Evaluate and improve the quality of neural style transfer

Article *in* Computer Vision and Image Understanding · March 2021

DOI: 10.1016/j.cviu.2021.103203

7 authors, including:

Zhizhong Wang
Zhejiang University
13 PUBLICATIONS **66** CITATIONS
SEE PROFILE

Wei Xing
Shandong University of Technology
109 PUBLICATIONS **582** CITATIONS
SEE PROFILE

Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Evaluate and Improve the Quality of Neural Style Transfer

Zhizhong Wang, Lei Zhao**, Haibo Chen, Zhiwen Zuo, Ailin Li, Wei Xing**, Dongming Lu

Zhejiang University, Zhejiang, Hangzhou, 310027, People's Republic of China

ABSTRACT

Recent studies have made tremendous progress in neural style transfer (NST) and various methods have been advanced. However, evaluating and improving the stylization quality remain two important open challenges. Committed to these two aspects, in this paper, we first decompose the quality of style transfer into three quantifiable factors, i.e., the content fidelity (CF), global effects (GE) and local patterns (LP). Then, two novel approaches are further presented for exploiting these factors to improve the stylization quality. The first, named cascade style transfer (CST), utilizes the factors to guide the cascade combination of existing NST methods to absorb their merits and avoid their own shortcomings. The second, dubbed multi-objective network (MO-Net), directly optimizes these factors to balance their performance and achieves more harmonious stylized results. Extensive experiments demonstrate the effectiveness and superiority of our proposed factors and methods.

1. Introduction

Given the content and style images, the gist of style transfer is to synthesize an image that preserves some notion of the content but carries characteristics of the style. Recently, the seminal work of (Gatys et al., 2016b) firstly captured the style of artistic images and transferred it to other images using convolutional neural networks (CNNs). Since then, various neural style transfer (NST) methods have been advanced and obtained visually pleasing results (Jing et al., 2017).

Despite the recent rapid progress, how to assess the quality of style transfer has always been an open problem. To our knowledge, there are few quantitative protocols to evaluate it apart from user studies. This may be due to the fact that styles are highly subjective and subtle to define, hence such effective metrics are hard to choose (Yeh et al., 2018). However, the tedious process of user studies restrains the quick adjustment of method, and the results are highly influenced by the subjective preference of the participants. Thus, measuring the stylization quality in a quantitative way is anticipated, and it is bound to have an important impact on the future research.

On the other hand, a lot of valuable efforts have been devoted to improving the stylization quality, however, each of the existing methods has its own limitations. For instance, as shown in the 2^{nd} column of Fig. 1, Avatar-Net (Sheng et al., 2018) and LST (Li et al., 2019) can transfer the global colors and rough textures well and preserve the clear content structures, but fail to capture enough local patterns (e.g., droplets). In the 3^{rd} column, Reshuffle (Gu et al., 2018) and CNNMRF (Li and Wand, 2016a) can migrate more local patterns, but the global effects (e.g., color and saturation) are not satisfactory. Moreover, some methods (e.g., WCT (Li et al., 2017b) and Doodle (Champandard, 2016)) excessively mimic the style patterns, thus severely damaging the content structure (see the 4^{th} column). Through in-depth analysis, we believe that the fundamental cause of

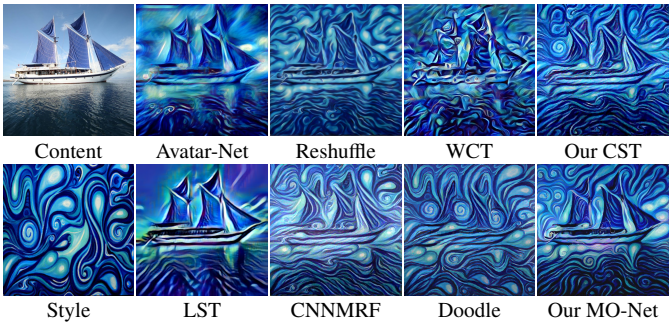

Fig. 1. Since the lack of a comprehensive consideration on stylization quality, existing NST methods often focus on limited aspects, while suffering from some problems on other aspects, such as deficient local patterns (2^{nd} column), insufficient global effects (3^{rd} column) and low content fidelity (4^{th} column). Our methods can avoid their own shortcomings and absorb all of their merits (last column).

**Corresponding author:
e-mail: cszhl@zju.edu.cn (Lei Zhao), wxing@zju.edu.cn (Wei Xing)

these problems is the lack of a comprehensive consideration on stylization quality, as some methods assume that the quality lies in global effects and content fidelity, while others consider it depends on the similarity of local patterns. Obviously, each of them has limitations. *Might there be a way to combine the strengths of both and avoid their own shortcomings?* The answer is yes, and this is what this paper is devoted to. Our results are shown in the last column of Fig. 1.

As the **first contribution** of this paper, we explore a quantitative way to comprehensively evaluate the quality of neural style transfer. Specifically, we first decompose the quality into three aspects, i.e., the content fidelity (CF), global effects (GE) and local patterns (LP). Then, for each aspect, we propose a quantifiable quality factor to measure its performance. Our proposed factors comprehensively take into account the structural retention of content, the transferred effects of global colors and textures, and the similarity and diversity of local style patterns, which cover the main aspects considered by different types of existing NST methods (Gatys et al., 2016b; Li and Wand, 2016a), and could better match the human perceptual judgments. These proposed factors can be served as a general quality benchmark for style transfer.

The **second contribution** of this work is designing two novel approaches for exploiting the aforementioned quality factors to improve the stylization quality. For the first approach, we directly benefit from existing methods, i.e., simply combine existing NST methods to absorb their merits. Concretely, we use a simple yet effective combination strategy which serially connects multiple NST methods by taking the output of the former as the input content of the latter, hence we name it *cascade style transfer (CST)*. The combination is guided by a novel *automatic quality ranker (AQR)*, which utilizes the proposed quality factors to assess the stylization quality of different NST methods, and follows some human observed heuristic knowledge to automatically determine the reasonable cascade order. For the second approach, we introduce a novel *multi-objective network (MO-Net)* by directly treating these quality factors as optimization objectives and optimizing them simultaneously. This multi-objective mechanism helps our *MO-Net* jointly optimize multiple objectives and balance their performance to achieve more harmonious stylized results. An *automatic parameter-tuning module* is also applied to remove the requirement for a human to manually adjust parameters.

To our best knowledge, our *CST* is the first to combine multiple existing NST approaches directly to improve the quality of style transfer. We conduct extensive qualitative and quantitative experiments to verify the effectiveness and superiority of our proposed factors and methods.

2. Related Work

Our proposed factors and methods can be related to most neural style transfer (NST) methods. Here we review some of the most representative ones. For a detailed review we refer to (Jing et al., 2017).

