Machine Learning (2024) 113:3485–3503
https://doi.org/10.1007/s10994-023-06435-5

Style spectroscope: improve interpretability and controllability through Fourier analysis

Zhiyu Jin[1] · **Xuli Shen**[1] · **Bin Li**[1] · **Xiangyang Xue**[1]

Received: 12 June 2023 / Revised: 25 August 2023 / Accepted: 7 October 2023 /
Published online: 9 January 2024
© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2024

Abstract
Universal style transfer (UST) infuses styles from arbitrary reference images into content images. Existing methods, while enjoying many practical successes, are unable of explaining experimental observations, including different performances of UST algorithms in preserving the spatial structure of content images. In addition, methods are limited to cumbersome global controls on stylization, so that they require additional spatial masks for desired stylization. In this work, we first provide a systematic Fourier analysis on a general framework for UST. We present an equivalent form of the framework in the frequency domain. The form implies that existing algorithms treat all frequency components and pixels of feature maps equally, except for the zero-frequency component. We connect Fourier amplitude and phase with a widely used style loss and a well-known content reconstruction loss in style transfer, respectively. Based on such equivalence and connections, we can thus interpret different structure preservation behaviors between algorithms with Fourier phase. Given the interpretations, we propose two plug-and-play manipulations upon style transfer methods for better structure preservation and desired stylization. Both qualitative and quantitative experiments demonstrate the improved performance of our manipulations upon mainstreaming methods without any additional training. Specifically, the metrics are improved by 6% in average on the content images from MS-COCO dataset and the style images from WikiArt dataset. We also conduct experiments to demonstrate (1) the above-mentioned equivalence, (2) the interpretability based on Fourier amplitude and phase and (3) the controllability associated with frequency components.

Keywords Universal style transfer · Fourier transform · Structure preservation · Phase and amplitude

1 Introduction

Style transfer deals with the problem of synthesizing an image which has the style characteristics from a style image and the content representation from a content image. The seminal work of Gatys et al. (2016) uses Gram matrices of feature maps to model style

Editors: Vu Nguyen, Dani Yogatama.

Extended author information available on the last page of the article

characteristics and optimizes reconstruction losses between the reference images and the stylized images iteratively. For the purpose of gaining vivid visual styles and less computation cost, more trained feed-forward networks are proposed (Chen et al., 2017; Dumoulin et al., 2017; Johnson et al., 2016; Li et al., 2017a; Li & Wand, 2016; Sheng et al., 2018; Wang et al., 2020). Recent works focus on arbitrary style transfer (Chen et al., 2021; Park & Lee, 2019), or artistic style ((Chen et al., 2021)). These works capture limited types of style and cannot well generalize to unseen style images (Hong et al., 2021).

To obtain the generalization ability for arbitrary style images, many methods are proposed for the task of universal style transfer (UST). Essentially, the main challenge of UST is to properly extract the style characteristics from style images and transfer them onto content images without any prior knowledge of target style. The representative methods of UST consider various notions of style characteristics. For example, AdaIN (Huang & Belongie, 2017a) aligns the channel-wise means and variances of feature maps between content images and style images, and WCT (Li et al., 2017b) further matches up the covariance matrices of feature maps by means of whitening and coloring processes, leading to more expressive colors and intensive stylization.

While both approaches and their subsequent works exhibit remarkable stylization capabilities, they demonstrate different behavior in terms of generation, including the retention of the underlying structure of content images. For example, it is observed that the operations performed by WCT might introduce structural artifacts and distortions. Many follow-up works focus on alleviating the problem of WCT (Chiu & Gurari, 2022; Li et al., 2018; Yoo et al., 2019), but seldom can analytically and systematically explain what makes the difference. In the field of UST, we need an analytical theory to bridge algorithms with experimental phenomena for better interpretability, potentially leading to better stylization controls. To this end, we resort to apply Fourier transform for deep analysis, aiming to find new equivalence in the frequency domain and bring new interpretations and practical manipulations to existing style transfer methods.

In this work, we first revisit and expand the framework by Li et al. (2017b) which unifies several well-known UST methods. Based on the framework, we derive an equivalent form for it in the frequency domain, which has the same simplicity with its original form in the spatial domain. Accordingly, the derived result demonstrates that these UST methods perform a uniform transformation in the frequency domain except for the origin point. Furthermore, these UST methods transform frequency components (excluding the zero-frequency component) and spatial pixels of feature maps in an identical manner. Thus, these UST methods perform manipulations on the whole frequency domain instead of specific subsets of frequencies (either high frequencies or low frequencies).

Secondly, through the lens of the Fourier transform, we further explore the relation of Fourier phase and amplitude with key notions in style transfer, and then we present new interpretations based on the equivalence we have. On one hand, we prove that a content reconstruction loss between two feature maps reaches a local minimum when they have identical Fourier phase, which implies that the Fourier phase of feature maps contributes to the structure of stylized results. On the other hand, we prove that the Fourier amplitude of feature maps determines the style loss between feature maps, which implies that Fourier amplitude contributes to the intensity information of stylization presentations in images. Next, We demonstrate that WCT does not preserve the Fourier phase of feature maps

compared with MCCNet and AdaIN, and we interpret the different behaviors between the UST methods in structure preservation as a consequence of their different treatment with the Fourier phase of feature maps.

Thirdly, based on the connection we establish between style transfer and Fourier transfer, we propose two manipulations on the frequency components of feature maps: (1) a phase replacement operation to keep the phase of feature maps unchanged during stylization for better structure preservation. (2) a feature combination operation to assign different weights to different frequency components of feature maps for desired stylization. We then conduct extensive experiments to validate their efficacy.

The contributions of this paper are summarized as follows:

- *Equivalence* We present a theoretically equivalent form for several state-of-the-art UST methods in the frequency domain and reveal their effects on frequencies. We conduct corresponding experiments to validate the equivalence.
- *Interpretability* We connect Fourier amplitude and phase with key losses in style transfer and present new interpretations on different behaviors of UST methods. The interpretations are validated by experiments.
- *Controllability* We propose two manipulations for structure preservation and desired stylization. We have experimental validation for their efficacy and controllability.

2 Preliminaries

2.1 Fourier transform

The Fourier transform has been widely used for the analysis of the frequency components in signals, including images and feature maps in the shallow layers of neural networks. Given an image $F \in \mathbb{R}^{C\times H\times W}$, the discrete Fourier transform (DFT) (Jenkins & Desai, 1986) decomposes it into a unique representation $\mathcal{F} \in \mathbb{C}^{C\times H\times W}$ in the frequency domain as follows:

$$\mathcal{F}_{u,v} = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} F_{h,w} e^{-j2\pi\left(u\frac{h}{H}+v\frac{w}{W}\right)}, \quad j^2 = -1, \tag{1}$$

where (h, w) and (u, v) are the indices on the spatial dimensions and the frequency dimensions, respectively. Since images and feature maps consist of multiple channels, we here apply the Fourier transform upon each channel separately and omit the explicit notation of channels. Each frequency component $\mathcal{F}_{u,v}$ can be decomposed into its amplitude $|\mathcal{F}_{u,v}|$ and its phase $\angle\mathcal{F}_{u,v}$:

$$|\mathcal{F}_{u,v}| = \sqrt{\left(\mathcal{R}_{u,v}\right)^2 + \left(\mathcal{I}_{u,v}\right)^2}, \quad \angle\mathcal{F}_{u,v} = \texttt{atan2}(\mathcal{I}_{u,v}, \mathcal{R}_{u,v}), \tag{2}$$

where $\mathcal{R}_{u,v}$ and $\mathcal{I}_{u,v}$ are the real part and the imaginary part of the complex frequency component $\mathcal{F}_{u,v}$, respectively. Intuitively, as for images, amplitude carries much of intensity information, including the contrast or the difference between the brightest and darkest

peaks of images, and phase crucially determines the spatial content of images (Gonzalez & Woods, 2008).

