Understanding Generalized Whitening and Coloring Transform for Universal Style Transfer

Tai-Yin Chiu
University of Texas as Austin
chiu.taiyin@utexas.edu

Abstract

Style transfer is a task of rendering images in the styles of other images. In the past few years, neural style transfer has achieved a great success in this task, yet suffers from either the inability to generalize to unseen style images or fast style transfer. Recently, an universal style transfer technique that applies zero-phase component analysis (ZCA) for whitening and coloring image features realizes fast and arbitrary style transfer. However, using ZCA for style transfer is empirical and does not have any theoretical support. In addition, other whitening and coloring transforms (WCT) than ZCA have not been investigated. In this report, we generalize ZCA to the general form of WCT, provide an analytical performance analysis from the angle of neural style transfer, and show why ZCA is a good choice for style transfer among different WCTs and why some WCTs are not well applicable for style transfer.

1. Introduction

Style transfer is a task of synthesizing an image whose content comes from a target content image and style comes from another texture image. At the early stage, one successful method for style transfer is image quilting [4], which considers spatial maps of certain quantities (such as image intensity and local image orientation) over both the texture image and the content image. Another method [8] makes style transfer possible by making analogy between images and their artistically filtered version. Though these methods produce good results, they suffer from only using low level image features.

Later a seminal work of neural style transfer [5] leverages the power of convolutional neural network to extract features of an image that decouple and well represent visual styles and contents. Style transfer [6] is then achieved by jointly minimizing the feature loss [17] and the style loss formulated as the difference of Gram matrices. This optimization is solved by iteration [6, 12, 16, 18]. Thus, while having remarkable results, it suffers from computational inefficiency. To overcome this issue, a few methods [10, 13, 20] that use pre-computed neural networks to accelerate the style transfer were proposed. However, these methods are limited by only one transferrable style and cannot generalize to other unseen styles. StyleBank [1] addresses this limit by controlling the transferred style with style filters. Whenever a new style is needed, it can be learned into filters while holding the neural network fixed. Another method [3] proposes to train a conditional style transfer network that uses conditional instance normalization for multiple styles. Besides these two, more methods [2, 7, 21] to achieve arbitrary style transfer are proposed. However, they partly solve the problem and are still not able to generalize to every unseen style.

More recently, several methods exploring the second-order statistics of content image features and style image features came up for universal style transfer. AdaIN [9] tries to match up the variances of the stylized image feature and the style image feature. A method [14] that uses zero-phase component analysis (ZCA), a special kind of whitening and coloring transform (WCT), for feature transformation further focuses on covariances of image features. While AdaIN enjoys more computational efficiency and ZCA method synthesizes images visually closer to a considered style, Avatar-Net [19] aims to find a balance between them.

Even though the ZCA method for feature transformation produces good stylized images, it still remains an empirical method and lacks any theoretical analyses. Also, the report [14] does not mention the performance of other WCT methods. Here we generalize ZCA to the general form of WCT for style transfer. Furthermore, we incorporate the idea of neural style transfer to analytically discuss the performance of WCT on style transfer. From the analysis, we explain why ZCA is good for style transfer among different WCTs and show that not every WCT is well applicable to style transfer. In experiments, we study five natural WCT methods [11] associated with ZCA, principal component analysis (PCA), Cholesky decomposition, standardized ZCA,

and standardized PCA. The experiments show that PCA and standardized PCA give bad results while others lead to perceptually meaningful images, which is consistent with our theory.

2. Background

Since this report mainly inherits the ideas of neural style transfer and universal style transfer with ZCA, here we briefly introduce them in the following.

2.1. Neural style transfer

The paper by Gatys et al.[6] first introduced an effective method for style transfer using neural networks. The key finding of this work is that by processing images using convolutional neural network the style representation and the content representation of an image can be decoupled. This means that the style or the content of an image can be altered independently to generate another perceptually meaningful image.

The capability of separating content from style was observed from the VGG-19 network, a convolutional neural network targeting object recognition. VGG is composed of a series of convolutional layers followed by three fully connected layers. At the output of each convolutional layer is a feature map for the input image. For two images with similar contents, their feature maps extracted from a higher convolutional layer should be closer than the feature maps extracted from a lower convolutional layer, so that the final fully connected layer can classify them into the same cetegory based on their similar features from the highest convolutional layer. This implies that the feature map from a higher convolutional layer can be used as a content representation of an image. Suppose $\phi_j(I)$ is the feature map of an input image I at the j-th convolutional layer of the VGG-19 network. A feasible content representation of I could be extracted, for example, from the layers $j = relu4\_1$ or $relu4\_2$.

On the other hand, the computation of a style representation is an inspiration from the primary visual system where correlations between neurons are computed. Let $\phi_j(I)$ be the feature map of an image I at j-th convolutional layers of shape $h_j(I) \times w_j(I) \times k_j$, where $h_j(I)$, $w_j(I)$, and k_j are the height, width, and channel length of the feature map. For each layer, say j-th layer for example, we can define the Gram matrix $\mathbf{G}_j(I)$ of shape $k_j \times k_j$, where the (α, β) component is the correlation between the channels α and β and is given by

$$\mathbf{G}_j(I)_{\alpha,\beta} = \sum_{h=1}^{h_j(I)} \sum_{w=1}^{w_j(I)} \phi_j(I)_{h,w,\alpha} \phi_j(I)_{h,w,\beta}. \quad (1)$$

By reshaping $\phi_j(I)$ into a matrix $\mathbf{F}_j(I)$ of shape $k_j \times h_j(I)w_j(I)$, the Gram matrix can be written in a concise form $\mathbf{F}_j\mathbf{F}_j^{\mathrm{T}}$. In fact, such a Gram matrix can serve as a style representation for an image I. Furthermore, the Gram matrix from a higher convolutional layer captures a coarser style representation of I, while the one from a lower layer captures a finer style representation. For convenience, when we mention a feature map later, it refers to the reshaped feature matrix $\mathbf{F}$ instead of the original feature tensor ϕ.

To synthesize an image I_o with the content from the image I_c and the style from the image I_s, one has to find an optimal I_o to minimize the content loss and the style loss:

$$\begin{aligned} \arg\min_{I_o} & \frac{1}{n_l} ||\mathbf{F}_l(I_o) - \mathbf{F}_l(I_c)||_F^2 \\ & + \sum_{j \in \Omega} \lambda_j || \frac{1}{n_j} \mathbf{G}_j(I_o) - \frac{1}{m_j} \mathbf{G}_j(I_s) ||_F^2, \end{aligned} \quad (2)$$

where F indicates the Frobenius norm, l denotes some high l-th convolutional layer of VGG-19 network for evaluating content loss, Ω is a predefined set of convolutional layers for evaluating style loss, λ_j's are scaling factors, and $n_j = h_j(I_c)w_j(I_c)$ and $m_j = h_j(I_s)w_j(I_s)$.

2.2. Universal style transfer with ZCA

Unlike neural style transfer where learning from content images and style images is necessary to optimize Eq. 2, [14] proposes a learning-free scheme and formulates style transfer as an image reconstruction process. In particular, four autoencoders for general image reconstruction are built. Each encoder is a part of pre-trained VGG-19 network that encompasses the input layer to the $reluN\_1$ layer ($N = 1, 2, 3$ or 4) and is kept fixed during training process, while the corresponding decoder is structurally symmetrical to the encoder network. The autoencoder network is trained by minimizing the reconstruction loss $||I_r - I_i||_F^2$, where I_i and I_r are an input image and the reconstruction image. After training, each autoencoder can be used for single-level style transfer, and the four autoencoders can be cascaded to perform multi-level style transfer to achieve better synthetic images.