Gram-based Methods. The seminal work of Gatys et al. (Gatys et al., 2015b, 2016b, 2015a) opened up the era of NST, and provided a gold standard baseline for many subsequent methods (Johnson et al., 2016; Ulyanov et al., 2016; Dumoulin et al., 2017; Chen et al., 2017b; Wang et al., 2017; Huang and Belongie, 2017; Li et al., 2017b; Zhang and Dana, 2018; Shen et al., 2018; Li et al., 2019; Zhao et al., 2020; Lu et al., 2019; Park and Lee, 2019; Song et al., 2019). These methods use the correlations (i.e., Gram matrix) of the features extracted by convolutional neural networks (CNNs) to represent the style of an image, which could produce visually compelling results, especially for artistic styles. Based on them, several methods further improved on many other aspects, including diversity (Ulyanov et al., 2017; Li et al., 2017a; Wang et al., 2020b; Chen et al., 2020), photorealism (Luan et al., 2017; Li et al., 2018; Yoo et al., 2019), video (Chen et al., 2017a), stereoscopic 3D (Chen et al., 2018a), model compression (Wang et al., 2020a), user control (Gatys et al., 2017; Lu et al., 2017), and style classification (Chu and Wu, 2018).

Patch-based Methods. Another concurrent work is patch-based style transfer. (Li and Wand, 2016a,b) first combined Markov Random Fields (MRFs) and CNNs to match the most similar local neural patches of the style and content images. Following this line, many methods (Champandard, 2016; Chen and Schmidt, 2016; Liao et al., 2017; Gu et al., 2018; Mechrez et al., 2018; Sheng et al., 2018; Wang et al., 2020c; Yao et al., 2019; Kolkin et al., 2019; Zhang et al., 2019; Wang et al., 2021) are further proposed to improve the performance on quality and many other aspects.

Quantitative Evaluation Methods. Previous work mainly conduct user studies to evaluate the stylization quality, which may be influenced by the subjective preference of the participants. Recently, several methods began to use some more objective indicators for this purpose. (Li et al., 2018) used the similarity of the boundary maps to evaluate the content fidelity. (Yeh et al., 2018) evaluated the style effect and content perception based on the extent to which the style was transferred and the extent to which the transferred image decomposed into the content objects. Moreover, (Gu et al., 2018) and (Lu et al., 2019) directly exploited the content and style loss defined in (Gatys et al., 2016b) or (Li and Wand, 2016a) to evaluate the content and style quality. These evaluation methods either focus on the limited aspects of quality, or only apply to a certain type of style transfer, which cannot comprehensively match the human perceptual judgments in practice.

Unlike existing approaches, our quantitative evaluation is based on three quantifiable quality factors, which could comprehensively measure the quality of style transfer from different aspects (we also conduct user studies and compare with them in later Section 6.2). Guided by them, our CST seeks a common way to improve the stylization quality by simply combining existing NST methods. Moreover, our MO-Net further optimizes the quality by directly treating these quality factors as optimization objectives.

3. Quantitative Evaluation of Style Transfer

Quantitatively evaluating the quality of style transfer has long been an open problem in the community, and so far no

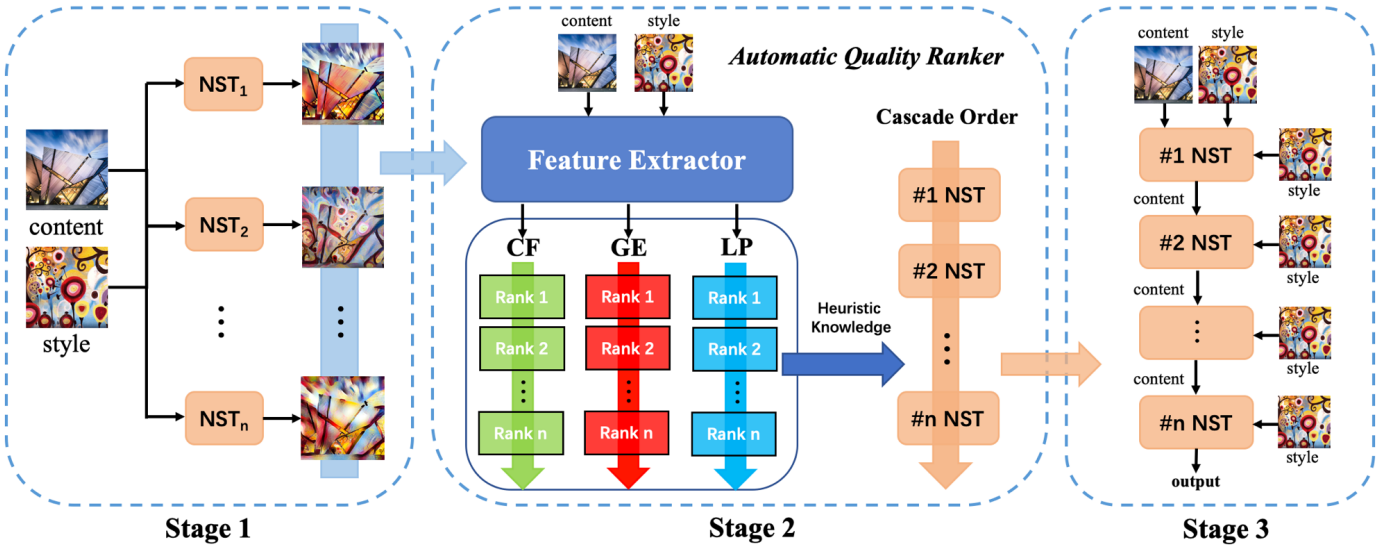

Fig. 2. Overview of our cascade style transfer (CST) method. Stage 1: Given a pair of inputs, generate stylized results by different NST methods. Stage 2: Rank these NST methods according to the scores of content fidelity (CF), global effects (GE) and local patterns (LP) by an *automatic quality ranker (AQR)*. Some heuristic knowledge is exploited to determine a reasonable cascade order. Stage 3: Combine these NST methods in the cascade order by using the stylized result of the former as the input content of the latter.

unified criterion except for user studies has been reached. Inspired by the observations of the main aspects considered by humans when they participate in user studies, we propose three quantifiable factors to measure these aspects for better matching the human perceptual judgments.