2.2 A unified framework for universal style transfer

To better demonstrate the connection between style transfer and the Fourier transform, a unified framework of different style transfer methods is preferred to serve as a bridge. Given a content image I^c and a style image I^s, we denote the feature maps of I^c and I^s as $F^c \in \mathbb{R}^{C\times H^c\times W^c}$ and $F^s \in \mathbb{R}^{C\times H^s\times W^s}$ respectively, where C denotes the number of channels, H^c (H^s) the height and W^c (W^s) the width. For a majority of UST methods, their goal is to transform the content image feature maps F^c into stylized feature maps F^{cs}, whose first-order and second-order statistics are aligned with those of the style image feature maps F^s. Accordingly, their methods mainly depend on the corresponding channel-wise mean vectors $\mu^c, \mu^s \in \mathbb{R}^C$ and the covariance matrices $\Sigma^c, \Sigma^s \in \mathbb{R}^{C\times C}$ of F^c and F^s, respectively.

A framework is proposed in (Lu et al., 2019) for unifying several well-known methods [AdaIN (Huang & Belongie, 2017a), WCT (Li et al., 2017b), and OptimalWCT (Lu et al., 2019)] under the same umbrella. To expand upon this framework, we have identified several following methods that could be integrated into the framework, including LinearWCT (Li et al., 2019), MAST (Huo et al., 2021) and MCCNet (Deng et al., 2021), which provide additional insights for the refinement and improvement of the existing framework. Specifically, each pixel $F^c_{h,w}$ of F^c is first centralized by subtracting the mean vector μ^c, where h and w are indices on spatial dimensions. Then the framework linearly transforms $F^c_{h,w}$ with the transformation matrix $T \in \mathbb{R}^{C\times C}$ and re-centers $F^c_{h,w}$ by adding the mean vector μ^s of the style. Each pixel $F^{cs}_{h,w} \in \mathbb{R}^C$ of stylized feature maps can be represented as follows:

$$F^{cs}_{h,w} = T\left(F^c_{h,w} - \mu^c\right) + \mu^s, \tag{3}$$

where the transformation matrix T has multiple forms based on a variety of configurations of different methods. We here demonstrate the relation between the unified framework and several methods in details.

1. *AdaIN* In *Adaptive Instance Normalization* (AdaIN) (Huang & Belongie, 2017a), the transformation matrix $T = \mathtt{diag}(\Sigma^s)/\mathtt{diag}(\Sigma^c)$, where $\mathtt{diag}(\Sigma)$ denotes the diagonal matrix of a given matrix Σ and / denotes the element-wise division. Because of the characteristics of diagonal matrices, only the means and variances within each single feature channel of F^{cs} are matched up to those of F^s, ignoring the correlation between channels.
2. *WCT* Instead of shifting a single set of intra-channel statistics, Li et al. (2017b) propose a *Whitening and Coloring Transform* (WCT) that focuses further on the alignment of covariance matrices. Similar with AdaIN, the transformation matrix for WCT is $T = (\Sigma^s)^{\frac{1}{2}}(\Sigma^c)^{-\frac{1}{2}}$, leading to well-aligned second-order statistics.
3. *OptimalWCT* Similarly, OptimalWCT (Lu et al., 2019) is proposed to derive a closed-form solution for T without the help of an optimization process:

$$T = (\Sigma^c)^{-\frac{1}{2}}\left((\Sigma^c)^{\frac{1}{2}}\Sigma^s(\Sigma^c)^{\frac{1}{2}}\right)^{\frac{1}{2}}(\Sigma^c)^{-\frac{1}{2}}. \tag{4}$$

Their method reaches the theoretical local minimum for the content loss $\mathcal{L}_c = \|F^c - F^{cs}\|_F^2$, which is widely-used in style transfer (Gatys et al., 2016; Huang & Belongie, 2017a; Lu et al., 2019) for the structure preservation of content images.

4. *LinearWCT* While WCT generates stylized images more expressively, it is still computationally expensive because of the high dimensions of feature maps in neural networks. Li et al. (2019) propose LinearWCT to use light-weighted neural networks to model the linear transformation T by optimizing the Gram loss, known as a widely-used style reconstruction objective function:

$$T = \arg\min_T \|F^{cs}F^{cs\top} - F^sF^{s\top}\|_F^2, \tag{5}$$

where $F^{cs}F^{cs\top}$ is the Gram matrix for F^{cs} and $\|\cdot\|_F^2$ denotes the squared Frobenius norm of the differences between given matrices.

5. *MAST* Different with WCTs and derivative works, Huo et al. (2021) view style transfer as an alignment of two multi-manifold distributions and propose a Manifold Alignment based Style Transfer (MAST) method with transformation matrix:

$$T = \left(F^cD^cF^{c\top}\right)^{-1}\left(F^cU^{cs}F^{s\top}\right). \tag{6}$$

where U_{cs} are regularized affinity matrices to indicate the neighbors in feature space for manifold alignment.

6. *MCCNet* Deng et al. (2021) propose to realign and mix style features based on their similarity to content features with transformation matrix:

$$T = I + \mathrm{diag}\big(\Sigma\big(WF_s\big)\big), \tag{7}$$

where I is the identity matrix, W is a learnable diagonal matrix for weighting and $\Sigma(WF_s)$ represents the covariance matrix of WF_s.

3 Method

In this section, we first show an equivalent form of the framework in the frequency domain. In this way, all the methods based on the framework in Sect. 2.2 can be interpreted as effecting on the frequency domain. We further connect amplitude and phase with existing concepts in the context of style transfer, and explain why WCT might not preserve the structure of content images. Finally, we propose two operations for better structure preservation and desired stylization.

3.1 The equivalent form of the framework in the frequency domain

We theoretically analyze the unified framework from the angle of 2-D DFT. We denote the DFT of F^{cs} as $\mathcal{F}^{cs} \in \mathbb{C}^{C\times H^c\times W^c}$, where $\mathbb{C}$ is the set of complex numbers. According to the unified framework in Eq. (3), we can derive each complex frequency component $\mathcal{F}^{cs}_{u,v}$ as:

$$
\begin{aligned}
\mathcal{F}^{cs}_{u,v} &= \sum_{h=0}^{H^c-1}\sum_{w=0}^{W^c-1} F^{cs}_{h,w} e^{-j2\pi\left(u\frac{h}{H^c}+v\frac{w}{W^c}\right)} \\
&= \sum_{h=0}^{H^c-1}\sum_{w=0}^{W^c-1} [T(F^c_{h,w}-\mu^c)+\mu^s] e^{-j2\pi\left(u\frac{h}{H^c}+v\frac{w}{W^c}\right)} \\
&= T\underbrace{\sum_{h=0}^{H^c-1}\sum_{w=0}^{W^c-1} F^c_{h,w} e^{-j2\pi\left(u\frac{h}{H^c}+v\frac{w}{W^c}\right)}}_{\begin{cases} H^cW^c\mu^c, & \text{if } u=v=0; \\ \mathcal{F}^c_{u,v}, & \text{else} \end{cases}} + [\mu^s - T\mu^c]\underbrace{\sum_{h=0}^{H^c-1}\sum_{w=0}^{W^c-1} e^{-j2\pi\left(u\frac{h}{H^c}+v\frac{w}{W^c}\right)}}_{\begin{cases} H^cW^c, & \text{if } u=v=0; \\ 0, & \text{else} \end{cases}} \\
&= \begin{cases} H^cW^c\mu^s, & \text{if } u=v=0; \\ T\mathcal{F}^c_{u,v}, & \text{else} \end{cases} = \begin{cases} \left(\frac{H^cW^c}{H^sW^s}\right)\mathcal{F}^s_{0,0}, & \text{if } u=v=0; \\ T\mathcal{F}^c_{u,v}, & \text{else} \end{cases},
\end{aligned}
\tag{8}
$$

