. The second column of each group shows the segmentation maps of the content and style images from the method of Luan [1]. The third column of each group shows the MCIM of the content and style images from the proposed method, respectively. Note that the pseudo-colors in the MCIMs are arbitrarily selected to represent the indices of the largest color basis. Thus the color itself is not important. It is the matching between the content MCIM and the style MCIM that matters.

sparse constraint, the representations are more discriminative, which allows for diverse local transfer in a global consistent fashion. Since the transfer of each context is mainly determined by dominant color bases that have larger representation values, the visualization mechanism is designed based on such dominant bases, which is referred to as the "major color index map (MCIM)." The MCIM of each image is constructed by the "index" of the largest representation of each pixel, which indicates the most important color basis for that given pixel. Mathematically, it can be expressed as

$$t = \arg\max_{j} \{s_1, \cdots, s_j, \cdots, s_k\} \tag{15}$$

where k is the number of color bases. With the indices of the largest representation of all the pixels, we are able to define the MCIM of the entire image.

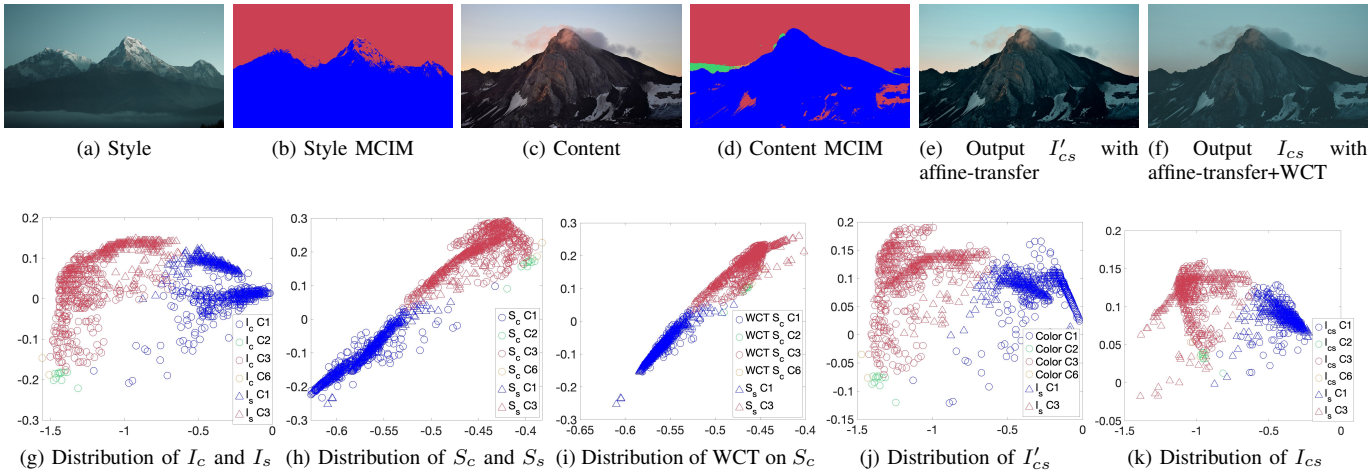
The MCIMs of toy examples are shown in Fig. 14, where each pseudo-color indicates a single index. Note that, since some objects may have more than one major color basis, the index map only roughly segments the image. Nonetheless, MCIM allows us to perform in-depth visual inspection on how the proposed method works in different scenarios. From Fig. 14, we can observe that, the extracted representations from similar objects or parts in the image pairs are context-sensitive and matched in different scenarios, even if the matched objects/parts are with different colors. This is the main reason why NL-MAT can realize local transfer while preserving global consistency, as shown in Figs. 10 to 13. It is worth mentioning that, the proposed NL-MAT is unsupervised and does not need any additional models or steps for segmentation to perform photorealistic stylization. As a comparison, even though the segmentation method adopted by Luan *et al*. [1] is trained in a supervised way, it may not handle unknown objects or objects with complex components as shown in Fig. 14, which may affect the stylization performance.

2) Reasoning of Scheme: Since the style transfer problem can be explained as mapping the distribution of the content image to that of the style image [8], the stylized result should have similar distribution to that of the style image. To further illustrate why the proposed scheme works, we draw the distributions of intermediate results as well as each context, by way of MCIM, to visualize the effect of stylization in Fig. 15.