Content Fidelity (CF). An impressive style transfer result should carry enough characteristics of the style while preserving clearly perceived structure of the content. Profited from the high-level semantic features extracted by deep convolutional neural networks (Simonyan and Zisserman, 2014), we define a content fidelity (CF) factor to measure the faithfulness to content characteristics in multiple scales.

$$CF(\vec{x},\vec{c}) = \frac{1}{N}\sum_{l=1}^{N}\frac{f_l(\vec{x})\cdot f_l(\vec{c})}{\| f_l(\vec{x})\| \cdot \| f_l(\vec{c})\|} \tag{1}$$

where $\vec{c}$ and $\vec{x}$ are the content image and stylized result, respectively. $f_l(\cdot)$ denotes the deep feature activations extracted from layer l. N is the number of different layers. We adopt cosine similarity since it is independent of feature dimensions and numerical values of feature points, which can better measure the consistency of the two features in the holistic direction rather than the absolute difference in numerical values, as discussed in (Kolkin et al., 2019). Multi-scale feature similarities are computed to assess the content fidelity more comprehensively.

Global Effects (GE). When comparing the transferred style with the style image, people often appreciate the global effects first. The global similarity will leave a preliminary impression on humans' visual perception, so it is an important factor for measuring the stylization quality. By interviewing the people involved in our user studies, we learned that there are two main aspects that may affect their assessment on global effects, i.e., colors and textures. Therefore, here our global effect (GE) factor also includes these two aspects, namely, global colors (GC) and holistic textures (HT). For global colors, similar to (Gatys et al., 2016a), we also resort to the color information in RGB color space, but instead of matching the mean and covariance of the color pixels, we directly compare the cosine similarity of the color histograms:

$$GC(\vec{x},\vec{s}) = \frac{1}{3}\sum_{c=1}^{3}\frac{hist_c(\vec{x})\cdot hist_c(\vec{s})}{\| hist_c(\vec{x})\| \cdot \| hist_c(\vec{s})\|} \tag{2}$$

where $\vec{s}$ is the style image, $hist_c(\cdot)$ denotes the color histogram vector obtained in channel c.

For holistic textures, we utilize the style representation (Gram matrix) proposed by (Gatys et al., 2016b), which has been widely proved to be effective in expressing the holistic textures of the images.

$$HT(\vec{x},\vec{s}) = \frac{1}{N}\sum_{l=1}^{N}\frac{\mathcal{G}(f_l(\vec{x}))\cdot\mathcal{G}(f_l(\vec{s}))}{\| \mathcal{G}(f_l(\vec{x}))\| \cdot \| \mathcal{G}(f_l(\vec{s}))\|} \tag{3}$$

where $\mathcal{G}(\cdot)$ denotes the Gram matrix calculation. We also compute the multi-layer cosine similarities to better assess the holistic textures in multiple levels.

Finally, since these two aspects are considered equally important, our GE factor is simply the average of them:

$$GE(\vec{x},\vec{s}) = \frac{1}{2}(GC(\vec{x},\vec{s}) + HT(\vec{x},\vec{s})) \tag{4}$$

Local Patterns (LP). After roughly perceiving the global effects, people would zoom in to observe the similarity of some local style patterns, such as brush strokes, exquisite motifs, detailed textures, etc. Therefore, we define a local pattern (LP) factor to measure the quality of this aspect. Also inspired by the elements considered in human evaluation, our LP factor consists of two parts, one is to assess the similarity of the local pattern counterparts directly, and the other is to compare

the diversity of the retrieved pattern categories. We believe that a good stylized result should be similar to the style image not only in the corresponding local patterns, but also in the diversity of the retrieved pattern categories. Specifically, like those in patch-based methods (Li and Wand, 2016a; Chen and Schmidt, 2016), we first extract a set of 3×3 neural patches for multi-scale features $f_l(\vec{x})$ and $f_l(\vec{s})$, denoted by $\{\Phi_i^l(\vec{x})\}_{i\in n_x}$ and $\{\Phi_j^l(\vec{s})\}_{j\in n_s}$, where n_x and n_s are the numbers of extracted patches. For each patch $\Phi_i^l(\vec{x})$, we determine a closest-matching style patch $\Phi_{CM(i)}^l(\vec{s})$ based on the following normalized cross correlation measure:

$$CM(i) := \arg\max_{j=1,\ldots,n_s} \frac{\Phi_i^l(\vec{x}) \cdot \Phi_j^l(\vec{s})}{\| \Phi_i^l(\vec{x}) \| \cdot \| \Phi_j^l(\vec{s}) \|} \tag{5}$$

Then, we define the first part of our LP factor as follows:

$$LP_1(\vec{x}, \vec{s}) = \frac{1}{Z} \sum_{l=1}^{N} \sum_{i=1}^{n_x} \frac{\Phi_i^l(\vec{x}) \cdot \Phi_{CM(i)}^l(\vec{s})}{\| \Phi_i^l(\vec{x}) \| \cdot \| \Phi_{CM(i)}^l(\vec{s}) \|} \tag{6}$$

where $Z = N \times n_x$. This part measures the similarity of the corresponding local patterns of $\vec{x}$ and $\vec{s}$.

For the second part, we define it by comparing the categories of different neural patches in $\{\Phi_{CM(i)}^l(\vec{s})\}$ and $\{\Phi_j^l(\vec{s})\}$ directly.

$$LP_2(\vec{x}, \vec{s}) = \frac{1}{N} \sum_{l=1}^{N} \frac{t_{cm}^l}{t_s^l} \tag{7}$$

where t_{cm}^l and t_s^l are the numbers of different patches in $\{\Phi_{CM(i)}^l(\vec{s})\}$ and $\{\Phi_j^l(\vec{s})\}$, respectively.

Finally, since these two aspects are also considered equally important, our LP factor is simply the average of them:

$$LP(\vec{x}, \vec{s}) = \frac{1}{2}(LP_1(\vec{x}, \vec{s}) + LP_2(\vec{x}, \vec{s})) \tag{8}$$

Differences from Existing Methods. Basically, the proposed quality factors have borrowed some previous knowledge from the Gram-based (Gatys et al., 2016b) and patch-based (Li and Wand, 2016a) methods. However, there are three main differences between them: (1) First, the goal of our factors is to quantitatively evaluate the quality of style transfer, while existing methods are dedicated to achieving style transfer or improving the stylization quality. (2) Second, existing methods only consider the limited aspects of our factors, and usually measure the distances in Euclidean space, which highly depends on the numerical values of feature points. By contrast, our factors use the cosine scores to directly measure the similarity in the holistic direction, and is independent of feature dimensions and numerical values. (3) Third, we introduce a global color factor (Eq. 2) to measure the low-level similarity in RGB color space, and add a patch diversity factor (Eq. 7) to compute the diversity similarity of retrieved local pattern categories, which are two important aspects considered by humans but often not reflected in the existing methods. That is to say, our metrics cover the diverse aspects considered by previous methods and humans, which could help unify the quintessence and methodology in this field and consider the problem of style transfer more comprehensively.