where u and v are indices upon the frequency dimensions, and $\mathcal{F}^c$ and $\mathcal{F}^s$ are the DFTs of F^c and F^s, respectively. According to the Fourier transform in Eq. (1), $\mathcal{F}_{0,0} = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} F_{h,w} = HW\mu$. Thus, we have $\mathcal{F}^{cs}_{u,v} = H^cW^c\mu^s = \left(\frac{H^cW^c}{H^sW^s}\right)\mathcal{F}^s_{0,0}$ when $u=v=0$. Therefore, in the frequency domain, style transfer methods based on the unified framework are essentially linear transformations on $\mathcal{F}^c$ except for the zero-frequency component $\mathcal{F}^c_{0,0}$, which is replaced with the re-scaled zero-frequency component of $\mathcal{F}^s$.

From Eq. (8), we find that each individual frequency component (excluding the zero-frequency component) has an identical linear transformation with pixels on the feature maps. In this way, there is no entanglement between different frequencies in the process of style transfer. Thus, it is feasible to treat and manipulate each frequency component of $\mathcal{F}^{cs}$ as an individual for practical usage. Therefore, it is justified that mainstream methods in Sect. 2.2 for UST are not sole transfer on specific subsets of frequencies (either high frequencies or low frequencies), but essentially on the whole frequency domain.

3.2 Connections and interpretations: amplitude and phase

To better bridge style transfer with the Fourier transform, we connect phase and amplitude with a reconstruction loss and a widely-used style loss in style transfer, respectively.

Phase and the content loss We here demonstrate the relation between phase and the content loss, which widely serves as a construction loss for optimizing the distances of spatial arrangement between stylized images I^{cs} and content images I^c. Given their feature maps $F^{cs}, F^c \in \mathbb{R}^{C\times H\times W}$, corresponding DFTs $\mathcal{F}^{cs}, \mathcal{F}^c$, Fourier amplitude $|\mathcal{F}^{cs}|, |\mathcal{F}^c|$ and Fourier phase $\angle\mathcal{F}^{cs}, \angle\mathcal{F}^c$, the content loss between F^{cs} and F^c can be derived as:

$$
\begin{aligned}
\mathcal{L}_c &= \sum_{k=0}^{C-1}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\left(F^{cs}_{k,h,w}-F^c_{k,h,w}\right)^2 = \frac{1}{HW}\sum_{k=0}^{C-1}\sum_{u=0}^{H-1}\sum_{v=0}^{W-1}|\mathcal{F}^{cs}_{k,u,v}-\mathcal{F}^c_{k,u,v}|^2 \\
&= \frac{1}{HW}\sum_{k=0}^{C-1}\sum_{u=0}^{H-1}\sum_{v=0}^{W-1}\left[|\mathcal{F}^{cs}_{k,u,v}|^2+|\mathcal{F}^c_{k,u,v}|^2-2|\mathcal{F}^{cs}_{k,u,v}||\mathcal{F}^c_{k,u,v}|\cos\left(\angle\mathcal{F}^{cs}_{k,u,v}-\angle\mathcal{F}^c_{k,u,v}\right)\right],
\end{aligned}
\tag{9}
$$

where the second equality is held by the Parseval's theorem and k, (h, w) and (u, v) are indices on channels, spatial dimensions and frequency dimensions, respectively. When F^{cs}

is optimized for the content loss, since $|\mathcal{F}^{cs}_{k,u,v}|$ and $|\mathcal{F}^{c}_{k,u,v}|$ are non-negative numbers, the content loss $\mathcal{L}_c$ reaches a local minimum when $\angle\mathcal{F}^{cs}_{k,u,v} = \angle\mathcal{F}^{c}_{k,u,v}$ for all (k, u, v). Furthermore, whenever $\angle\mathcal{F}^{cs}_{k,u,v}$ gets closer to $\angle\mathcal{F}^{c}_{k,u,v}$, the content loss decreases, demonstrating the crucial role of the phase of feature maps in determining the spatial information of corresponding decoded images. Therefore, we can interpret the structure preservation abilities of methods from the perspective of Fourier phase. Furthermore, we can manipulate Fourier phase for better performances in structure preservation.

Interpretations on structure preservation Based on the equivalent form in Eq. (8) and the relation between Fourier phase and the content loss, we can give interpretations to different behaviors of methods in structure preservation. Concerning AdaIN, WCT and MCCNet as instances of the equivalent framework in the frequency domain, we have $\mathcal{F}^{cs}_{u,v} = T\mathcal{F}^{c}_{u,v}$ when $(u, v) \neq (0, 0)$. Note that for AdaIN and MCCNet (the latter is directly designed for the better preservation of the content structure), their transformation matrices $T = \mathtt{diag}(\Sigma^s)/\mathtt{diag}(\Sigma^c)$ and $T = I + \mathrm{diag}\big(\Sigma\big(WF_s\big)\big)$ are real diagonal matrices, which have the same scaling upon the real part and the imaginary part of $\mathcal{F}^c$. In this way, AdaIN and MCCNet preserve the phase in each feature channel and keep the content loss of feature maps in a local minimum. As a result, they have the ability to better preserve the content structure. While WCT provides a non-diagonal matrix $T = (\Sigma^s)^{\frac{1}{2}}(\Sigma^c)^{-\frac{1}{2}}$ for transformation, the information between different channels is consequently entangled, the phase of each channel is disturbed and the content loss after the process of WCT is likely to increase much more than the ones after the process of AdaIN and MCCNet. Additionally, similar ideas emerge in Huo et al. (2021) with explanations from the perspective of self-similarity, which accords with our interpretations. In this way, WCT needs more efforts to preserve the spatial information of content images, resulting in its less appealing performances in structure preservation.

Amplitude and the style loss We theoretically demonstrate the connection between the Fourier amplitude of feature maps and a style loss. Given feature maps $F \in \mathbb{R}^{C\times H\times W}$, corresponding Fourier amplitude $|\mathcal{F}|$ and Fourier phase $\angle\mathcal{F}$ of their DFT $\mathcal{F}$, the style loss between F^{cs} and F^s can be derived as:

$$
\begin{aligned}
\mathcal{L}_s &= \|\mu^{cs} - \mu^s\|_2 + \|\sigma^{cs} - \sigma^s\|_2 \\
\|\mu^{cs} - \mu^s\|_2 &= \sum_{k=0}^{C}\left(\mu^{cs}_k - \mu^s_k\right)^2 = \frac{1}{H^2W^2}\sum_{k=0}^{C}\left(|\mathcal{F}^{cs}_{k,0,0}| - |\mathcal{F}^{s}_{k,0,0}|\right)^2 \\
\|\sigma^{cs} - \sigma^s\|_2 &= \frac{1}{H^2W^2}\sum_{k=0}^{C}\left[\sum_{h=0}^{H}\sum_{w=0}^{W}\left(F^{cs}_{k,h,w}\right)^2 - \sum_{h=0}^{H}\sum_{w=0}^{W}\left(F^{s}_{k,h,w}\right)^2\right]^2 \\
&= \frac{1}{H^2W^2}\sum_{k=0}^{C}\left[\sum_{h=0}^{H}\sum_{w=0}^{W}\left(\mathcal{F}^{cs}_{k,u,v}\mathcal{F}^{cs*}_{k,u,v} - \mathcal{F}^{s}_{k,u,v}\mathcal{F}^{s*}_{k,u,v}\right)\right]^2 \\
&= \frac{1}{H^2W^2}\sum_{k=0}^{C}\left[\sum_{h=0}^{H}\sum_{w=0}^{W}\left(|\mathcal{F}^{cs}_{k,u,v}|^2 - |\mathcal{F}^{s}_{k,u,v}|^2\right)\right]^2,
\end{aligned}
\tag{10}
$$

where $(*)$ represents complex conjugate. The observation that the style loss can be expressed as the squared sum of squared differences between Fourier amplitude components suggests that these components have a direct influence on the style loss. Therefore, if we only manipulate the Fourier phase of the DFTs of feature maps and keep Fourier

amplitude unchanged, it can be expected that the stylization intensity and presentations of the corresponding decoded images are roughly the same.