The inner-workings of an autoencoder for single-level style transfer is illustrated as follows. The encoder of the autoencoder is used to extract a feature map for an input image. For style transfer, a special kind of whitening and coloring transform (WCT) that uses zero-phase component analysis (ZCA) is applied for feature transformation. The transformed feature is then converted back to a perceptually meaningful image by the decoder. Specifically, given a content image I_c and a style image I_s, the extracted features by the encoder are $\mathbf{F}_c$ of shape $k \times n$ and $\mathbf{F}_s$ of shape $k \times m$, respectively. $\mathbf{F}_c$ and $\mathbf{F}_s$ are first subtracted by their means such that the centralized features $\bar{\mathbf{F}}_c$ and $\bar{\mathbf{F}}_s$ have a zero mean. Then eigen-decomposition is applied to their covariance matrices and derive $\frac{1}{n}\bar{\mathbf{F}}_c\bar{\mathbf{F}}_c^{\mathrm{T}} = \mathbf{E}_c\mathbf{\Lambda}_c\mathbf{E}_c^{\mathrm{T}}$ and

$\frac{1}{m}\bar{\mathbf{F}}_s\bar{\mathbf{F}}_s^{\mathrm{T}} = \mathbf{E}_s\mathbf{\Lambda}_s\mathbf{E}_s^{\mathrm{T}}$. The whitening step transforms $\bar{\mathbf{F}}_c$ to an uncorrelated feature $\tilde{\mathbf{F}}_c$ ($\frac{1}{n}\tilde{\mathbf{F}}_c\tilde{\mathbf{F}}_c^{\mathrm{T}} = \mathbf{I}$) according to Eq. 3:

$$\tilde{\mathbf{F}}_c = \mathbf{E}_c\mathbf{\Lambda}_c^{-\frac{1}{2}}\mathbf{E}_c^{\mathrm{T}}\bar{\mathbf{F}}_c. \tag{3}$$

The coloring step transforms $\tilde{\mathbf{F}}_c$ to $\bar{\mathbf{F}}_{zca}$ such that $\frac{1}{n}\bar{\mathbf{F}}_{zca}\bar{\mathbf{F}}_{zca}^{\mathrm{T}} = \frac{1}{m}\bar{\mathbf{F}}_s\bar{\mathbf{F}}_s^{\mathrm{T}}$ according to Eq. 4:

$$\bar{\mathbf{F}}_{zca} = \mathbf{E}_s\mathbf{\Lambda}_s^{\frac{1}{2}}\mathbf{E}_s^{\mathrm{T}}\tilde{\mathbf{F}}_c. \tag{4}$$

$\bar{\mathbf{F}}_{zca}$ is finally re-centered to $\mathbf{F}_{zca}$ by adding the mean of $\mathbf{F}_s$, which finishes the whole WCT. After feature transformation, the decoder converts the transformed feature $\mathbf{F}_{zca}$ to an image with the content from I_c and the style from I_s.

Moreover, better results can be achieved by multi-level style transfer: the autoencoder associated with the *relu4_1* layer takes I_c and I_s as inputs and produces a synthetic image I_4. Then I_4 as a content image and I_s are passed to the autoencoder associated with the *relu3_1* to generate a synthetic image I_3. Repeating this procedure until a synthetic image I_1 is generated from the autoencoder associated with the *relu1_1*. We will explain why multi-level style transfer works better later in Sec 3.4.

3. Methods

In [14], it uses ZCA, a special case of WCT, to realize style transfer. However, it mentions little about whether other WCT methods can work well for style transfer nor does it analytically discusses the performance of WCT. Here we consider the generalized WCT scheme for style transfer, providing a theory from the perspective of neural style transfer to explain why ZCA is a good way for this task and why some other WCTs might not.

3.1. Whitening transform

Suppose $x = (x_1, \ldots, x_d)^{\mathrm{T}}$ is a random vector to be whitened, and $\mu = \mathbb{E}[x]$ is the mass center of x. While x is not necessary to be centralized by subtracting μ from x for a whitening purpose, we follow the results of [14] where the whitening transform is done on centered signals. Therefore, in the following x is assumed to be centered with $\mu = 0$.

A whitening transform is a linear operation that brings a random vector x with covariance matrix $Cov(x) = \mathbf{\Sigma}$ to another random vector $z = (z_1, \ldots, z_d)^{\mathrm{T}}$ with an identity covariance matrix $Cov(z) = \mathbf{I}$. Specifically, the linear operation is defined by a $d \times d$ matrix $\mathbf{W}$ that converts x to $z = \mathbf{W}x$ that satisfies

$$\mathbb{E}[zz^{\mathrm{T}}] = \mathbf{W}\,\mathbb{E}[xx^{\mathrm{T}}]\mathbf{W}^{\mathrm{T}} = \mathbf{W}\mathbf{\Sigma}\mathbf{W}^{\mathrm{T}} = \mathbf{I}. \tag{5}$$

It follows that $\mathbf{W}\mathbf{\Sigma}\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{W}$ and thus

$$\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{\Sigma}^{-1}. \tag{6}$$

Accordingly,

$$\mathbf{W} = \mathbf{U}_1\mathbf{\Sigma}^{-\frac{1}{2}} = \mathbf{U}_1\mathbf{E}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{E}^{\mathrm{T}}, \tag{7}$$

where the eigen-decomposition of $\mathbf{\Sigma}$ is $\mathbf{\Sigma} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^{\mathrm{T}}$, and $\mathbf{U}_1$ is an orthogonal matrix. Different choices of $\mathbf{U}_1$ define different whitening transforms.

Besides, in certain situations it is more convenient to work with the standardized random vector $y = \mathbf{V}^{-\frac{1}{2}}x$ with $\mathbf{V}$ being the diagonal variance matrix $diag(\mathbf{\Sigma})$. Similar to Eq. 7, the whitening matrix $\mathbf{W}_y$ for y is written as $\mathbf{W}_y = \mathbf{U}_2\mathbf{P}^{-\frac{1}{2}}$, where $\mathbf{U}_2$ is an orthogonal matrix and $\mathbf{P}$ is the covariance matrix of y and also the correlation matrix of x, i.e., $\mathbf{P} = \mathbb{E}[yy^{\mathrm{T}}] = \mathbf{V}^{-\frac{1}{2}}\,\mathbb{E}[xx^{\mathrm{T}}]\mathbf{V}^{-\frac{1}{2}} = \mathbf{V}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{V}^{-\frac{1}{2}}$.