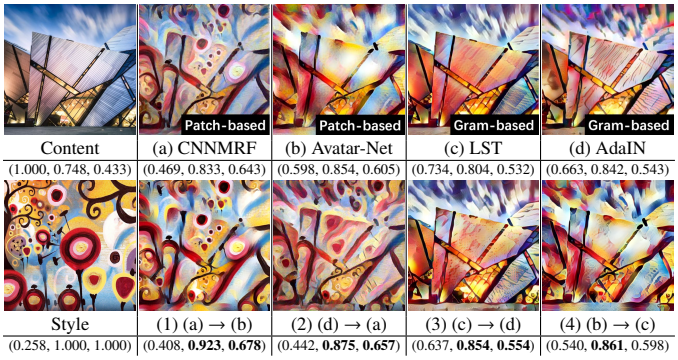

Fig. 3. Some examples of different cascade combinations (bottom row, 1-4). The original results of different NST methods are shown in the top row (a-d). We also put the type of each method in the lower right corner. For all results, we compute the quality scores of content fidelity (CF), global effects (GE) and local patterns (LP), and show the score triplet (CF, GE, LP) below each picture. For each combination, scores with promotions are marked in bold.

Quality Benchmark for Style Transfer. Since our proposed quality factors cover the comprehensive aspects considered by not only previous style transfer methods but also human beings, they can also be served as a general quality benchmark for newly proposed style transfer methods. Different from previous assessment methods (e.g., (Rosin et al., 2017) for portraits' non-photorealistic rendering assessment and (Aydın et al., 2014) for aesthetics judgment) based on sets of benchmark images or only low-level features such as sharpness, depth, clarity, tone and colorfulness, our factors consider not only low-level features (e.g., global colors and low-level textures), but also high-level features (e.g., semantic structures, high-level textures and local patterns) extracted from CNNs, thus providing a more thorough assessment on style transfer quality. Though the proposed factors are mainly based on the knowledge used in previous methods, they also apply to other style transfer methods which might improve the stylization in ways not reflected before, since each of our factors corresponds to a specific human interpretable physical meaning and involves a clear definition which has been proved effective and reasonable in many existing works. Nevertheless, adding other fine-grained factors such as brushstroke/fine-structure fidelity, clarity or contrast statistics may further enrich the evaluation, which is also what we plan to do in the future.

4. Cascade Style Transfer

We propose a common approach, namely, cascade style transfer (CST), for exploiting the aforementioned quality factors to guide the combination of existing NST methods, so as to further improve the quality of style transfer. The pipeline is depicted in Fig. 2, which involves the following three stages:

Stage 1: Employ different NST methods to generate stylized results for a given pair of content and style images.

Stage 2: An automatic quality ranker (AQR) is used to determine a reasonable cascade order for these NST methods. It first adopts a deep feature extractor to extract the features and evaluate the stylization quality, and then ranks these methods according to the similarity scores of content fidelity (CF), global

effects (GE) and local patterns (LP), as described in Section 3. To help give some useful guidance, we empirically summarize several heuristic knowledge by observing the results of a large number of different cascade combinations in advance. As style transfer is a highly subjective task, embedding this human observed heuristic knowledge would be necessary and beneficial. The knowledge is concluded into the following three rules, and some examples are shown in Fig. 3.

Rule (1): For CF, the score after cascade is usually lower than the individual score of each node method. That is to say, the content fidelity is a generally negative growth factor in the cascade process, and the upper bound of final CF score depends on the lowest one of all node methods.

Rule (2): For GE, the cascade score is usually increasing, provided the node methods do not have too low GE scores. That is to say, the global effect is a generally positive growth factor, and different methods can promote each other, even if the GE score of the latter is slightly lower than that of the former. However, the growth rate will decrease with the increase of the score.

Rule (3): For LP, the situation depends on the type of each node method. Patch-based methods often bring a positive gain regardless of their scores, while gram-based methods improve the performance only if they have higher LP scores.

Finally, these heuristic rules are utilized in the procedure of determining the reasonable cascade order, as depicted in Algorithm 1 (see the following paragraphs for analysis).

Stage 3: Following the obtained cascade order, users can combine different NST methods by using the stylized result of the former as the input content of the latter.

Analysis of CST:

(1) Explanations of Algorithm 1.

Since in the context of this work we aim to improve the quality of *artistic* style transfer, our CST regards the three factors as equally important. Therefore, the intention behind our Algorithm 1 is to determine the appropriate cascade order that maximizes the CF, GE and LP scores of the last synthesized image. Unfortunately, according to our observed heuristic knowledge (*Rule 1-3*), these three factors do not always promote each other. In fact, CF factor usually contradicts the other two due to the accumulative content loss caused by further stylization (*Rule 1*). And the mutual promotion between GE and LP factors is also affected by the orders and types of the cascaded methods (*Rule 2-3*). Therefore, *the key challenge is to balance the performance of CF, GE and LP factors in our CST*. Based on this analysis, our Algorithm 1 tries to achieve it from two aspects. The first is a warm-up preparation (step 2), we reorder the node methods according to their *average* ranking of CF, GE and LP factors, and prepare it as the initial cascade order. This warm-up operation, which ranks the methods from high to low according to the average performance, can help us make a comprehensive evaluation of each method in advance, and provide a better initial point for the following screening and cascading process. We will compare it with other schemes to prove its superiority in later Section 6.3. The second is a conditional screening process (step 3-13), we exploit the cascading characteristics of GE and LP factors in our heuristic knowledge (*Rule 2-3*) to select the methods that meet the promotion conditions and then add them to the final cascade list. This screening process can select the methods that truly benefit the cascade, and remove the methods that may damage the results, so as to optimize the stylization to the greatest extent.

Algorithm 1 Procedure to determine the cascade order.

1: Initialize an empty list O to store the cascade order;
2: Reassemble the node methods according to their average ranking of CF, GE and LP (if the average ranking is the same, the one with higher CF score is preferred), store the results in list R;
3: **for** $i = 1$ to $len(R)$ **do**
4: **if** $i == 1$ **then**
5: $O.append(R_i)$
6: **else if** $Type(R_i) == Patch\_based$ **then**
7: $O.append(R_i)$
8: **else if** $Type(R_i) == Gram\_based$ **then**
9: **if** $Rank_{LP}(R_i) < Rank_{LP}(R_{i-1})$ **then**
10: $O.append(R_i)$
11: **end if**
12: **end if**
13: **end for**

(2) How to avoid the accumulative content loss?

Intuitively, greater stylization effects (higher GE and LP scores) naturally lead to lower content retention (lower CF scores). This means accumulative content loss may be caused in our CST chain. To alleviate this problem, two strategies are adopted in our CST. The first is the original node method selection. Considering that there are a large number of NST methods available, selecting those methods that can generate results with higher CF scores as the nodes of our CST will directly benefit CST to better preserve content. The second is the reordering and screening process used in Algorithm 1. Reordering the original node methods according to the average ranking of CF, GE and LP factors can provide a good initial cascade order, and coupled with the subsequent screening process, CST can discard those methods that may severely harm the content or stylization effects, and only a few methods that really improve the performance are retained, thus greatly suppressing the accumulation of content loss.