The distributions are visualized by projecting the raw image (I_c, I_s) or representations (S_c, S_s) onto a two-dimensional space using the singular value decomposition (SVD) method. Fig. 15g shows the distributions of both the raw content and style images before stylization. The projected pixels are colored according to the contexts identified by MCIM, with circles indicating those from the content image and triangles for those from the style image. For example, the blue and red circles denote the pixels from the mountain area (I_c C1) and the sky area (I_c C3) of the content image, respectively. We can observe from Fig. 15g that, the blue circles (I_c C1) and blue triangles (I_s C1), although indicating the same context (*i.e*., the mountain), belong to the two different blue clusters. This indicates that the distributions of the raw image pair are quite different from each other. Fig. 15h presents the distributions of the representations S_c and S_s for the content and style images, respectively. We can observe that the same context of the S_c and S_s, *e.g*., the blue circles (S_c C1) and blue triangles (S_s C1), overlap into one blue cluster. This indicates that when we project the raw image onto the representation space with the stick-breaking encoder, affine-transfer decoder, sparse constraint and mutual discriminative network, the representations of the same context in the content and style images reveal similar characteristics. This is in consistent with the previous analysis and the matching contexts in the MCIM.

By feeding the extracted representations S_c to the affine-transfer decoder, we can achieve preliminary stylized result I'_{cs} as shown in Fig. 15e. From Fig. 15j, we can observe that the distribution of I'_{cs} is closer to that of the style image, as compared to that of the content image. Nonetheless, there are still two apparent blue clusters in the distribution plot. To further match the representations, we conduct WCT on S_c and show their distributions in Fig. 15i. We observe that the same contexts of S_c move closer to that of the S_s, presenting one dense blue cluster. Therefore, with the WCT on S_c, the distribution of the generated result I_{cs} (*e.g*., the blue circle cluster) is quite similar to that of the style image I_s (*e.g*., the blue triangle cluster) as shown in Fig. 15k. As a result, the transferred image shown in Fig. 15f carries the style of Fig. 15a better than I'_{cs}.

Due to the complexity of real applications, the context of the content and style images may not be matched perfectly. For example, from Fig. 15d, the content MCIM shows a context area marked in green that does not have any matching areas



(a) Style (b) Style MCIM (c) Content (d) Content MCIM (e) Output I'_{cs} with affine-transfer (f) Output I_{cs} with affine-transfer+WCT

(g) Distribution of I_c and I_s (h) Distribution of S_c and S_s (i) Distribution of WCT on S_c (j) Distribution of I'_{cs} (k) Distribution of I_{cs}

Fig. 15: Reasoning of the proposed NL-MAT using MCIM and the distribution plots masked by MCIM. The raw images (I_c, I_s) and representations (S_c, S_s) are projected onto a two-dimensional space using SVD. Different colors indicate different context of the MCIM. C# denotes the index of the color bases for the corresponding context. Circles and triangles denote the pixels or representations belonging to the content and style images, respectively. The distribution of the stylized image is similar to that of the style image with the affine-transfer, and it became closer with WCT. Note that for each MCIM masked region (or context), we vectorize pixels in that region into a column vector and pick every 1000th element of the vector for display purpose, so that the density change of the distribution is more easily observed.

in the style MCIM. Nonetheless, from Fig. 15g, we observe this green area denoted by green circles (I_c C2) is closer to the red circles indicating the sky (I_c C3). From Fig. 15h, we see that the extracted representations of C2 (green circles S_c C2) are also close to that of C3 in the content image (sky red circles). When WCT is performed on the representations of C2, the green cluster stay close to that of the C3 of the style image (red triangle, I_s C3). Hence the C2 area of the stylized image looks still natural as shown in Fig. 15f.