Since the whitened vector $\mathbf{W}_y y = \mathbf{U}_2\mathbf{P}^{-\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}}x$, we can alternatively express the whitening matrix $\mathbf{W}$ for x as

$$\mathbf{W} = \mathbf{U}_2\mathbf{P}^{-\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}}. \tag{8}$$

In this report we study five natural whitening transformations which are succinctly representable in the form defined in either Eq. 7 or Eq. 8. First recall [14] that uses ZCA whitening transformation for style transfer. ZCA whitening matrix $\mathbf{W}^{zca}$ is given by

$$\mathbf{W}^{zca} = \mathbf{\Sigma}^{-\frac{1}{2}} = \mathbf{E}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{E}^{\mathrm{T}}, \tag{9}$$

which corresponds to $\mathbf{U}_1 = \mathbf{I}$ in Eq. 7. Closely related to ZCA whitening, PCA whitening matrix $\mathbf{W}^{pca}$ is defined as

$$\mathbf{W}^{pca} = \mathbf{E}^{\mathrm{T}}\mathbf{\Sigma}^{-\frac{1}{2}} = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{E}^{\mathrm{T}}. \tag{10}$$

The major difference between $\mathbf{W}^{zca}$ and $\mathbf{W}^{pca}$ is that ZCA finally rotates back to the original coordinate by $\mathbf{E}$ after the rotation by $\mathbf{E}^{\mathrm{T}}$ followed by the scaling $\mathbf{\Lambda}^{-\frac{1}{2}}$.

If ZCA transform or PCA transform is applied to standardized vectors, we can have the standardized version of ZCA transform matrix $\mathbf{W}_{std}^{zca}$ or PCA transform matrix $\mathbf{W}_{std}^{pca}$:

$$\mathbf{W}_{std}^{zca} = \mathbf{P}^{-\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}} = \mathbf{E}_p\mathbf{\Lambda}_p^{-\frac{1}{2}}\mathbf{E}_p^{\mathrm{T}}\mathbf{V}^{-\frac{1}{2}}, \tag{11}$$

$$\mathbf{W}_{std}^{pca} = \mathbf{E}_p^{\mathrm{T}}\mathbf{P}^{-\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}} = \mathbf{\Lambda}_p^{-\frac{1}{2}}\mathbf{E}_p^{\mathrm{T}}\mathbf{V}^{-\frac{1}{2}}, \tag{12}$$

where the eigen-decomposition $\mathbf{P} = \mathbf{E}_p\mathbf{\Lambda}_p\mathbf{E}_p^{\mathrm{T}}$ is used. Note that $\mathbf{W}_{std}^{zca}$ and $\mathbf{W}_{std}^{pca}$ correspond to $\mathbf{U}_2 = \mathbf{I}$ and $\mathbf{U}_2 = \mathbf{E}_p^{\mathrm{T}}$ in Eq. 8, respectively.

The last natural whitening is Cholesky whitening, whose name comes from the Cholesky decomposition of $\mathbf{\Sigma}^{-1}$: $\mathbf{\Sigma}^{-1} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$. Comparing it to Eq. 6, we can derive the Cholesky whitening matrix $\mathbf{W}^{chol}$ to be

$$\mathbf{W}^{chol} = \mathbf{L}^{\mathrm{T}}. \tag{13}$$

Note that $\mathbf{W}^{chol}$ corresponds to $\mathbf{U}_1 = \mathbf{L}^{\mathrm{T}}\mathbf{\Sigma}^{\frac{1}{2}}$ in Eq. 7. In addition, it can be verified that the standardized version $\mathbf{W}_{std}^{chol}$ is also $\mathbf{L}^{\mathrm{T}}$.

3.2. Coloring transform

A coloring transform is a reversed procedure of the corresponding whitening transform. Specifically in single-level style transfer, a content feature $\mathbf{F}_c = [f_1^c, \ldots, f_n^c]_{k\times n}$ and a style feature $\mathbf{F}_s = [f_1^s, \ldots, f_m^s]_{k\times m}$ are extracted by an encoder composed of a part of VGG-19 network. $\mathbf{F}_c$ and $\mathbf{F}_s$ are then centralized to $\bar{\mathbf{F}}_c = [f_1^c - \bar{f}^c, \ldots, f_n^c - \bar{f}^c]$ and $\bar{\mathbf{F}}_s = [f_1^s - \bar{f}^s, \ldots, f_m^s - \bar{f}^s]$ by subtraction of their means $\bar{f}^c = \frac{1}{n}\sum_{i=1}^n f_i^c$ and $\bar{f}^s = \frac{1}{m}\sum_{i=1}^n f_i^s$, respectively. After the computation of the covariances $\mathbf{\Sigma}_c = \frac{1}{n}\bar{\mathbf{F}}_c\bar{\mathbf{F}}_c^{\mathrm{T}}$ and $\mathbf{\Sigma}_s = \frac{1}{m}\bar{\mathbf{F}}_s\bar{\mathbf{F}}_s^{\mathrm{T}}$, for each of them we can derive a whitening transform matrix $\mathbf{W}_c$ or $\mathbf{W}_s$ according to the methods introduced in Sec. 3.1. The whitened and colored feature $\mathbf{F}_{wct}$ to be decoded is then derived by

$$\mathbf{F}_{wct} = \mathbf{W}_s^{-1}\mathbf{W}_c\bar{\mathbf{F}}_c + [\bar{f}^s, \ldots, \bar{f}^s]_{k\times n}, \tag{14}$$

where $\mathbf{W}_c$ is the whitening transform, $\mathbf{W}_s^{-1}$ is the coloring transform, and the addition of the $k \times n$ matrix $[\bar{f}^s, \ldots, \bar{f}^s]$ is the re-centering step.

To comply with the formulation in Sec. 3.1, we define f^c as the random vector representing the n examples f_i^c, $i = 1, \ldots, n$, and f^s as the random vector representing the m examples f_i^s, $i = 1, \ldots, m$. Whitening matrices $\mathbf{W}_c$ and $\mathbf{W}_s$ then can be derived from the covariances $\mathbf{\Sigma}_c = \mathbb{E}[(f^c - \bar{f}^c)(f^c - \bar{f}^c)^{\mathrm{T}}]$ and $\mathbf{\Sigma}_s = \mathbb{E}[(f^s - \bar{f}^s)(f^s - \bar{f}^s)^{\mathrm{T}}]$, where $\bar{f}^c = \mathbb{E}[f^c]$ and $\bar{f}^s = \mathbb{E}[f^s]$. Therefore, the random vector f^{wct} that represents the columns of $\mathbf{F}_{wct}$ is given by

$$f^{wct} = \mathbf{W}_s^{-1}\mathbf{W}_c(f^c - \bar{f}^c) + \bar{f}^s. \tag{15}$$

Equipped with Eq. 15, we can analyze the performance of WCTs for style transfer.

3.3. Analysis of WCTs for single-level style transfer

From the point of view of neural style transfer, a single-level style transfer in [14] can actually be regarded as a way that uses ZCA to provide a fast and approximate solution to $\min_{\mathbf{F}} l(\mathbf{F}; \mathbf{F}_c, \mathbf{F}_s)$ with $l(\mathbf{F}; \mathbf{F}_c, \mathbf{F}_s)$ defined as

$$l(\mathbf{F}; \mathbf{F}_c, \mathbf{F}_s) = \underbrace{\frac{1}{n}||\mathbf{F} - \mathbf{F}_c||_F^2}_{\textit{content loss}} + \lambda \underbrace{||\frac{1}{n}\mathbf{F}\mathbf{F}^{\mathrm{T}} - \frac{1}{m}\mathbf{F}_s\mathbf{F}_s^{\mathrm{T}}||_F^2}_{\textit{style loss}}, \tag{16}$$

where $\mathbf{F}_c$ of shape $k \times n$ and $\mathbf{F}_s$ of shape $k \times m$ are extracted at the output of either *relu4_1*, *relu3_1*, *relu2_1*, or *relu1_1* layer of VGG network, and k is the number of channels at that layer. Furthermore, it can be proved that if a WCT is used to approximate the solution, the loss function $l(\mathbf{F}; \mathbf{F}_c, \mathbf{F}_s)$ of the approximated solution $\mathbf{F}_{wct}$ is bounded.