(3) How many and which NST methods should be selected?

Explicitly, the number of methods in our CST can be arbitrary if the cost of time is not considered (we will analyze the efficiency in later Section 6.2). There is no specific requirement on the selection of them, because these methods will be automatically ranked by our AQR, and screened by our Algorithm 1. That is to say, the users do not need to try out different methods first to include in the CST, since the methods that may harm the final results would be discarded automatically by our Algorithm 1. Nonetheless, if users could filter the methods in advance (e.g., selecting those methods with higher average scores or a much higher score on a certain factor), the quality of the final results generated by CST could be further improved.

(4) In Fig. 3, why only explore the cascade combination of two methods?

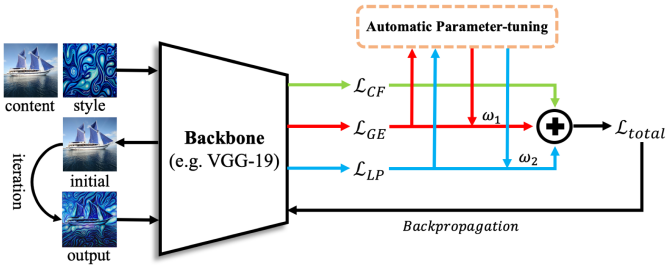

Fig. 4. Overview of our multi-objective network (MO-Net), which exploits an automatic parameter-tuning module to determine the parameters automatically.

The answer is no matter how many methods are cascaded, they can be decomposed into cascades between two methods. Therefore, we only observe the heuristic knowledge from the most basic cases that two methods are cascaded. This knowledge can be extended to cascade arbitrary number of methods, which will be verified in later Section 6, where our CST are used to cascade five NST methods.

5. Multi-objective Network

As mentioned in Section 1, an alternative way to take advantage of the proposed quality factors is directly treating them as optimization objectives. For this target, we further design a multi-objective network (MO-Net).

As shown in Fig. 4, our MO-Net follows the iterative optimization framework of (Gatys et al., 2016b), and our total loss function $\mathcal{L}_{total}$ is the simple linear combination of $\mathcal{L}_{CF}$, $\mathcal{L}_{GE}$ and $\mathcal{L}_{LP}$, which are obtained by replacing the cosine similarities in the corresponding quality factors (Eq. 1−8) with the squared-error losses.

However, how to balance the performance of these objectives remains an intractable problem. Previous methods such as (Gatys et al., 2016b) would often manually tune parameters on a *per-image* basis, which is quite tedious (Risser et al., 2017). To release this heavy burden for users, we embed an automatic parameter-tuning module to automatically determine the parameters for different content and style inputs.

The parameters here refer to the coefficients multiplied by the individual loss terms in our total loss. We dynamically adjust them based on the magnitude of the loss values. Concretely, since our generated image is initialized from the content image (we will discuss it in later Section 6.3), we default the parameter of $\mathcal{L}_{CF}$ to 1 and use the automatic parameter-tuning module to automatically determine the parameters of $\mathcal{L}_{GE}$ and $\mathcal{L}_{LP}$, as depicted in Algorithm 2. We execute it before each iteration of backpropagation. This dynamic and automatic process ensures that the values of $\mathcal{L}_{GE}$ and $\mathcal{L}_{LP}$ are remained at the similar order of magnitude, thus alleviating the imbalance problem in the optimization process.

Analysis of MO-Net:

(1) Explanations of Algorithm 2.

Similar to our CST, the focus of MO-Net is also the *artistic* style transfer, so we regard the three factors as equally important. *The key challenge here also lies in the balance of their performance.* Different from CST, in MO-Net, the performance of CF, GE and LP factors is determined by their corresponding loss terms. As acknowledged by (Risser et al., 2017), the optimization would be more mathematically well-founded if the non-gradient information such as the magnitude of the losses can be dynamically tuned. Therefore, the intention behind Algorithm 2 is to achieve this target. As we have mentioned previously, since we use the content image as the initialization, the parameter of $\mathcal{L}_{CF}$ is default to 1, and we only need to adjust the parameters of $\mathcal{L}_{GE}$ and $\mathcal{L}_{LP}$ to balance the performance of GE and LP factors. By fully observing a large number of manually adjusted optimal parameters, we empirically find that maintaining the values of $\mathcal{L}_{GE}$ and $\mathcal{L}_{LP}$ to the similar order of magnitude can achieve the most balanced performance. Hence, the parameters ω_1 and ω_2 of $\mathcal{L}_{GE}$ and $\mathcal{L}_{LP}$ are recurrently divided by 10 in our Algorithm 2. We dynamically execute it before each iteration of backpropagation. Based on this, the automatic parameter-tuning module can generalize MO-Net to arbitrary content-style image pairs, thus eliminating the tedious manual adjustment for different inputs.

Algorithm 2 Procedure of automatic parameter-tuning.

Initialize the parameter $\omega_1 = 1$ and $\omega_2 = 1$ for $\mathcal{L}_{GE}$ and $\mathcal{L}_{LP}$, respectively;
if $\mathcal{L}_{GE} > \mathcal{L}_{LP}$ **then**
  while $\omega_1\mathcal{L}_{GE} > \mathcal{L}_{LP}$ **do**
    $\omega_1 = \omega_1/10$
  end while
else
  while $\omega_2\mathcal{L}_{LP} > \mathcal{L}_{GE}$ **do**
    $\omega_2 = \omega_2/10$
  end while
end if

Note that the parameters here do not refer to the importance of the factors, because in this task we regard them as equally important. Actually, the importance of our factors can be reflected by the magnitude of each factor loss, therefore we can denote the loss magnitude ratio (e.g., $\frac{\mathcal{L}_{GE}}{\mathcal{L}_{LP}}$) as the factor importance ratio (which we will use in later Section 6.3). Obviously, there is a lot of room to manipulate the importance weights for our factors. As a result, this framework (as well as our CST) can be further applied to other tasks (e.g., photo-realistic style transfer) if appropriate factor importance weights are used (e.g., give greater importance weight to CF factor). We will demonstrate and discuss the importance trade-offs between our factors in later Section 6.3.

(2) Contributions of our MO-Net.

The main innovation of our MO-Net is the newly defined multiple objectives and the automatic parameter-tuning module, instead of the multi-objective framework itself (since it follows that of (Gatys et al., 2016b)). Its most important contribution is providing an optimization way to demonstrate what a unique role played by each of our quality factors, and how to exploit these factors to improve the stylization quality, which will be demonstrated in later Section 6.3.