D. Ablation Study on Affine-Transfer Decoder

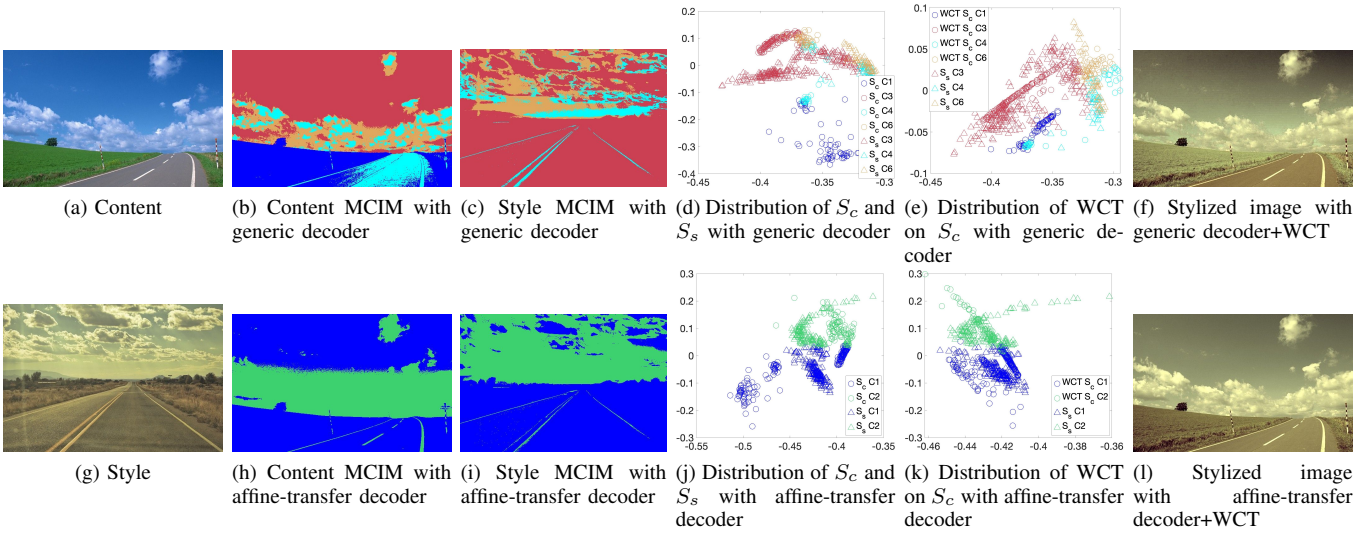
One of the contributing factors to the effectiveness of the non-local representation scheme is the proposed affine-transfer decoder, which allows the network to learn the color bases of both the content and style images as well as the transfer between them. This decoder, along with the mutual-discriminative network and the sparse constraint, allows the network to match representations of similar contexts regardless of their actual color content.

To demonstrate the importance of the affine-transfer decoder, we replace it with a generic fully-connected decoder and show the results in Fig. 16. We observe that, with a shared generic decoder for both the content and style images, the network could not match the representations of the image pair well even with the enforced constraints. For example, we can observe from Fig. 16d that the extracted representations of the "grass" from the content image, blue circles (S_c C1), are far away from that of the same context from the style image, red triangles (S_s C3). This scattered distribution indicates the extracted representations are not matched well. On the contrary, as shown in Fig. 16j, with the affine transfer decoder, the representations of the "grass" from the content image (blue circles S_c C1) and the style image (blue triangles S_s C1) are much closer to each other as compared to the distribution in Fig. 16d, showing a better match of similar contexts between the content and style images. Thus, when WCT is applied, the distributions of the representations from the affine-transfer decoder are closer as compared to those from the generic decoder, as shown in Figs. 16k and 16e, respectively. As a result, the proposed method is able to obtain more photorealistic image carrying more natural styles as compared to the one with the generic decoder.

E. Ablation Study on Context-sensitive Local Color Transfer

In addition to the affine-transfer decoder evaluated in Sec. V-D, the two other important components of the proposed method are the sparse constraint, used to increase the discriminative capacity of local-color transfer, and the mutual discriminative network, used to enforce the representations of the content and style images to have context correspondence. To evaluate the contribution of these two components, we perform ablation study by changing the weight parameters, α on sparsity constraint and λ on mutual information as in Eq. 14. The evaluation results are demonstrated in Fig. 17.