3.3 Training-free manipulations on stylized feature maps in the frequency domain

The equivalent form in Eq. (8) and abovementioned connections enable further manipulations for better structure preservation or desired stylization. We propose two simple but effective operations upon the frequency components of feature maps called phase replacement and frequency combination, both of which are training-free and plug-and-play for the style transfer models.

3.3.1 Phase replacement

Given the DFT of the content feature maps $\mathcal{F}^c$ and the DFT of the stylized feature maps $\mathcal{F}^{cs}$, we calculate the phase of $\mathcal{F}^c$, denoted as $\angle\mathcal{F}^c \in [0, 2\pi)^{C\times H^c\times W^c}$ and the amplitude of $\mathcal{F}^{cs}$, denoted as $|\mathcal{F}|^{cs} \in \mathbb{R}_+^{C\times H^c\times W^c}$, where $\mathbb{R}_+$ denotes the set of non-negative real numbers. We then reconstruct $\mathcal{F}^{cs}$ as:

$$\mathcal{F}^{cs}_{u,v} = |\mathcal{F}|^{cs}_{u,v} \odot \cos\angle\mathcal{F}^{c}_{u,v} + j|\mathcal{F}|^{cs}_{u,v} \odot \sin\angle\mathcal{F}^{c}_{u,v}, \tag{11}$$

where cos and sin are element-wise operators on vectors (e.g., $\cos\boldsymbol{\phi} = [\cos\boldsymbol{\phi}^1, \ldots, \cos\boldsymbol{\phi}^C]$), $\odot$ is the element-wise multiplication and j is the imaginary unit. Based on the connections established in Sect. 3.2, when we replace the phase of $\mathcal{F}^{cs}$ with $\angle\mathcal{F}^c$, the content loss between F^{cs} and F^c is reduced and in this way, the structure of content images is more preserved in F^{cs}. In addition, since the amplitude of $\mathcal{F}^{cs}$ is not changed, the style loss stays unchanged and so does the style information of the generated results.

We provide the sketched style transfer methods with phase replacement as Algorithm 1 shows. It is worth noting that the phase replacement works in a plug-and-play manner controlled by users flag and we perform Fast Fourier Transform only for once thanks to the equivalence proved in Eq. (8). As a result, we achieve controllable structure preservation without any additional training with limited cost.

Algorithm 1 Style transfer methods with phase replacement.

Require: The content images $\mathbf{I}^c$, the style images $\mathbf{I}^s$, the trained encoder $\mathbf{E}^{(i)}$, the trained decoder $\mathbf{D}^{(i)}$, the total number of iterations n, the flag for phase replacement $flag$ and the function f to compute transformation matrices $\mathbf{T}$.

for $i = 1$ to n **do**
  Encode content and style images: $\mathbf{F}^c \Leftarrow \mathbf{E}^{(i)}(\mathbf{I}^c)$ and $\mathbf{F}^s \Leftarrow \mathbf{E}^{(i)}(\mathbf{I}^s)$.
  Calculate the transformation matrices: $\mathbf{T} \Leftarrow f(\mathbf{F}^c, \mathbf{F}^s)$.
  if $flag$ is True **then**
    Calculate the FFT of the encoded content feature maps: $\mathcal{F}^c \Leftarrow \texttt{FFT}(\mathbf{F}^c)$.
    Calculate the stylized FFT with Eq. (8): $\mathcal{F}^{cs} \Leftarrow \mathbf{T}\mathcal{F}^{cs}$ and $\mathcal{F}^{cs}_{0,0} \Leftarrow H^cW^c\mu^s$.
    Replace the phase of the stylized FFT: $\mathcal{F}^{cs} \Leftarrow \mathcal{F}^c/\|\mathcal{F}^c\| \odot \|\mathcal{F}^{cs}\|$.
    Calculate the Inverse FFT of the stylized FFT: $\mathbf{F}^{cs} \Leftarrow \texttt{InvFFT}(\mathcal{F}^{cs})$.
  else
    Calculate the stylized feature maps: $\mathbf{F}^{cs} \Leftarrow \mathbf{T}(\mathbf{F}^c - \mu^c) + \mu^s$
  end if
  Decode $\mathbf{F}^{cs}$ as the content images for next iteration: $\mathbf{I}^c \Leftarrow \mathbf{D}^{(i)}(\mathbf{F}^{cs})$.
end for

3.3.2 Frequency combination

To accommodate different requirements from users, appropriate control on stylization is needed for practical usage. Plenty of works for style transfer use linear combination of content feature maps F^c and stylized feature maps F^{cs} as shown in Eq. (12):

$$F^{cs} = \alpha F^{cs} + (1 - \alpha)F^c, \tag{12}$$

where α is the weight for controlling on the stylization. In this way, all the global characteristics of images (e.g., the sharp edges of trees and the smooth background of sky) are combined uniformly. While in most cases, users are expecting for customized global changes on images (e.g., having the details of trees less stylized but keeping the sky moderately stylized). Since high frequencies determine the details and low frequencies determine the overview of images, we can accommodate the customized needs of users with a combination of frequencies in different proportions.

Given the DFT of content feature maps $\mathcal{F}^c$ and the DFT of stylized feature maps $\mathcal{F}^{cs}$, we first rearranges their frequency components with the zero-frequency components in the center point (u_0, v_0), following a common technique in digital image processing. In this way, the frequency components close to (u_0, v_0) are low-frequency components whereas the rest of components represent high frequencies. Next, we combine $\mathcal{F}^{cs}$ and $\mathcal{F}^c$ using a weighting function $\alpha : \mathbb{R}^2 \rightarrow [0, 1]$:

$$\mathcal{F}^{cs}_{u,v} = \alpha(u, v)\mathcal{F}^{cs}_{u,v} + [1 - \alpha(u, v)]\mathcal{F}^{c}_{u,v}, \tag{13}$$

where α serves as the stylization weighting function dependent on the indices (u, v). For example, if users want to have the details less stylized, higher frequencies of $\mathcal{F}^{cs}$ need to be less weighted, and accordingly a lower value of α can be set for (u, v) indexing higher frequencies. In practice, the function α is set to be controlled by a hyper-parameter σ:

$$\alpha(u, v) = \exp\left[-\frac{\left(u - u_0\right)^2 + \left(v - v_0\right)^2}{\sigma}\right], \tag{14}$$

where σ represents the degree for combining the low frequencies of $\mathcal{F}^{cs}$. When σ gets larger, the value of $\alpha(u, v)$ increases for every (u, v). In this way, more low frequencies of $\mathcal{F}^{cs}$ (indexed by (u, v) close to (u_0, v_0)) are gradually kept. We provide the sketched style transfer methods with our frequency combination as Algorithm 2 shows.