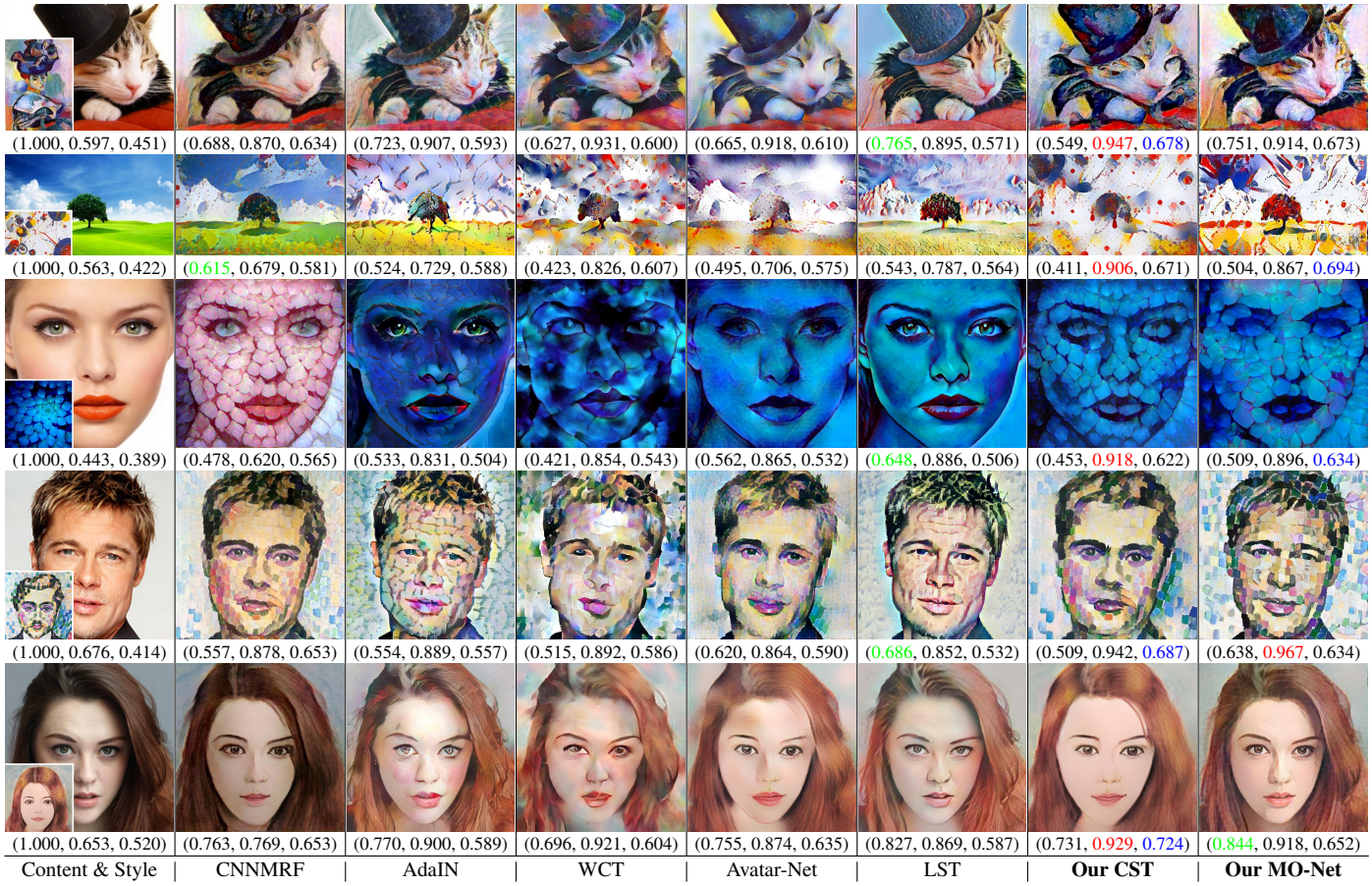

Fig. 5. Qualitative comparisons with five other representative NST methods. All results are obtained by running the author-released codes with default configurations. For all results (including the content images), we show the quality score triplet (CF, GE, LP) below each picture. For each group, the highest CF, GE and LP scores are marked in green, red and blue, respectively.

Table 1. Average stylization ranking scores of seven approaches.

Method	CNNMRF	AdaIN	WCT	Avatar-Net	LST	Our CST	Our MO-Net
Average Rank / User Study	5.286	4.254	3.912	3.819	4.072	**3.684**	**2.442**
Average Rank / Quality Scores	5.028	4.420	4.016	3.725	4.235	**3.551**	**2.580**

6. Experimental Results

6.1. Implementation Details of AQR

In AQR of our CST, we adopt VGG-19 (Simonyan and Zisserman, 2014) as the feature extractor since this model has been proved to be much more effective than others in extracting style information (Nikulin and Novak, 2016; Santurkar et al., 2019). Multi-scale features are first extracted from multiple layers $\ell \in \{Conv\mathbf{k}\_1\}_{\mathbf{k}=2}^{5}$. To evaluate the CF score, we utilize the features of $\{Conv\mathbf{k}\_1\}_{\mathbf{k}=4}^{5}$ as they mainly map the high-level semantic information. For GE and LP scores, we evaluate on the features of $\{Conv\mathbf{k}\_1\}_{\mathbf{k}=2}^{5}$ (we remove the lowest $Conv1\_1$ feature because we have evaluated the low level color information individually) and $\{Conv\mathbf{k}\_1\}_{\mathbf{k}=3}^{4}$ respectively, as suggested by (Gatys et al., 2016b) and (Li and Wand, 2016a).

6.2. Comparison with Prior Arts

To demonstrate the effectiveness and superiority of our methods, we compare them with five other representative NST methods (CNNMRF (Li and Wand, 2016a), AdaIN (Huang and Belongie, 2017), WCT (Li et al., 2017b), Avatar-Net (Sheng et al., 2018) and LST (Li et al., 2019)). For fair comparison, our CST is also used to cascade these methods.