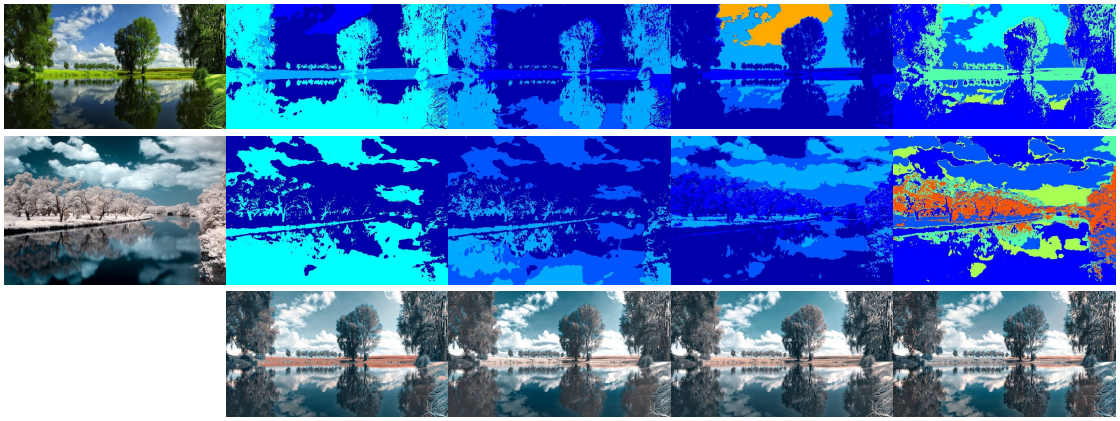
We can observe that, when both α and λ are zero, *i.e.*, no sparsity or mutual information loss is considered, the proposed method could still generate reasonably good stylized images with spatial structure well preserved, showing the effectiveness of the non-local representation. However, the color was not adequately transferred, as can be seen from the color of the tree and the grass in Fig. 17b. When we gradually increase the sparse parameter α, the discriminative capacity is increased as can be observed from the MCIMs (Figs. 17c-17e) with more subtle segments. Also with such sparsity, we observe local colors start to change drastically but with global consistency. However, since the network does not encourage the correspondence between the representations of the content and style images as $\lambda = 0$, the color may not be transferred adequately when the representations do not match. When we increase the parameter λ to include the



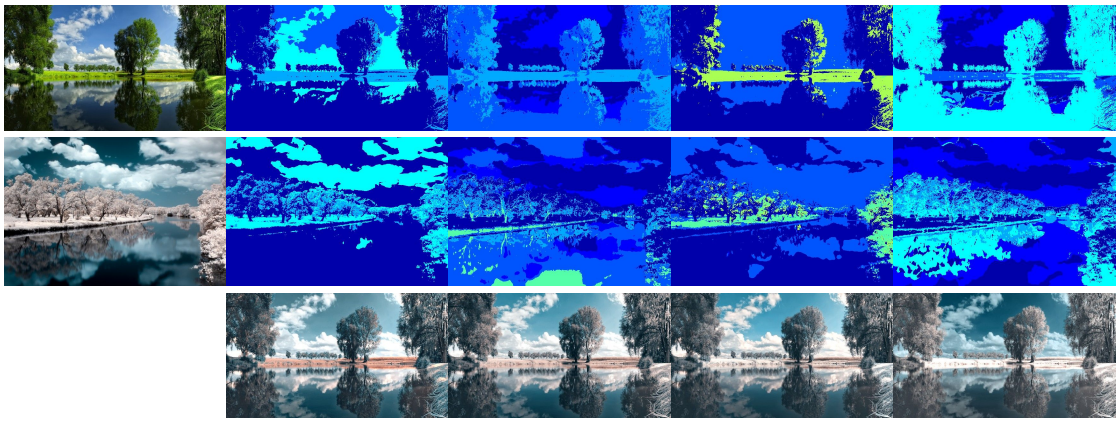
(a) Content (b) Content MCIM with generic decoder (c) Style MCIM with generic decoder (d) Distribution of S_c and S_s with generic decoder (e) Distribution of WCT on S_c with generic decoder (f) Stylized image with generic decoder+WCT

(g) Style (h) Content MCIM with affine-transfer decoder (i) Style MCIM with affine-transfer decoder (j) Distribution of S_c and S_s with affine-transfer decoder (k) Distribution of WCT on S_c with affine-transfer decoder (l) Stylized image with affine-transfer decoder+WCT

Fig. 16: The effect of affine-transfer decoder.