Algorithm 2 Style transfer methods with frequency combination.

Require: The content images $\mathbf{I}^c$, the style images $\mathbf{I}^s$, the trained encoder $\mathbf{E}^{(i)}$, the trained decoder $\mathbf{D}^{(i)}$, the total number of iterations n, the flag for phase replacement $flag$ and the function f to compute transformation matrices $\mathbf{T}$.

for $i = 1$ to n **do**
  Encode content and style images: $\mathbf{F}^c \Leftarrow \mathbf{E}^{(i)}(\mathbf{I}^c)$ and $\mathbf{F}^s \Leftarrow \mathbf{E}^{(i)}(\mathbf{I}^s)$.
  Calculate the transformation matrices: $\mathbf{T} \Leftarrow f(\mathbf{F}^c, \mathbf{F}^s)$.
  if $flag$ is True **then**
    Calculate the FFT of the encoded content feature maps: $\mathcal{F}^c \Leftarrow \texttt{FFT}(\mathbf{F}^c)$.
    Calculate the stylized FFT with Eq. (8): $\mathcal{F}^{cs} \Leftarrow \mathbf{T}\mathcal{F}^{cs}$ and $\mathcal{F}^{cs}_{0,0} \Leftarrow H^c W^c \mu^s$.
    Calculate the frequency combination weight $\alpha(u, v)$ in Eq. (14).
    Combine the frequency of FFTs: $\mathcal{F}^{cs}_{u,v} \Leftarrow \alpha(u,v)\mathcal{F}^{cs}_{u,v} + [1 - \alpha(u,v)]\mathcal{F}^{c}_{u,v}$.
    Calculate the Inverse FFT of the stylized FFT: $\mathbf{F}^{cs} \Leftarrow \texttt{InvFFT}(\mathcal{F}^{cs})$.
  else
    Calculate the stylized feature maps: $\mathbf{F}^{cs} \Leftarrow \mathbf{T}(\mathbf{F}^c - \mu^c) + \mu^s$.
  end if
  Decode $\mathbf{F}^{cs}$ as the content images for next iteration: $\mathbf{I}^c \Leftarrow \mathbf{D}^{(i)}(\mathbf{F}^{cs})$.
end for

4 Experiments

In this section, we first introduce our method specification and implementation details. We implement our method upon the state-of-the-art style transfer methods and make comparison in terms of visual effect and structure preservation.

Method specification Based on the equivalence and connections mentioned above, the proposed phase replacement manipulation performs on the basis of UST algorithms in the frequency domain. In practice, we choose to implement our method in conjunction with WCT (Li et al., 2017b), OptimalWCT (Lu et al., 2019) and MAST (Huo et al., 2021). We skip AdaIN (Huang & Belongie, 2017b) and MCCNet (Deng et al., 2021) because their phase of stylization results has been the same with that of content images. We also pass over LinearWCT (Li et al., 2019) because its official implementation does not strictly follow the unified framework in Eq. (5). In regards to WCT, OptimalWCT and MAST, we adopt the phase replacement onto three mentioned algorithms by substituting the Fourier phase of stylized feature maps with that of content feature maps. Since the phase replacement can optimize the content loss to a local minimum according to Eq. (9), the structure of content images is preserved and the overwhelming stylization is alleviated. Finally, the proposed method uses inverse discrete Fourier transform to reverse the frequency components back to the spatial domain. It is worth noting that the our phase replacement needs no additional training or fine-tuning and works in a plug-and-play manner, which could be activated to the discretion of users.

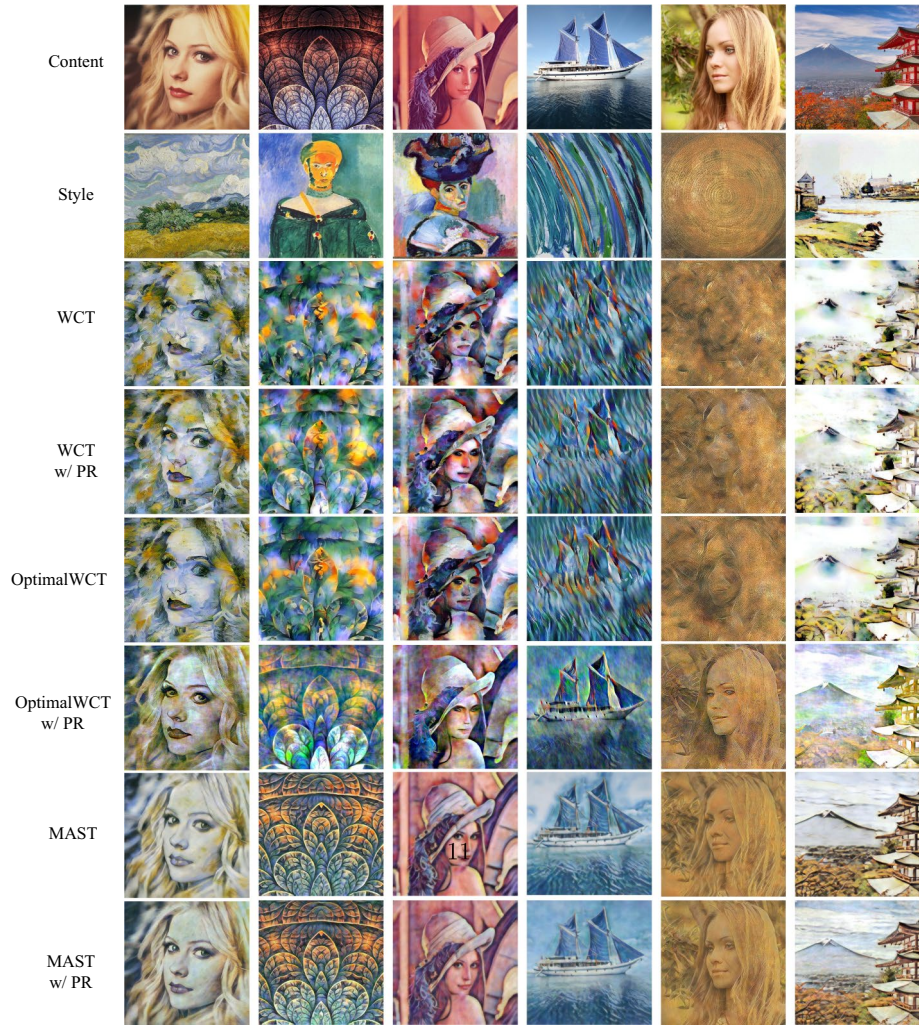

Fig. 1 Qualitative comparison on the state-of-the-art UST algorithms. The proposed phase replacement (*abbr.* PR) could improve the preservation of content structure while maintaining the stylization presentations

4.1 Qualitative comparison

We first conduct a qualitative comparison on visual effect and structure preservation. To be more specific, we implement our phase replacement manipulation upon mentioned methods and visualize the results for the comparison.

Results In Fig. 1, we show some visualization results of the qualitative comparison between the state-of-the-art UST methods with and without the proposed phase replacement manipulation [i.e., WCT (Li et al., 2017b), OptimalWCT (Lu et al., 2019) and

MAST (Huo et al., 2021)]. We observe that WCT and OptimalWCT could introduce intensive but distorted artistic style and yield images less similar with content images in structure (e.g., 2nd and 4th columns). The visual effect of their stylization results is less attractive especially when content images belong to human portraits (e.g., 1st, 3rd, and 5th columns). MAST does preserve the spatial structure of content images, but the stylization does not well in the shadow and illumination. (e.g., 1st and 4th columns). Comparatively, the proposed method improves the preservation of the spatial structure for content images, including the details (e.g., the spatial arrangements of human faces and the contours of leaves and ships) and the overview (e.g., better presentation of blue sky and more vivid illustration of human faces) of images. Additionally, we observe that with better spatial arrangements, phase replacement generally keeps the stylization intensity, including the color and the contrast for stylization, which accords with our theoretical findings in Eq. (10).