Qualitative Evaluations. As shown in Fig. 5, the results of CNNMRF (Li and Wand, 2016a) can preserve complete local style patterns such as leaf-like motifs in the 3^{rd} row (higher LP scores), but the global effects (e.g., colors) are not plausible enough (lower GE scores). AdaIN (Huang and Belongie, 2017) and LST (Li et al., 2019) can transfer the global effects more faithfully (higher GE scores), but the results cannot manifest the exquisite local style details (lower LP scores). WCT (Li et al., 2017b) and Avatar-Net (Sheng et al., 2018) perform better in both global effects and local patterns, but may destroy some content structures or introduce hazy blocks (lower CF scores). By contrast, our CST can transfer adequate local style patterns while better preserving the overall feels of the exemplar styles, so it obtains much higher GE and LP scores than existing meth-

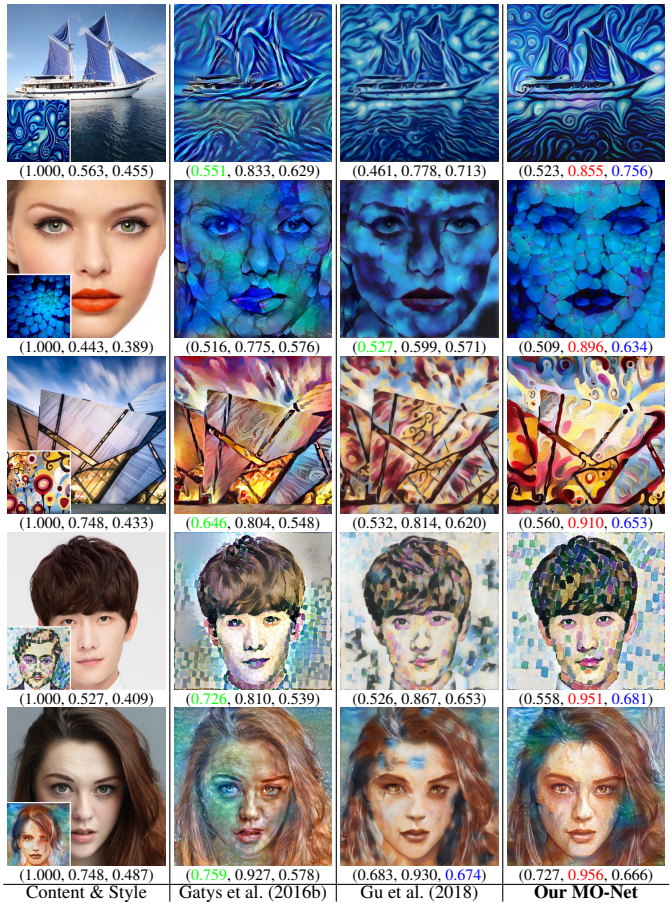

Fig. 6. Additional comparison results of our MO-Net and two other optimization-based NST methods (Gatys et al., 2016b; Gu et al., 2018). All results are obtained by running the author-released codes with default configurations. For all results (including the content images), we show the quality score triplet (CF, GE, LP) below each picture. For each group, the highest CF, GE and LP scores are marked in green, red and blue, respectively.

ods. Moreover, by simultaneously optimizing these three quality factors, our MO-Net well balances the performance of content fidelity, global effects and local patterns, which can be reflected by the higher average scores.

User Study vs. Quality Scores. For quantitative evaluation, a user study is also conducted and compared with our proposed quality scores. In user study, we use 100 groups of images and randomly select 30 groups for each subject. All seven results in each group are presented side-by-side and in a random order. Subjects are given unlimited time to rank the score from 1 to 7 (1 is the best, 7 is the worst) according to preference. We show the average ranking scores over 50 participants in the 2^{nd} row of Table 1. We also use our AQR to rank the same results provided to subjects, and show the average ranking scores in the 3^{rd} row (for each result, we compute its average ranking on the CF, GE and LP scores). Overall, the scores obtained from user study are close to those from our AQR, and subjects prefer our results more than others.

Actually, as we have discussed in Section 5, there is a lot of room to manipulate the importance weights for our quality factors. Since in this paper we focus on the *artistic* style transfer,

Table 2. Average stylization ranking scores of three approaches.

Method	(Gatys et al., 2016b)	(Gu et al., 2018)	Our MO-Net
A. R. / User Study	2.133	2.267	**1.600**
A. R. / Quality Scores	2.199	2.333	**1.467**

we empirically regard these factors as equally important and directly compute the average scores for comparison. *However, a more meaningful practice may be to explore the proper combination of these three factors that best fits the user study results, which we leave as a future work.*

Efficiency Analysis. With a 3.3 GHz hexa-core CPU, our AQR takes around 2 minutes to calculate the quality scores for a stylized image with 512×512 pixels. The speed of MO-Net is comparable to that of (Gatys et al., 2016b). These processes can be further accelerated, but currently they are beyond the scope of this work.

Preference between CST and MO-Net. Our CST and MO-Net provide two different ways to exploit the proposed factors to improve the quality of style transfer. It would be helpful to discuss the preference between them for further practical usage. As demonstrated in the previous experiments (Fig. 5), the results generated by our CST exhibit much more plausible stylization effects (higher GE and LP scores) but the content retention is relatively lower (lower CF scores). MO-Net achieves more balanced performance (higher average scores), but the stylization effects are not as vivid as CST. Therefore, if the users are pursuing higher stylization effects which do not have high demand for content maintenance, CST may be a more appropriate choice. Otherwise, using MO-Net can bring more satisfactory balanced results. On the other hand, in terms of efficiency, since CST involves some node method selection and automatic quality ranking processes, it may need more time and labor to generate exquisite results.

Additional Evaluation of MO-Net. Since our MO-Net follows the main framework of (Gatys et al., 2016b), and (Gu et al., 2018) also shares a similar idea which optimizes both parametric/Gram-based and non-parametric/patch-based losses, we make an additional comparison to them so as to demonstrate the superiority of our MO-Net and discuss their differences. The qualitative results are shown in Fig. 6, we can observe that though (Gatys et al., 2016b) achieves better content fidelity (higher CF scores) and moderate global effects (acceptable GE scores), the stylization is often unsatisfying due to the poor performance in local patterns (lower LP scores). By optimizing both parametric and non-parametric losses, (Gu et al., 2018) improves the performance of local patterns (higher LP scores), but the inconsistent color and saturation with the style image may reduce the GE scores (e.g., the top two images), and the CF scores may also be reduced due to the introduced content artifacts (e.g., the bottom three images). By contrast, our MO-Net well balances the performance of these three aspects by achieving much higher GE and LP scores as well as acceptable CF scores.

An additional user study vs. quality scores following the previous settings (but only 3 methods are ranked and 15 par-

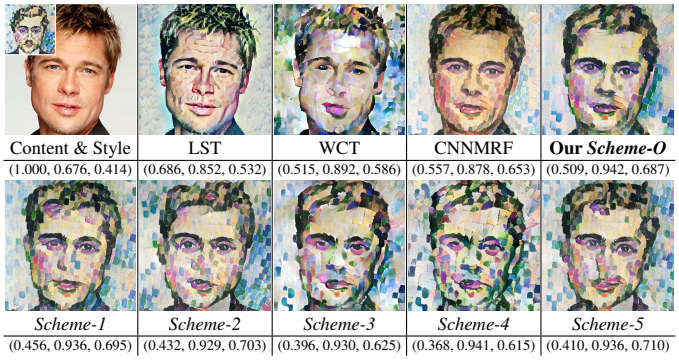

Fig. 7. Comparisons of different cascade schemes. The results of individual methods with the highest CF, GE or LP scores are shown in the 2^{nd}-4^{th} columns of the top row, respectively. We show the quality score triplet (CF, GE, LP) below each picture.