(a) Content and style images (b) $\alpha = 0$, $\lambda = 0$ (c) $\alpha = 0.0001$, $\lambda = 0$ (d) $\alpha = 0.01$, $\lambda = 0$ (e) $\alpha = 0.1$, $\lambda = 0$

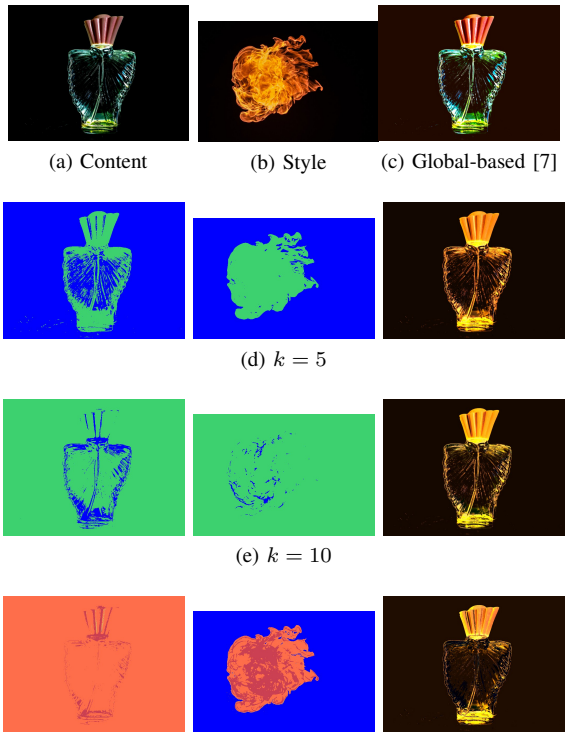


(f) Content and style images (g) $\alpha = 0.01$, $\lambda = 0.0001$ (h) $\alpha = 0.01$, $\lambda = 0.01$ (i) $\alpha = 0.01$, $\lambda = 1$ (j) $\alpha = 0.01$, $\lambda = 5$

Fig. 17: Effects of including the sparsity constraint and the mutual discriminative network by adjusting the parameters α and λ, respectively. (b)-(e): stylized images with $\lambda = 0$ and α changed from $\{0, 0.0001, 0.01, 0.1\}$, respectively. Top: MCIM of the content image. Middle: MCIM of the style image. Bottom: stylized images. (g)-(j): stylized images with $\alpha = 0.01$ and λ changed from $\{0.0001, 0.01, 1, 5\}$. Note mainly the color changes of the tree, sky, and grass in the stylized image as the parameters are adjusted.

mutual information loss, the extracted representations are more correlated with each other, resulting in a more photorealistic local style transfer, as shown in Figs. 17g-17j. We observe that the color of the trees starts showing the snowy effect and the color of grass is mostly white. The color transfer even affects the reflection of the trees in the water, as shown in Figs. 17i-17j.

F. How to Choose the Number of Color Bases



(a) Content (b) Style (c) Global-based [7]

(d) $k = 5$

(e) $k = 10$

(f) $k = 15$

Fig. 18: Stylized images with different numbers of color bases k with and small sparse constraint $\alpha = 0.001$. In (d), (e), and (f), from left to right: MCIM of the content (left) and style (middle) images, and the resulting stylized image with the corresponding k value.

Generally speaking, to extract accurate color bases, the number of bases k should be large enough to encompass the different contexts so as to reconstruct the RGB images with high fidelity. Large k also allows more flexible local color transfer for different contexts. In general, $k = 10$ works well for both datasets adopted in this paper. However, if there is less number of colors in the image pair and the sparse constraint is set to a small value, *e.g.*, $\alpha = 0.001$, we find that the stylization is more effective when k is smaller, as shown in Fig. 18. In this toy example, the number of colors in the image pair is small, thus, $k = 5$ is sufficient to extract a set of effective color bases, as shown in Fig. 18d. However, as we gradually increase the k value with α fixed at 0.001, the extracted representations start losing the discriminative power and the matching between representations of the content and style images start to deteriorate, as shown in Fig. 18f. This is because for images with only a few different colors, if k is set to a large value, without effective sparse constraints, it tends to learn duplicated color bases. Thus, the representations will not be context-sensitive, which would affect the style transfer. Therefore, in this case, we can either increase α or decrease k for an effective transfer. It is worth mentioning that, even with $k = 15$, the stylized results from the proposed method is still more effective than that from the global-based method, as shown in Fig. 18c.