Theorem 3.1. *Given a single-level style transfer that is formulated as the minimization of $l(\mathbf{F}; \mathbf{F}_c, \mathbf{F}_s)$. If $\mathbf{F} = \mathbf{F}_{wct}$ which is computed according to Eq. 14, then the style loss is zero and the content loss is bounded by the means and covariances of $\mathbf{F}_c$ and $\mathbf{F}_s$.*

Proof. First recall that a whitening matrix should satisfy Eq. 6. Therefore, we have $\mathbf{\Sigma} = (\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1} = \mathbf{W}^{-1}(\mathbf{W}^{\mathrm{T}})^{-1}$. Thus

$$\mathbf{\Sigma}_c = \mathbb{E}[(f^c - \bar{f}^c)(f^c - \bar{f}^c)^{\mathrm{T}}] = \mathbf{W}_c^{-1}(\mathbf{W}_c^{\mathrm{T}})^{-1}, \tag{17a}$$

$$\mathbf{\Sigma}_s = \mathbb{E}[(f^s - \bar{f}^s)(f^s - \bar{f}^s)^{\mathrm{T}}] = \mathbf{W}_s^{-1}(\mathbf{W}_s^{\mathrm{T}})^{-1}. \tag{17b}$$

Let's begin with the second term of $l(\mathbf{F}_{wct})$:

$$\begin{aligned}&\frac{1}{n}\mathbf{F}_{wct}\mathbf{F}_{wct}^{\mathrm{T}} - \frac{1}{m}\mathbf{F}_s\mathbf{F}_s^{\mathrm{T}}\\ =&\frac{1}{n}\sum_{i=1}^n f_i^{wct}(f_i^{wct})^{\mathrm{T}} - \frac{1}{m}\sum_{i=1}^m f_i^s(f_i^s)^{\mathrm{T}}\\ =&\mathbb{E}[f^{wct}(f^{wct})^{\mathrm{T}}] - \mathbb{E}[f^s(f^s)^{\mathrm{T}}],\end{aligned} \tag{18}$$

where $\mathbb{E}[f^{wct}(f^{wct})^{\mathrm{T}}]$ equals to

$$\begin{aligned}&\mathbf{W}_s^{-1}\mathbf{W}_c\,\mathbb{E}[(f^c - \bar{f}^c)(f^c - \bar{f}^c)^{\mathrm{T}}]\mathbf{W}_c^{\mathrm{T}}(\mathbf{W}_s^{-1})^{\mathrm{T}}\\ &+\mathbf{W}_s^{-1}\mathbf{W}_c(\mathbb{E}[f^c] - \bar{f}^c)(\bar{f}^s)^{\mathrm{T}}\\ &+\bar{f}^s(\mathbb{E}[f^c] - \bar{f}^c)^{\mathrm{T}}\mathbf{W}_c^{\mathrm{T}}(\mathbf{W}_s^{-1})^{\mathrm{T}} + \bar{f}^s(\bar{f}^s)^{\mathrm{T}},\end{aligned} \tag{19}$$

which can be reduced to

$$\mathbf{W}_s^{-1}(\mathbf{W}_s^{-1})^{\mathrm{T}} + 0 + 0 + \bar{f}^s(\bar{f}^s)^{\mathrm{T}} = \mathbb{E}[f^s(f^s)^{\mathrm{T}}], \tag{20}$$

where the identities $(\mathbf{W}_s^{-1})^{\mathrm{T}} = (\mathbf{W}_s^{\mathrm{T}})^{-1}$ and $\mathbb{E}[(f^s - \bar{f}^s)(f^s - \bar{f}^s)^{\mathrm{T}}] = \mathbb{E}[f^s(f^s)^{\mathrm{T}}] - \bar{f}^s(\bar{f}^s)^{\mathrm{T}}$ are used. Hence the second term of $l(\mathbf{F}_{wct})$ is $\mathbb{E}[f^s(f^s)^{\mathrm{T}}] - \mathbb{E}[f^s(f^s)^{\mathrm{T}}] = 0$. This proves that the style loss is zero if $\mathbf{F}_{wct}$ is used.

Next we focus on the first term:

$$\begin{aligned}&\frac{1}{n}||\mathbf{F}_{wct} - \mathbf{F}_c||_F^2 = \mathrm{tr}[\frac{1}{n}(\mathbf{F}_{wct} - \mathbf{F}_c)(\mathbf{F}_{wct} - \mathbf{F}_c)^{\mathrm{T}}]\\ =&\mathrm{tr}[\frac{1}{n}\sum_{i=1}^n (f_i^{wct} - f_i^c)(f_i^{wct} - f_i^c)^{\mathrm{T}}]\\ =&\mathrm{tr}[\mathbb{E}[(f^{wct} - f^c)(f^{wct} - f^c)^{\mathrm{T}}]]\end{aligned} \tag{21}$$

With $f^{c-\bar{c}} \triangleq f^c - \bar{f}^c$ and $f^{c-\bar{s}} \triangleq f^c - \bar{f}^s$, Eq. 21 can be expanded as

$$\begin{aligned}&\mathrm{tr}[\mathbf{W}_s^{-1}\mathbf{W}_c\,\mathbb{E}[f^{c-\bar{c}}f^{c-\bar{c}\mathrm{T}}]\mathbf{W}_c^{\mathrm{T}}(\mathbf{W}_s^{-1})^{\mathrm{T}}]\\ &-\mathrm{tr}[\mathbf{W}_s^{-1}\mathbf{W}_c\,\mathbb{E}[f^{c-\bar{c}}f^{c-\bar{s}\mathrm{T}}]]\\ &-\mathrm{tr}[\mathbb{E}[f^{c-\bar{s}}f^{c-\bar{c}\mathrm{T}}]\mathbf{W}_c^{\mathrm{T}}(\mathbf{W}_s^{-1})^{\mathrm{T}}]\\ &+\mathrm{tr}[\mathbb{E}[f^{c-\bar{s}}f^{c-\bar{s}\mathrm{T}}]].\end{aligned} \tag{22}$$