4.2 Quantitative comparison

In addition to the qualitative comparison, we conduct quantitative comparison on visual effect, structure preservation and computing time. In particular, we implement the proposed phase replacement manipulation upon mentioned methods and present corresponding quantitative scores with multiple metrics to comprehensively evaluate the performance.

Metrics For the quantitative comparison, we adopt three objective metrics: (1) Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), which aims to judge the perceptual similarity between two images. (2) the Peak Signal-to-noise ratio (PSNR), which is widely used to measure the quality of image reconstruction. (3) the Structural Simiarlity Index (SSIM) (Wang et al., 2004), which is a well known quality metric for the similarity between images. As for the visual effect, we also include human evaluation scores on stylization artistic presentation and overall visual preference, which is representative in image generation tasks.

Evaluation details We randomly choose 15 images from MS-COCO dataset (Lin et al., 2014) as content images and 20 images from WikiArt dataset (Nichol, 2016) as style

Table 1 The LPIPS scores (↓), the PSNR scores (↑), the SSIM scores (↑) and the user study results [artistic scores (↑) and preference ratings (↑)] for different UST methods

Methods	LPIPS (↓)	PSNR (↑)	SSIM (↑)	Art. (↑)	Pref. (↑)
AdaIN	0.630	11.382	0.309	5.727	6.588
MCCNet	0.612	11.872	0.404	6.875	7.188
LinearWCT	0.591	11.771	0.398	7.142	7.235
WCT	0.737	9.473	0.237	6.121	6.628
WCT w/PR	**0.704**	**10.032**	**0.254**	6.143	**6.978**
OptimalWCT	0.710	9.627	0.252	6.676	6.779
OptimalWCT w/PR	**0.687**	**10.334**	**0.275**	6.339	**6.952**
MAST	0.538	12.568	0.465	7.443	7.413
MAST w/PR	**0.466**	**13.763**	**0.489**	7.312	**7.677**

PR, Art. and Pref. represent the acronym for Phase Replacement, Artistic scores and Preference scores, respectively. Our proposed phase replacement manipulation maintains the artistic ratings and improves all the other metrics (denoted in bold fonts) on the performance of structure preserving and visual effect, which accords with our theoretical findings and interpretations

Table 2 The time cost (seconds) for different UST methods with and without phase replacement (*abbr.* PR)

Method	WCT	OptimalWCT	MAST
Time	0.3167	0.6247	0.058
Time (w/PR)	0.3273	0.6321	0.065

It is worth noting that the additional time cost for phase replacement is limited (less than 0.01 s) and could be activated to the discretion of users

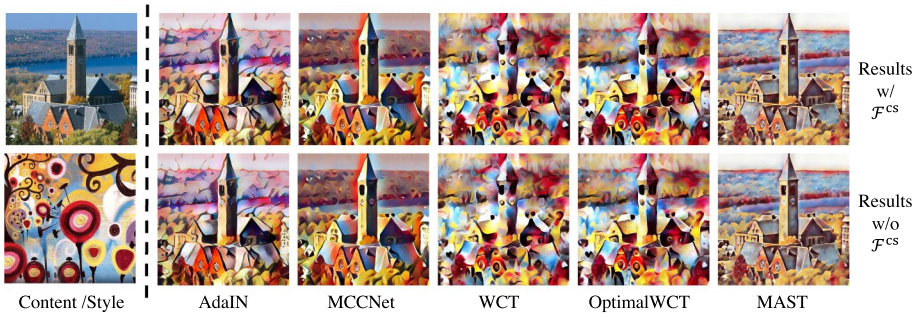

Fig. 2 Visualized results on the validation of equivalence. Results w/$\mathcal{F}^{cs}$ are synthesized following Eq. (8) in the frequency domain, which perform equivalent stylization with the results of original methods, denoted as Results w/o $\mathcal{F}^{cs}$. The results validate the equivalence established in Sect. 3.1

images. Then we synthesize 300 groups of stylized images with each group corresponding to a combined content-style image pair. All the objective metrics are evaluated and averaged upon these image groups. In regard to the human evaluation, we invite 30 participants to rate the synthesized image groups given each content-style pair considering the artistic effect and visual preference. Overall, we get 300 ratings and calculate the average rating for artistic scores and human preference scores.

Results and analysis As shown in Table 1, the proposed phase replacement maintains the artistic scores of the mainstream style transfer methods and improves all the other metrics by an average of 6%, including 7.03% in LPIPS scores (better perceptual similarity with content images), 7.58% in PSNR scores (less introduced noise), 7.15% in SSIM scores (better structural similarity with content images) and 3.79% in human preference scores (better visual effect). The results show that phase replacement could improve the performances in structure preservation and visual effect due to the introduction of Fourier phase of content images, while maintaining the artistic intensity of stylization at the same time by keeping the original Fourier amplitude. The improved metrics (denoted in bold fonts) accord with our theoretical findings in Sect. 3 and fully validate our interpretations on the structure preservation behaviors between methods.

Time costs Regarding the computing time, we evaluate the average time cost for style transfer methods with and without the proposed phase replacement. As shown in Table 2, the additional time cost for phase replacement is limited (less than 0.01 s) and users could choose to activate the manipulation since the manipulation is implemented in a plug-and-play manner. It is worth noting that due to the equivalence proved in Eq. (8), we perform Fast Fourier Transform only for content feature maps alone. As a result, the additional time cost for phase replacement is largely cut off and therefore becomes limited.

5 Discussions

In this section, we conduct more experiments to validate the equivalence presented in Eq. (8), the interpretations on Fourier amplitude and phase introduced in Sect. 3.2, and the efficacy of manipulations proposed in Sect. 3.3.

Validation of equivalence On the validation of the equivalence stated in our theoretical findings, we implement multiple style transfer method in the frequency domain based on Eq. (8), including AdaIN, WCT, OptimalWCT and MAST. As shown in Fig. 2, it can be observed that these implemented UST methods in the frequency domain produce the same visual effect with original methods. This observation validates the proposed equivalence, which lays solid foundation for the following interpretations and manipulations.

Interpretations on amplitude and phase To validate the roles of amplitude and phase, we replace the Fourier amplitude or the Fourier phase of stylized feature maps in each layer during the stylization and present the results in Fig. 3. It can be observed that feature maps with the same Fourier phase produce images with highly similar spatial arrangements. This observation matches up with our interpretations on phase, provided by its connection with