VI. Conclusion and Future Work

To tackle the problem of photorealistic style transfer, we proposed a non-local representation scheme realized with a mutual affine-transfer network (NL-MAT). To the best of our knowledge, this work represents the first attempt to address the photorealistic style transfer problem through a non-local representation model. The proposed scheme successfully decouples the image pairs into non-local representations and color information, with a stick-breaking encoder and an affine-transfer decoder. By enforcing the sparsity with the entropy function and representation correspondence with the mutual discriminative network, the method is able to extract context-sensitive and matched representations. This largely facilitates context-correspondent local style transfer in a global-consistent fashion. Experimental results demonstrated that the proposed NL-MAT is able to generate photorealistic photos without abrupt color changes or needing any additional models for segmentation or classification.

The proposed scheme works well in most scenarios, even when there are some mismatches between the content and style images, as discussed in Sec. V-C. Nonetheless, NL-MAT does have its limitations and may fail on some challenging image pairs. For example, if the distributions of the objects in the content and style images are very different, it could result in large semantic mismatching. Another type of typical failure cases may occur when different semantic contexts in the image actually possess very similar color. The failure cases are further discussed in the supplementary file (Sec. 3). In our future work, we will exploit the usage of prior knowledge serving as additional physical constraints to regulate the learning process and enforce semantic matching.

Acknowledgment

The authors would like to thank all the developers of the evaluated methods who kindly offer their codes or results, and the anonymous reviewers who have helped us greatly in improving the quality of this paper.

References

[1] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep photo style transfer," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4990–4998, 2017.

[2] R. Mechrez, E. Shechtman, and L. Zelnik-Manor, "Photorealistic style transfer with screened poisson equation," *arXiv preprint arXiv:1709.09828*, 2017.

[3] Y. Li, M.-Y. Liu, X. Li, M.-H. Yang, and J. Kautz, "A closed-form solution to photorealistic image stylization," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 453–468, 2018.

[4] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.

[5] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," *Advances in neural information processing systems*, pp. 386–396, 2017.
[6] J. Yoo, Y. Uh, S. Chun, B. Kang, and J.-W. Ha, "Photorealistic style transfer via wavelet transforms," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 9036–9045, 2019.
[7] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.
[8] F. Pitie, A. C. Kokaram, and R. Dahyot, "N-dimensional probability density function transfer and its application to color transfer," *Tenth IEEE International Conference on Computer Vision (ICCV)*, vol. 2, pp. 1434–1439, 2005.
[9] X. Li, S. Liu, J. Kautz, and M.-H. Yang, "Learning linear transformations for fast arbitrary style transfer," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
[10] J. Liao, Y. Yao, L. Yuan, G. Hua, and S. B. Kang, "Visual attribute transfer through deep image analogy," *arXiv preprint arXiv:1705.01088*, 2017.
[11] M. He, J. Liao, D. Chen, L. Yuan, and P. V. Sander, "Progressive color transfer with dense semantic correspondences," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 2, p. 13, 2019.
[12] J. An, H. Xiong, J. Huan, and J. Luo, "Ultrafast photorealistic style transfer via neural architecture search," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 10 443–10 450, 2020.
[13] S. Bae, S. Paris, and F. Durand, "Two-scale tone management for photographic look," vol. 25, no. 3, pp. 637–645, 2006.
[14] F. Pitié, A. C. Kokaram, and R. Dahyot, "Automated colour grading using colour distribution transfer," *Computer Vision and Image Understanding*, vol. 107, no. 1-2, pp. 123–137, 2007.
[15] D. Freedman and P. Kisilev, "Object-to-object color transfer: Optimal flows and smsp transformations," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 287–294, 2010.
[16] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski, "Edge-preserving decompositions for multi-scale tone and detail manipulation," vol. 27, no. 3, p. 67, 2008.
[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
[18] I. Omer and M. Werman, "Color lines: image specific color representation," *Proceedings of IEEE computer society conference on Computer vision and pattern recognition (CVPR)*, pp. 946–953, 2004.
[19] P.-Y. Laffont, A. Bousseau, S. Paris, F. Durand, and G. Drettakis, "Coherent intrinsic images from photo collections," *ACM Transactions on Graphics*, vol. 31, no. 6, 2012.
[20] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," *2009 IEEE 12th International Conference on Computer Vision (ICCV)*, 2010.
[21] J. Sethuraman, "A constructive definition of dirichlet priors," *Statistica sinica*, pp. 639–650, 1994.
[22] E. Nalisnick and P. Smyth, "Deep generative models with stick-breaking priors," *ICML*, 2017.
[23] Y. Qu, H. Qi, and C. Kwan, "Unsupervised sparse dirichlet-net for hyperspectral image super-resolution," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2511–2520, 2018.
[24] P. Kumaraswamy, "A generalized probability density function for double-bounded random processes," *Journal of Hydrology*, vol. 46, no. 1-2, pp. 79–88, 1980.
[25] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia, "Incorporating second-order functional knowledge for better option pricing," *Advances in neural information processing systems*, pp. 472–478, 2001.
[26] J. Han and C. Moraga, "The influence of the sigmoid function parameters on the speed of backpropagation learning," *From Natural to Artificial Neural Computation*, pp. 195–201, 1995.
[27] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.
[28] S. Huang and T. D. Tran, "Sparse signal recovery via generalized entropy functions minimization," *arXiv preprint arXiv:1703.10556*, 2017.
[29] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image and vision computing*, vol. 21, no. 11, pp. 977–1000, 2003.
[30] J. Woo, M. Stone, and J. L. Prince, "Multimodal registration via mutual information incorporating geometric and spatial context," *IEEE Transactions on Image Processing*, vol. 24, no. 2, pp. 757–769, 2015.
[31] I. Belghazi, S. Rajeswar, A. Baratin, R. D. Hjelm, and A. Courville, "Mine: mutual information neural estimation," *arXiv preprint arXiv:1801.04062*, 2018.
[32] M. D. Donsker and S. S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time. iv," *Communications on Pure and Applied Mathematics*, vol. 36, no. 2, pp. 183–212, 1983.
[33] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," *arXiv preprint arXiv:1808.06670*, 2018.
[34] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $\ell 2$, 1-norms minimization," *Advances in neural information processing systems*, pp. 1813–1821, 2010.