The first trace is equal to $\mathrm{tr}[\mathbf{W}_s^{-1}(\mathbf{W}_s^{-1})^{\mathrm{T}}] = \mathrm{tr}[\mathbf{\Sigma}_s]$. The second and third trace terms are equivalent, since $\mathrm{tr}[\mathbf{A}] = \mathrm{tr}[\mathbf{A}^{\mathrm{T}}]$ for any arbitrary matrix $\mathbf{A}$. Since $\mathbb{E}[f^{c-\bar{c}}f^{c-\bar{s}\mathrm{T}}]$ equals to

$$\begin{aligned}&\mathbb{E}[f^c f^{c\mathrm{T}}] - \mathbb{E}[f^c]\bar{f}^{s\mathrm{T}} - \bar{f}^c\,\mathbb{E}[f^c]^{\mathrm{T}} + \bar{f}^c\bar{f}^{s\mathrm{T}}\\ =&\mathbb{E}[f^c f^{c\mathrm{T}}] - \bar{f}^c\bar{f}^{c\mathrm{T}} = \mathbf{\Sigma}_c = \mathbf{W}_c^{-1}(\mathbf{W}_c^{-1})^{\mathrm{T}},\end{aligned} \tag{23}$$

the second and third trace terms become $\mathrm{tr}[\mathbf{W}_s^{-1}(\mathbf{W}_c^{-1})^{\mathrm{T}}]$. For the fourth trace, we observe that $\mathbb{E}[f^{c-\bar{s}}f^{c-\bar{s}\mathrm{T}}]$ can be further written as

$$\begin{aligned}&\mathbb{E}[f^c f^{c\mathrm{T}}] - \mathbb{E}[f^c]\bar{f}^{s\mathrm{T}} - \bar{f}^s\,\mathbb{E}[f^c]^{\mathrm{T}} + \bar{f}^s\bar{f}^{s\mathrm{T}}\\ =&\mathbf{W}_c^{-1}(\mathbf{W}_c^{-1})^{\mathrm{T}} + \bar{f}^c\bar{f}^{c\mathrm{T}} - \bar{f}^c\bar{f}^{s\mathrm{T}} - \bar{f}^s\bar{f}^{c\mathrm{T}} + \bar{f}^s\bar{f}^{s\mathrm{T}}\\ =&\mathbf{W}_c^{-1}(\mathbf{W}_c^{-1})^{\mathrm{T}} + (\bar{f}^c - \bar{f}^s)(\bar{f}^c - \bar{f}^s)^{\mathrm{T}},\end{aligned} \tag{24}$$

taking trace of which, the fourth trace term turns to be $\mathrm{tr}[\mathbf{W}_c^{-1}(\mathbf{W}_c^{-1})^{\mathrm{T}}] + ||\bar{f}^c - \bar{f}^s||_2^2$. Putting everything above together, we have the loss $l(\mathbf{F}_{wct}; \mathbf{F}_c, \mathbf{F}_s)$ to be

$$\begin{aligned}&l(\mathbf{F}_{wct}; \mathbf{F}_c, \mathbf{F}_s) = ||\bar{f}^c - \bar{f}^s||_2^2 +\\ &\mathrm{tr}[\mathbf{W}_s^{-1}(\mathbf{W}_s^{-1})^{\mathrm{T}} - 2\mathbf{W}_s^{-1}(\mathbf{W}_c^{-1})^{\mathrm{T}} + \mathbf{W}_c^{-1}(\mathbf{W}_c^{-1})^{\mathrm{T}}].\end{aligned} \tag{25}$$

Moreover, by exploiting Cauchy-Schwarz inequality

$$\mathrm{tr}[\mathbf{W}_s^{-1}(\mathbf{W}_c^{-1})^{\mathrm{T}}] \geq -\sqrt{\mathrm{tr}[\mathbf{W}_s^{-1}\mathbf{W}_s^{-1\mathrm{T}}]}\sqrt{\mathrm{tr}[\mathbf{W}_c^{-1}\mathbf{W}_c^{-1\mathrm{T}}]} \tag{26}$$

and the identities $\mathbf{\Sigma}_c = \mathbf{W}_c^{-1}\mathbf{W}_c^{-1\mathrm{T}}$ and $\mathbf{\Sigma}_s = \mathbf{W}_s^{-1}\mathbf{W}_s^{-1\mathrm{T}}$, we can derive the inequality

$$l(\mathbf{F}_{wct}) \leq ||\bar{f}^c - \bar{f}^s||_2^2 + (\sqrt{\mathrm{tr}(\mathbf{\Sigma}_s)} + \sqrt{\mathrm{tr}(\mathbf{\Sigma}_c)})^2, \tag{27}$$

where the upper bound is associated with the means and covariances of $\mathbf{F}_c$ and $\mathbf{F}_s$. ■

Theorem 3.1 says that whichever WCT is used for single-level style transfer, there is always no style loss and the upper bound for content loss only depends on the content feature and the style feature but is irrespective of the WCT used. The bounded content loss implies that WCT can capture the general appearance of a content image. However, whether the details of a content can hold still depends on the WCT used. To evaluate the performance of different WCTs on style transfer, we have to come up with a simple and yet indicative quantitative metric. To this end, following Theorem. 3.1 we have a corollary as follows.

Corollary 3.1.1. $\mathrm{tr}[\mathbf{W}_s^{-1}(\mathbf{W}_c^{-1})^{\mathrm{T}}]$ *can be used as a score function to evaluate style transfer results using WCT: the higher value of* $\mathrm{tr}[\mathbf{W}_s^{-1}(\mathbf{W}_c^{-1})^{\mathrm{T}}]$*, the better the performance of the WCT used for style transfer.*

Proof. Recall Eq. 25 where $\bar{f}^c$, $\bar{f}^s$, $\mathbf{\Sigma}_c = \mathbf{W}_c^{-1}\mathbf{W}_c^{-1\mathrm{T}}$, and $\mathbf{\Sigma}_s = \mathbf{W}_s^{-1}\mathbf{W}_s^{-1\mathrm{T}}$ all depend only on the content and the style images and are irrelevant to the WCT used. Therefore, a WCT that minimizes the loss function $l(\mathbf{F_{wct}})$ more corresponds to a higher value of $\mathrm{tr}[\mathbf{W}_s^{-1}(\mathbf{W}_c^{-1})^{\mathrm{T}}]$. ■

With Corollary 3.1.1, we can have a sense for why ZCA is a good choice for style transfer. If ZCA (refer to Eq. 9) is used, the score is

$$\begin{aligned}&\mathrm{tr}[\mathbf{\Sigma}_s^{\frac{1}{2}}\mathbf{\Sigma}_c^{\frac{1}{2}}] = \mathrm{tr}[\mathbf{E}_s\mathbf{\Lambda}_s^{\frac{1}{2}}\mathbf{E}_s^{\mathrm{T}}\mathbf{E}_c\mathbf{\Lambda}_c^{\frac{1}{2}}\mathbf{E}_c^{\mathrm{T}}]\\ =&\mathrm{tr}[\sum_{i=1}^n \sigma_i^s e_i^s (e_i^s)^{\mathrm{T}} \sum_{j=1}^n \sigma_j^c e_j^c (e_j^c)^{\mathrm{T}}]\\ =&\sum_{i,j} \sigma_i^s\sigma_j^c (e_i^s)^{\mathrm{T}} e_j^c \times \mathrm{tr}[e_i^s (e_j^c)^{\mathrm{T}}]\\ =&\sum_{i,j} \sigma_i^s\sigma_j^c [(e_i^s)^{\mathrm{T}} e_j^c]^2,\end{aligned} \tag{28}$$

where σ_i^s's (σ_j^c's) and e_i^s's (e_j^c's) are the singular values and the corresponding eigenvectors of $\mathbf{\Sigma}_s^{\frac{1}{2}}$ ($\mathbf{\Sigma}_c^{\frac{1}{2}}$), respectively.