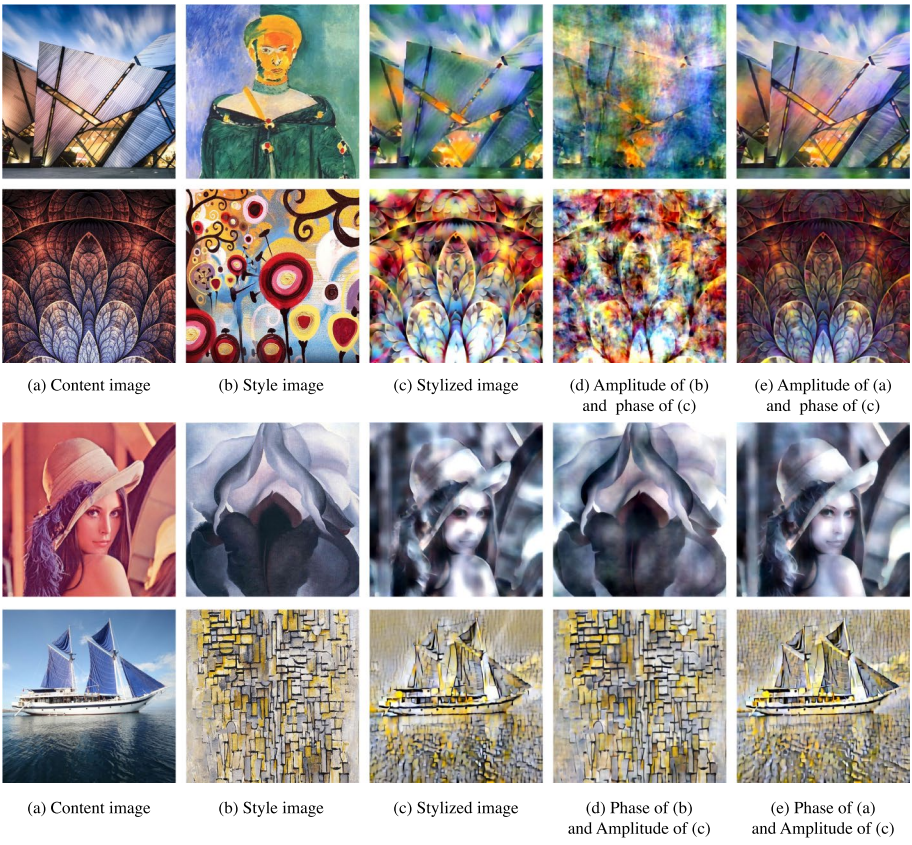

Fig. 3 Synthesized results with replaced Fourier amplitude or phase. The results validate the interpretable roles of Fourier amplitude (encoding the style presentations) and phase (attending to the image structure) in style transfer

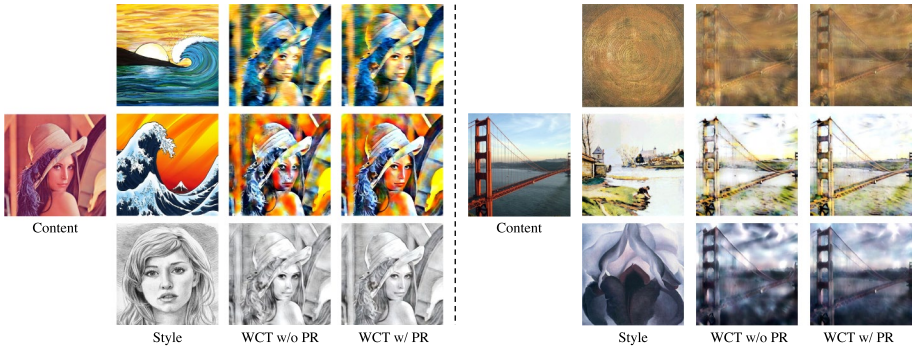

Fig. 4 Visualized results on the efficacy of phase replacement (*abbr.* PR). The results validate the efficacy of phase replacement which could effectively preserve the structure of content images and better arrange image in details (corresponding to phase of high frequencies) and overview presentation (corresponding to phase of low frequencies)

the content loss in Eq. (9). On the other hand, feature maps with same Fourier amplitude produces images with highly similar contrast and intensities in colors. This observation aligns with our interpretations on amplitude, which states that the Fourier amplitude of feature maps has strong connection with style losses in Eq. (10).

Efficacy of phase replacement We empirically display the effect of phase replacement in Sect. 3.3 for image stylization, whose results are shown in Fig. 4. It can be observed that for results without phase replacement, the details (e.g., contours of the eyes and the nose) and the overview (e.g., the sky and the sea) become messier and more distorted, yielding unappealing distortions. Considering these observations, we give the extended interpretation that phase replacement preserves the phase for both high frequencies and low frequencies, which are responsible for the spatial arrangement of the details and overview of images, respectively.

Additionally, we visualize the average Fourier amplitude of stylization results in Sect. 4.2 to validate that phase replacement does maintain the amplitude which encodes stylization presentation implied in Eq. (10). As shown in Fig. 5, the results with and without phase replacement bear a strong resemblance on the amplitude to each other. This validates the efficacy of phase replacement which effectively maintains the Fourier amplitude and thus maintains the stylization intensity and the presentations.

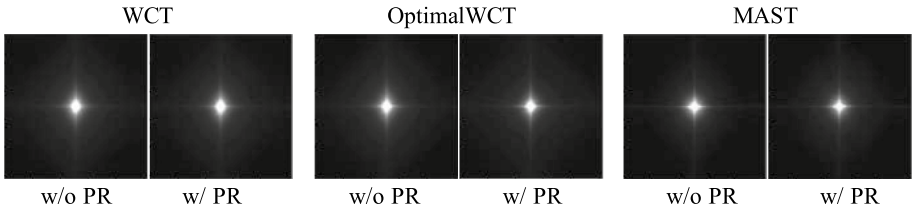

Fig. 5 Visualized Fourier amplitude of stylization results for the efficacy of phase replacement (*abbr.* PR). The observation is that the results with and without phase replacement bear a strong resemblance on the amplitude to each other. The observation validates the efficacy of phase replacement which effectively maintain the Fourier amplitude

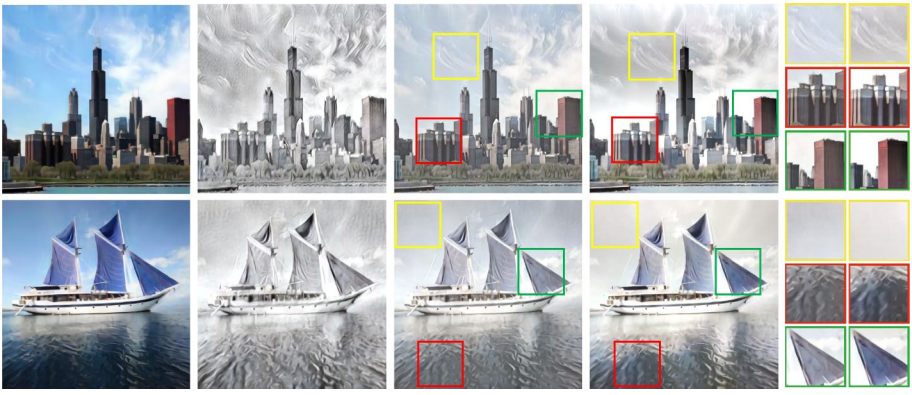

Fig. 6 Comparison between **c** linear combination and **d** frequency combination. It is worth noting that the proposed method helps the background remain stylized (e.g., the sky more sketch-stylized than (**c**)) while renders the details more realistic (e.g., the buildings and the sailboat more realistic than (**c**)). Note that the proposed method achieves this performance without any spatial masks to identify where to stylize. The hyper-parameter σ for results in (**d**) is set to 0.9

Efficacy of frequency combination To demonstrate the manipulations of frequency combination in Sect. 3.3, we present an example in Fig. 6. We choose the weighting function α in Eq. (14) and adjust the hyper-parameter σ for stylization controls. In Fig. 6, with an appropriate value of σ, frequency combination can have the details less stylized (e.g., more colorful and more vivid buildings and ships in the 4th column) while keeping the background moderately stylized (e.g., the sky with more intensive sketch style in the 4th column). In this way, users could have access to the customization of their own stylization results for various purposes by imposing different stylization effect upon local or global parts of images.