Ying Qu (IEEE Member since 2016) received the B.S. degree in automatics and M.S. degree in pattern recognition & artificial intelligence from Northeastern University, Shenyang, China in 2008 and 2010, respectively, and the Ph.D. degree in computer engineering from the University of Tennessee, Knoxville. She is currently working as a research associate in the Department of Electrical Engineering and Computer Science at the University of Tennessee, Knoxville. She was the recipient of the IEEE MIKIO Takagi Student Prize (Best Student Paper Awards) at the International Geoscience and Remote Sensing Symposium (IGARSS) in 2016. Her research interests include computer vision, remote sensing and artificial intelligence.

Zhenzhou Shao (IEEE Member since 2010) received the B.E. degree and M.E. degree in the Department of Information Engineering at Northeastern University, China, in 2007 and 2009, respectively, and the Ph.D. degree in mechanical engineering at the University of Tennessee, Knoxville, in 2013. He is currently the associate professor with the College of Information Engineering at Capital Normal University, China. His research interests include computer vision, machine learning and human-robot interaction.

Hairong Qi (IEEE Fellow since 2017) received the B.S. and M.S. degrees in computer science from Northern JiaoTong University, Beijing, China in 1992 and 1995, respectively, and the Ph.D. degree in computer engineering from North Carolina State University, Raleigh, in 1999. She is currently the Gonzalez Family Professor with the Department of Electrical Engineering and Computer Science at the University of Tennessee, Knoxville. Her current research interests are in advanced imaging and collaborative processing in resource-constrained distributed environment, hyperspectral image analysis, and automatic target recognition. Dr. Qi's research is supported by National Science Foundation (NSF), DARPA, Office of Naval Research (ONR), Department of Homeland Security (DHS), U.S. Army Space and Missile Defense Command, and U.S. Army Medical Research and Materiel Command. Dr. Qi is the recipient of the NSF CAREER Award. She also received the Best Paper Awards at the 18th International Conference on Pattern Recognition (ICPR) in 2006, the 3rd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC) in 2009, and IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensor (WHISPERS) in 2015. She is awarded the Highest Impact Paper from the IEEE Geoscience and Remote Sensing Society in 2012.