On the other hand, since generally $\mathbf{W}_s$ and $\mathbf{W}_c$ can be written as $\mathbf{W}_s = \mathbf{U}_s\mathbf{\Sigma}_s^{-\frac{1}{2}}$ and $\mathbf{W}_c = \mathbf{U}_c\mathbf{\Sigma}_c^{-\frac{1}{2}}$ with certain orthogonal matrices $\mathbf{U}_s$ and $\mathbf{U}_c$ (refer to Eq. 7), $\mathrm{tr}[\mathbf{W}_s^{-1}(\mathbf{W}_c^{-1})^{\mathrm{T}}]$ equals to $\mathrm{tr}[\mathbf{\Sigma}_s^{\frac{1}{2}}\mathbf{U}_s^{\mathrm{T}}\mathbf{U}_c\mathbf{\Sigma}_c^{\frac{1}{2}}]$. According to von Neumann's trace inequality, it can be bounded as follows:

$$|\mathrm{tr}[\mathbf{\Sigma}_s^{\frac{1}{2}}\mathbf{U}_s^{\mathrm{T}}\mathbf{U}_c\mathbf{\Sigma}_c^{\frac{1}{2}}]| \leq \sum_{i=1}^n \sigma_i^s\sigma_i^c = \sum_{i,j} \sigma_i^s\sigma_i^c[(e_i^s)^{\mathrm{T}}e_j^c]^2, \tag{29}$$

where we use the identity $\sum_j[(e_i^s)^{\mathrm{T}}e_j^c]^2 = ||e_i^s||_2^2 = 1, \forall i$. By replacing the factor σ_i^c of $\sum_{i,j}\sigma_i^s\sigma_i^c[(e_i^s)^{\mathrm{T}}e_j^c]^2$ with σ_j^c, it becomes the score of ZCA in Eq. 28. This implies that the score of ZCA is a good approximation of the upper bound and thus ZCA is a good choice for style transfer.

In contrast, when bad choices of $\mathbf{U}_s$ and $\mathbf{U}_c$ are used, it could result in negative scores and bad style transfer results, and PCA is one of such cases. If PCA (refer to Eq. 10) is used, the score is

$$\begin{aligned}&\mathrm{tr}[\mathbf{E}_s\mathbf{\Lambda}_s^{\frac{1}{2}}\mathbf{\Lambda}_c^{\frac{1}{2}}\mathbf{E}_c^{\mathrm{T}}]\\ =&\sum_{i=1}^n \sigma_i^s\sigma_i^c\mathrm{tr}[e_i^s(e_i^c)^{\mathrm{T}}] = \sum_{i=1}^n \sigma_i^s\sigma_i^c(e_i^s)^{\mathrm{T}}e_i^c.\end{aligned} \tag{30}$$

Eq. 30 could be a small or a negative value since $(e_i^s)^{\mathrm{T}}e_i^c$ could be negative. This implies that PCA is not a good option for style transfer.

Moreover, if the standardized version of ZCA (refer to Eq. 11) is used, the score is $\mathrm{tr}[\mathbf{V}_s^{\frac{1}{2}}\mathbf{P}_s^{\frac{1}{2}}\mathbf{P}_c^{\frac{1}{2}}\mathbf{V}_c^{\frac{1}{2}}]$, where $\mathbf{V}_s$ ($\mathbf{V}_c$) and $\mathbf{P}_s$ ($\mathbf{P}_c$) are the diagonal variance matrix and the correlation matrix of $\mathbf{F}_s$ ($\mathbf{F}_c$), respectively. Since a covariance matrix $\mathbf{\Sigma}$ connects to the corresponding correlation matrix $\mathbf{P}$ in a relation $\mathbf{\Sigma} = \mathbf{V}^{\frac{1}{2}}\mathbf{P}\mathbf{V}^{\frac{1}{2}}$, $\mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{V}^{\frac{1}{2}}\mathbf{P}^{\frac{1}{2}}\mathbf{U}$ with $\mathbf{U}$ being some orthogonal matrix. It can be shown that $\mathbf{U}$ is very close to the identity matrix, and thus we have $\mathbf{\Sigma}^{\frac{1}{2}} \approx \mathbf{V}^{\frac{1}{2}}\mathbf{P}^{\frac{1}{2}}$. This implies that the score $\mathrm{tr}[\mathbf{V}_s^{\frac{1}{2}}\mathbf{P}_s^{\frac{1}{2}}\mathbf{P}_c^{\frac{1}{2}}\mathbf{V}_c^{\frac{1}{2}}]$ is close to $\mathrm{tr}[\mathbf{\Sigma}_s^{\frac{1}{2}}\mathbf{\Sigma}_c^{\frac{1}{2}}]$, which is the score of ZCA. Hence, the

performance of standardized ZCA and ZCA on style transfer is similar. Besides, similar to the case of PCA, the standardized PCA could lead to a negative score and is not good for style transfer.

3.4. Multi-level style transfer

[14] empirically shows that multiple autoencoders for single-level style transfer can be cascaded to achieve better style transfer results. Here we propose an explanation from the perspective of neural style transfer: suppose we want to find a synthetic image I that minimizes the loss function below:

$$\frac{1}{n_4}||\mathbf{F}_4-\mathbf{F}_{4,c}||_F^2+\sum_{N=1}^{4}\lambda_N||\frac{1}{n_N}\mathbf{F}_N\mathbf{F}_N^{\mathrm{T}}-\frac{1}{m_N}\mathbf{F}_{N,s}\mathbf{F}_{N,s}^{\mathrm{T}}||_F^2, \tag{31}$$

where $\mathbf{F}_N = \mathbf{F}_N(I)$ is the feature of I extracted at the *reluN_1* layer of VGG-19 network, $\mathbf{F}_{N,c}$'s and $\mathbf{F}_{N,s}$'s are the features of the content image and the style image at different layers, respectively. We approach this optimization problem by first solving I_4 from the *relu4_1* part:

$$\frac{||\mathbf{F}_4(I_4)-\mathbf{F}_{4,c}||_F^2}{n_4}+\lambda_4||\frac{\mathbf{F}_4(I_4)\mathbf{F}_4(I_4)^{\mathrm{T}}}{n_4}-\frac{\mathbf{F}_{4,s}\mathbf{F}_{4,s}^{\mathrm{T}}}{m_4}||_F^2, \tag{32}$$

which is the single-level style transfer associated with the *relu4_1* layer and the solution is approximated using WCT.

Next we want to find another image I_3 on top of I_4 such that I_3 is close to I_4 and also optimizes the loss $\lambda_3||\frac{1}{n_3}\mathbf{F}_3\mathbf{F}_3^{\mathrm{T}}-\frac{1}{m_3}\mathbf{F}_{3,s}\mathbf{F}_{3,s}^{\mathrm{T}}||_F^2$. If I_3 is close to I_4, then I_3 is still suboptimal to Eq. 32. To account for this, instead of having a direct loss $||I_4-I_3||_F^2$, we require the features of I_4 and I_3 extracted at the *relu3_1* layer to be close. Overall, I_3 optimizes the following loss:

$$\frac{||\mathbf{F}_3(I_3)-\mathbf{F}_3(I_4)||_F^2}{n_3}+\lambda_3||\frac{\mathbf{F}_3(I_3)\mathbf{F}_3(I_3)^{\mathrm{T}}}{n_3}-\frac{\mathbf{F}_{3,s}\mathbf{F}_{3,s}^{\mathrm{T}}}{m_3}||_F^2, \tag{33}$$

which exactly corresponds to single-level style transfer associated with the *relu3_1* layer that takes I_4 as the content image and whose solution can be approximated by WCT.