Controllability of frequency combination It is worth noting that the linear combination in Eq. (12) can be viewed as a specialized instance of frequency combination by setting the weighting function $\alpha(u, v)$ as a simple scalar for any (u, v), which equivalently means to combine all the frequencies, including low frequencies and high frequencies, with the same weight α. Therefore, the controllability of the proposed frequency combination is better than that of linear combination considering more degrees of freedom brought by the frequency combination.

6 Conclusion

In this paper, we apply Fourier analysis to a unified framework of UST algorithms. We present the equivalent form of the framework and reveal the connections between the concepts of Fourier transform with those of style transfer. We give interpretations on the different performances between UST methods in structure preservation. We also present two operations for structure preservation and desired stylization. Extensive experiments are conducted to demonstrate (1) the equivalence between the framework and its proposed form, (2) the interpretability prompted by Fourier analysis upon style transfer and (3) the

controllability through manipulations on frequency components. Future work could further (1) expand work scope to include diverse artistic styles beyond WikiArt and (2) investigate model fine-tuning techniques, including leveraging Wasserstein distance.

Author contributions Zhiyu Jin and Xuli Shen mainly conducted experiments and wrote this manuscript. Bin Li guided the design of method and experiments. Xiangyang Xue provided suggestions for method improvement. All authors read and approved this manuscript.

Funding This work was supported in part by the National Natural Science Foundation of China (No. 62176061), STCSM project (No. 22511105000), UniDT's Cognitive Computing and Few Shot Learning Project, and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

Availability of data and materials All used data is publicly available.

Code availability The code will be available after this paper is accepted.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethical approval Not applicable.

Consent to participate Written informed consent was obtained from individual or guardian participants.

Consent for publication Not applicable.

References

Chen, D., Yuan, L., Liao, J., Yu, N., & Hua, G. (2017). Stylebank: An explicit representation for neural image style transfer. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2770–2779). https://doi.org/10.48550/arXiv.1703.09210.

Chen, H., Wang, Z., Zhang, H., Zuo, Z., Ailin Li, W. X., & Lu, D. (2021). Artistic style transfer with internal-external learning and contrastive learning. In *Advances in neural information processing systems*.

Chen, H., Zhao, L., Zhang, H., Wang, Z., Zuo, Z., Li, A., Xing, W., & Lu, D. (2021). Diverse image style transfer via invertible cross-space mapping. *In Proceedings of the international conference on computer vision (ICCV)*. https://doi.org/10.1109/ICCV48922.2021.01461

Chiu, T.-Y., & Gurari, D. (2022). Photowct 2: Compact autoencoder for photorealistic style transfer resulting from blockwise training and skip connections of high-frequency residuals. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)* (pp. 2978–2987). https://doi.org/10.1109/WACV51458.2022.00303.

Deng, Y., Tang, F., Dong, W., Huang, H., Ma, C., & Xu, C. (2021). Arbitrary video style transfer via multi-channel correlation. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, pp. 1210–1217). https://doi.org/10.48550/arXiv.2009.08003.

Dumoulin, V., Shlens, J., & Kudlur, M. (2017). Learned representation for artistic style. In *International conference on learning representations*. https://doi.org/10.48550/arXiv.1610.07629.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. https://doi.org/10.1109/CVPR.2016.265

Gonzalez, R. C., & Woods, R. E. (2008). *Digital image processing* (pp. 286–306). Prentice Hall, Upper Saddle River. http://www.amazon.com/Digital-Image-Processing-3rd-Edition/dp/013168728X.

Hong, K., Jeon, S., Yang, H., Fu, J., & Byun, H. (2021). Domain-aware universal style transfer. In *Proceedings of the international conference on computer vision (ICCV)* (pp. 14609–14617). https://doi.org/10.48550/arXiv.2108.04441.

Huang, X., & Belongie, S. (2017a). Arbitrary style transfer in real-time with adaptive instance normalization. In *International conference on computer vision (ICCV)* (pp. 1510–1519). https://doi.org/10.1109/ICCV.2017.167.

Huang, X., & Belongie, S. (2017b). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision* (pp. 1501–1510). https://doi.org/10.48550/arXiv.1703.06868.

Huo, J., Jin, S., Li, W., Wu, J., Lai, Y.-K., Shi, Y., & Gao, Y. (2021). Manifold alignment for semantically aligned style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 14861–14869). https://doi.org/10.48550/arXiv.2005.10777.

Jenkins, W. F., & Desai, M. D. (1986). The discrete frequency Fourier transform. *IEEE Transactions on Circuits and Systems, 33*, 732–734. https://doi.org/10.1109/TCS.1986.1085978

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision* (pp. 694–711). https://doi.org/10.48550/arXiv.1603.08155.

Li, C., & Wand, M. (2016). Combining Markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2479–2486). https://doi.org/10.48550/arXiv.1601.04589.

Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., & Yang, M.-H. (2017a). Diversified texture synthesis with feed-forward networks. In *IEEE conference on computer vision and pattern recognition (CVPR)*. https://doi.org/10.48550/arXiv.1703.01664.

Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., & Yang, M.-H. (2017b). Universal style transfer via feature transforms. In *Advances in neural information processing systems* (pp. 386–396). https://doi.org/10.48550/arXiv.1705.08086.

Li, Y., Liu, M.-Y., Li, X., Yang, M.-H., & Kautz, J. (2018). A closed-form solution to photorealistic image stylization. In *Proceedings of the European conference on computer vision (ECCV)*. https://doi.org/10.48550/arXiv.1802.06474.

Li, X., Liu, S., Kautz, J., & Yang, M.-H. (2019). Learning linear transformations for fast image and video style transfer. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 3804–3812). https://doi.org/10.1109/CVPR.2019.00393.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2014). Microsoft COCO: Common objects in context. arXiv:1405.0312.

Lu, M., Zhao, H., Yao, A., Chen, Y., Xu, F., & Zhang, L. (2019). A closed-form solution to universal style transfer. In *International conference on computer vision (ICCV)* (pp. 5951–5960). https://doi.org/10.48550/arXiv.1906.00668.

Nichol, K. (2016). *Painter by numbers* (Vol. 34). https://www.kaggle.com/c/painter-by-numbers.

Park, D. Y., & Lee, K. H. (2019). Arbitrary style transfer with style-attentional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 5880–5888). https://doi.org/10.48550/arXiv.1812.02342.

Sheng, L., Lin, Z., Shao, J., & Wang, X. (2018). Avatar-Net: Multi-scale Zero-shot style transfer by feature decoration. arXiv:1805.03857.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600–612. https://doi.org/10.1109/TIP.2003.819861

Wang, Z., Zhao, L., Chen, H., Qiu, L., Mo, Q., Lin, S., Xing, W., & Lu, D. (2020). Diversified arbitrary style transfer via deep feature perturbation. In *Proceedings of the IEEE international conference on computer vision* (pp. 7789–7798). https://doi.org/10.48550/arXiv.1909.08223.

Yoo, J., Uh, Y., Chun, S., Kang, B., & Ha, J.-W. (2019). Photorealistic style transfer via wavelet transforms. In: *International conference on computer vision (ICCV)* (pp. 9035–9044). https://doi.org/10.1109/ICCV.2019.00913.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *CoRR*. https://doi.org/10.48550/arXiv.1801.03924

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Zhiyu Jin[1] · Xuli Shen[1] · Bin Li[1] · Xiangyang Xue[1]

✉ Bin Li
libin@fudan.edu.cn

Zhiyu Jin
21210240058@m.fudan.edu.cn

Xuli Shen
xlshen20@fudan.edu.cn

Xiangyang Xue
xyxue@fudan.edu.cn

[1] School of Computer Science, Fudan University, 220 Handan Rd., Shanghai 200433, China