We can repeat the procedure for I_2 on top of I_3 and I_1 on top of I_2, and I_1 will be a suboptimal solution to Eq. 31. Effectively, we are approximating a solution to Eq. 31 by cascading four single-level style transfer autoencoders where each autoencoder takes the output image of the previous autoencoder as a content image .

This standpoint also explains why synthetic results are worse if the autoencoders are cascaded in a reverse order: first I_1 is generated from I_c and I_s using the autoencoder associated with the *relu1_1* layer, then I_1 and I_s are fed into the autoencoder associated with the *relu2_1* layer to generate I_2, and repeat the procedure until I_4 is generated by the autoencoder associated with the *relu4_1* layer. Compared to the original order where the content feature comes from the *relu4_1* layer (the $\frac{1}{n_4}||\mathbf{F}_4-\mathbf{F}_{4,c}||_F^2$ term in Eq. 31), the reverse order actually finds a suboptimal solution to the loss below:

$$\frac{1}{n_1}||\mathbf{F}_1-\mathbf{F}_{1,c}||_F^2+\sum_{N=1}^{4}\lambda_N||\frac{1}{n_N}\mathbf{F}_N\mathbf{F}_N^{\mathrm{T}}-\frac{1}{m_N}\mathbf{F}_{N,s}\mathbf{F}_{N,s}^{\mathrm{T}}||_F^2, \tag{34}$$

where the content information comes from *relu1_1* layer. As mentioned in Sec. 2.1, since a feature map from a higher convolutional layer better represents the content information of an image, the original order of cascade gives better synthetic results.

4. Experiments

4.1. Training details

We train the autoencoders on the MS-COCO dataset. MS-COCO dataset consists of 11.8K training images. Each image from the dataset is resized to 512×512 and randomly cropped to 256×256 as an input in a batch. For *relu4_1* and *relu3_1* cases, the autoencoders are trained with a batch size of 16 for 10 epochs, while for *relu2_1* and *relu1_1* cases, the autoencoders are trained for 5 epochs. We use Adam optimizer with learning rate 1×10^{-4} and without weight decay.

In the report [14], a decoder is structurally symmetric to the corresponding encoder in a way that a max-pooling layer in the encoder corresponds to an up-sampling layer in the decoder. However, a max-pooling operation in the encoder loses spatial information in feature maps, and the corresponding up-sampling operation in the decoder cannot recover the lost structures very well. Therefore, the decoded image will contain structural artifacts and distortion at boundaries [15]. To fix this, we use transposed convolutional layers as the symmetric part of max-pooling layers. Compared to up-sampling layers, transposed convolutional layers have adjustable parameters to flexibly learn to reconstruct images and avoid distortions.

4.2. Discussions

In Fig. 1 we demonstrate eight examples of style transfer based on five natural WCTs, which are associated with ZCA, PCA, standardized version of ZCA and PCA, and Cholesky decomposition, respectively. In the Table. 4.1 are the corresponding values of the score $\mathrm{tr}(\mathbf{W}_s^{-1}(\mathbf{W}_c^{-1})^{\mathrm{T}})$ at different single levels of style transfer. Note that comparing the values in different examples are meaningless, since a value depends on many factors, such as the sizes of the content image and the style image and the distribution of pixel values.

Let us focus on the results from PCA and PCA-std first. If we look at them closely, we can notice that there are

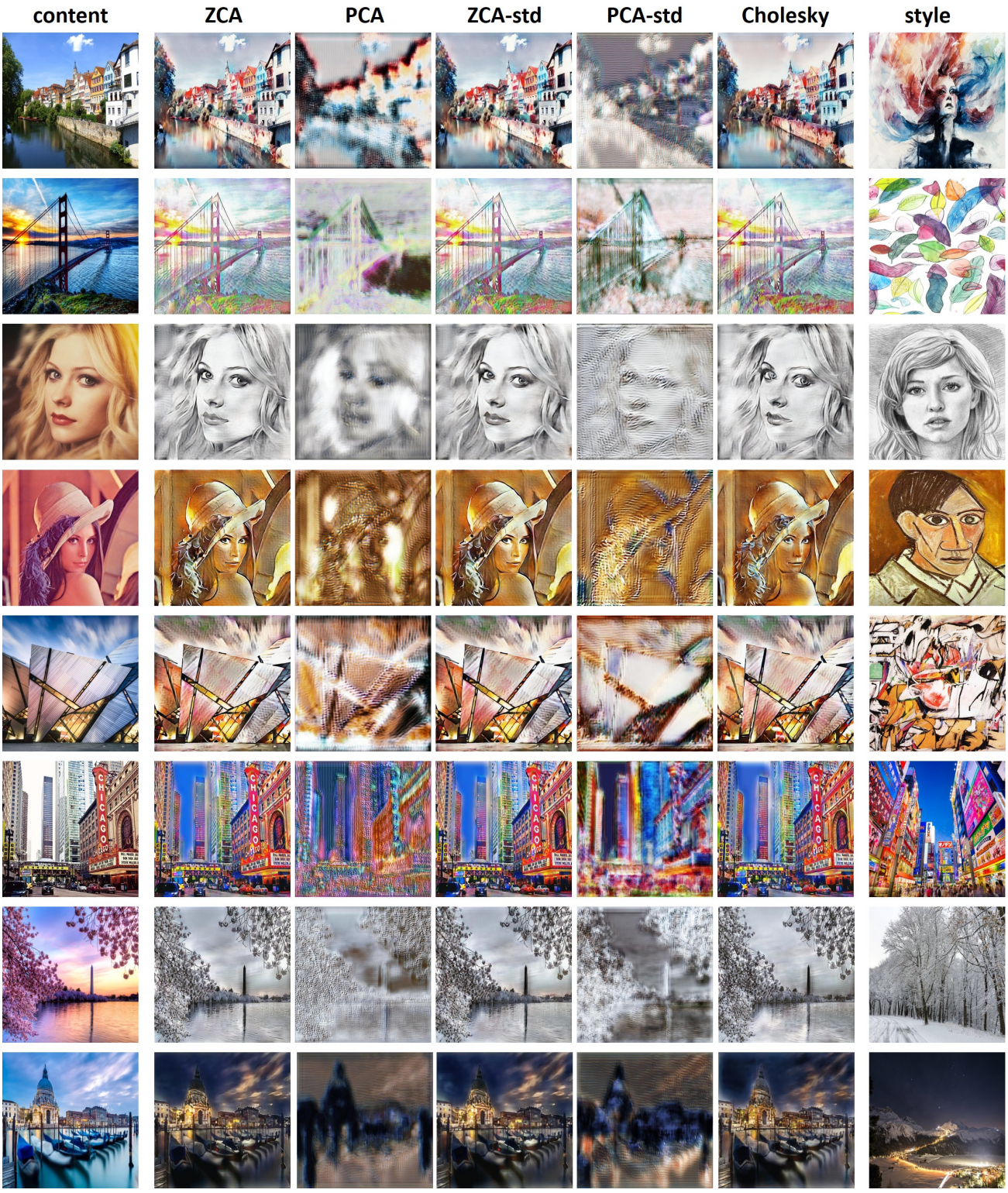

Figure 1. Examples of style transfer using five natural WCTs. The results from PCA and PCA-std are not good but still capture the general appearance of content images and the styles from style images, which can be explained by the bounded content loss and zero style loss in single-level style transfer. In contrast, the results from ZCA, ZCA-std, and Cholesky are much better, since the values of $\mathrm{tr}(\mathbf{W}_s^{-1}(\mathbf{W}_c^{-1})^{\mathrm{T}})$ for them are much higher as shown in Table 4.1.

Example	ZCA	PCA	ZCA-std	PCA-std	Cholesky
1	70177.1 / 2244.9	-8200.9 / -28.7	69862.4 / 2246.2	-7389.0 / -98.9	62073.5 / 2063.2
	464.0 / 8.5	-15.3 / 5.2	466.4 / 8.4	-22.6 / -3.7	442.8 / 8.2
2	83055.0 / 3349.3	359.5 / 140.3	83012.9 / 3358.8	5748.9 / 255.5	71536.8 / 2940.6
	688.4 / 9.6	-3.4 / -3.3	691.5 / 9.7	6.1 / -0.9	642.1 / 8.8
3	38127.9 / 1861.5	2267.4 / 144.3	37572.1 / 1852.1	6400.9 / -228.4	32533.6 / 1695.7
	404.6 / 3.9	79.8 / 2.7	402.9 / 3.8	-48.1 / -0.5	375.4 / 3.5
4	60312.3 / 2513.3	2079.1 / -20.2	60144.6 / 2533.0	6482.2 / -549.6	52955.3 / 2269.4
	616.0 / 7.6	43.9 / -0.3	620.8 / 7.7	0.8 / 2.9	587.6 / 6.9
5	92117.1 / 4942.3	-9806.4 / -386.3	91384.9 / 4969.0	4170.3 / 624.5	79736.5 / 4623.4
	1146.7 / 13.4	27.1 / 3.5	1151.0 / 13.5	39.0 / 1.5	1115.4 / 12.7
6	147059.6 / 5585.0	-15089.9 / 33.0	146909.1 / 5592.4	-6802.0 / -88.6	135144.2 / 5232.4
	1175.9 / 16.4	79.8 / -6.9	1183.1 / 16.5	162.9 / 7.0	1138.3 / 15.6
7	59170.7 / 2247.0	7361.8 / -8.3	59147.1 / 2264.9	12447.2 / -276.2	53560.3 / 2096.3
	647.1 / 6.3	-34.7 / -1.3	650.4 / 6.3	98.5 / -4.3	620.0 / 5.9
8	44762.8 / 1174.6	-6245.8 / -76.0	44515.5 / 1173.1	-10660.4 / -104.7	39738.1 / 1067.6
	255.8 / 4.5	-31.8 / 2.	255.3 / 4.5	15.1 / 2.5	241.2 / 4.3

Table 1. Values of $\mathrm{tr}(\mathbf{W}_s^{-1}(\mathbf{W}_c^{-1})^{\mathrm{T}})$ for the eight examples in Fig. 1. Each cell contains four values separated by slashes, and they correspond to the values from the single levels associated with the $reluN\_1$ layer, $N = 4, 3, 2, 1$, respectively.

some points that are consistent with our previous analysis. Apparently, the results from PCA and PCA-std are not good, but we can observe that the styles from the style images are transferred to the synthetic images to a certain extent. This is because from the perspective of neural style transfer, WCT does not cause any style information loss in single-level style transfer as shown in the proof for the Theorem 3.1. Moreover, we can observe that the synthetic images by PCA and PCA-std still capture the general appearance of the content images. This can be explained by the bounded content loss as proved in the Theorem 3.1. However, as explained in Eq. 30, since style transfer by PCA or PCA-std results in worse content losses, the details of content information are ruined in the synthetic images. This is justified in Table. 4.1, where in each example the values of $\mathrm{tr}(\mathbf{W}_s^{-1}(\mathbf{W}_c^{-1})^{\mathrm{T}})$ for PCA and PCA-std are much smaller than the values for other WCT methods and could even be negative.

Furthermore, Table. 4.1 shows that in each example the values of $\mathrm{tr}(\mathbf{W}_s^{-1}(\mathbf{W}_c^{-1})^{\mathrm{T}})$ for ZCA and ZCA-std are very close, which is consistent with the previous discussion. That is why the synthetic images by ZCA and ZCA-std look almost the same. Besides, we can notice that in each example the values of $\mathrm{tr}(\mathbf{W}_s^{-1}(\mathbf{W}_c^{-1})^{\mathrm{T}})$ for Cholesky decomposition method are slightly smaller than the values for ZCA and ZCA-std, implying that the results by Cholesky method could have been slightly worse than the ones by ZCA and ZCA-std. However, the visualization in Fig. 1 shows only little difference between them; one small but noticeable difference is in the third example of pencil sketch, where the left part in the Cholesky result is brighter.

In summary, even though there are many choices of WCT, such as standardized ZCA or Cholesky decomposition, for style transfer, from the angle of neural style transfer ZCA can always be the first good choice. In contrast, PCA, one of the most widely used whitening method, is not well applicable in style transfer.

5. Conclusion

In this report we study the general form of WCT for style transfer. We analyze the performance of WCT from the perspective of neural style transfer. From the analysis we show why some WCTs, especially ZCA, are good choices for style transfer among different WCTs and why some other WCTs might not be well applicable to style transfer. In experiments we study five natural WCTs and show that ZCA, standardized ZCA, and Cholesky decomposition for feature transformation can achieve good style transfer results while PCA and standardized PCA are not well applicable to style transfer.

References

[1] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stylebank: An explicit representation for neural image style transfer. In *Proc. CVPR*, volume 1, page 4, 2017. 1

[2] Tian Qi Chen and Mark Schmidt. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337*, 2016. 1

[3] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *Proc. of ICLR*, 2017. 1

[4] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346. ACM, 2001. 1

[5] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 262–270, 2015. 1

[6] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. 1, 2

[7] Golnaz Ghiasi, Honglak Lee, Manjunath Kudlur, Vincent Dumoulin, and Jonathon Shlens. Exploring the structure of a real-time, arbitrary neural artistic stylization network. *arXiv preprint arXiv:1705.06830*, 2017. 1

[8] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340. ACM, 2001. 1

[9] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1510–1519. IEEE, 2017. 1

[10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 1

[11] Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018. 1

[12] Chuan Li and Michael Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2479–2486, 2016. 1

[13] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016. 1

[14] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, pages 386–396, 2017. 1, 2, 3, 4, 6

[15] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. *arXiv preprint arXiv:1802.06474*, 2018. 6

[16] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017. 1

[17] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015. 1

[18] Eric Risser, Pierre Wilmot, and Connelly Barnes. Stable and controllable neural texture synthesis and style transfer using histogram losses. *arXiv preprint arXiv:1701.08893*, 2017. 1

[19] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8242–8250, 2018. 1

[20] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, pages 1349–1357, 2016. 1

[21] Hao Wang, Xiaodan Liang, Hao Zhang, Dit-Yan Yeung, and Eric P Xing. Zm-net: Real-time zero-shot image manipulation network. *arXiv preprint arXiv:1703.07255*, 2017. 1