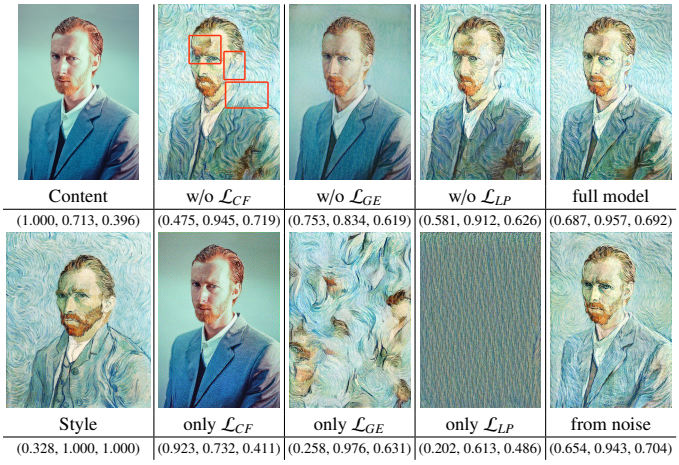

Fig. 8. Effect of different factors in MO-Net. The top row shows the results obtained by removing each factor separately in our default full model (initialized from the content image). The bottom row shows the results obtained by optimizing each factor individually to an extreme (initialized from white noise). Best viewed in conjunction with the quality scores (CF, GE, LP) below.

ticipants are recruited here) is also conducted to quantitatively evaluate these methods. The results shown in Table 2 verify the superiority of our MO-Net (note that the ranking scores of our MO-Net here is largely different from those in Table 1, because it is ranked among different approaches, and the number of approaches to be ranked is 3 vs. 7). The reasons of the superiority of our MO-Net may be (1) (Gatys et al., 2016b) regards the style transfer as a simple global statistical matching problem, which does not consider the local patterns, thus only transferring the global colors and rough textures. (2) (Gu et al., 2018) defines a non-differentiable reshuffle loss to transfer the local patterns under a global constraint restricting the usage time of each style patch, which does not explicitly optimize the global effects and the hard-determined usage parameter often causes less accurate matching, thus damaging the content structures. (3) Different from them, our MO-Net explicitly optimizes all considered aspects via differentiable losses and leverages a reasonable automatic parameter-tuning module, thereby well balancing the performance of CF, GE, and LP, and can generate much more satisfactory stylization results.

6.3. Ablation Study

Effect of Different Cascade Schemes. As analyzed earlier in Section 4, the key factor that influences the performance of our CST is the scheme to determine the cascade order. To verify the superiority of the proposed one (Algorithm 1, we denote it as *Scheme-O*), we compare it with five other schemes.

Scheme-1: Replace the warm-up (step 2) of *Scheme*-O with generating list R based on CF ranking only.

Scheme-2: Replace the warm-up (step 2) of *Scheme*-O with generating list R based on GE ranking only.

Scheme-3: Replace the warm-up (step 2) of *Scheme*-O with generating list R based on LP ranking only.

Scheme-4: Remove the conditional screening process (step 3-13) of *Scheme*-O and directly treat list R as O.

Scheme-5: Warm up *Scheme*-O with a low-to-high average ranking sequence by reversing list R in step 2.

The comparison results are shown in Fig. 7. We conduct experiments based on five NST methods used in the previous Section 6.2. As we can see in the 2^{nd}-4^{th} columns of the top row, the method that gets the highest score in one aspect (e.g., CF) often obtains lower scores in other aspects (e.g., GE and LP). With the proposed *Scheme-O*, our CST well balances the performance of these three aspects and produces much more stunning result (see the last column of the top row). It follows the main stream in list R, which is obtained based on the average ranking of CF, GE and LP scores (step 2 in Algorithm 1). This could provide a better initial point as all node methods have been ordered in advance according to their average performance before the next round of screening process. For comparison, we also show the results that generate R based on only one aspect (*Scheme-1* to *Scheme-3*). However, they all fail to balance the different aspects of performance as their results are often prone to have much lower CF scores (some contents are severely damaged, e.g., nose). This also occurs in *Scheme-4*, where we remove the constraints of our heuristic knowledge rules (step 3-13 in Algorithm 1) and directly treat list R as the final cascade order. It verifies the effectiveness and necessity of embedding the human observed heuristic knowledge. Moreover, we also try to warm up *Scheme-O* with a low-to-high average ranking sequence by reversing list R in step 2 (*Scheme-5*), unfortunately, the result shown in the last column of the bottom row still suffers from the content corruption problem, which may be attributed to the bad initial point provided by this way.

Effect of Different Factors in MO-Net. Our MO-Net contains three factors/objectives to optimize, i.e., $\mathcal{L}_{CF}$, $\mathcal{L}_{GE}$ and $\mathcal{L}_{LP}$. To demonstrate the role of each factor/objective, we conduct ablation experiments by removing them separately. As shown in the top row of Fig. 8, without $\mathcal{L}_{CF}$ (2^{nd} column), although the generated result still retains the main content structure (this is because we use the content image for initialization), it misses some content information or produces several unwanted artifacts (see red box area). In the 3^{rd} column, the global textures and colors are not fully transferred since the lack of $\mathcal{L}_{GE}$, while in the 4^{th} column, if there is no $\mathcal{L}_{LP}$, the resulting image cannot manifest the exquisite local style details (e.g., the details on the face).

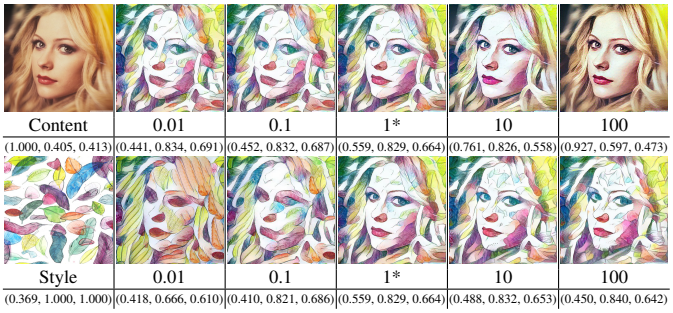

Fig. 9. Importance trade-offs between our factors in MO-Net. The top row shows the results obtained by varying the importance weight of CF factor. The bottom row shows the results obtained by changing the importance weight ratio of GE and LP factors. Our default setting is marked with *. Best viewed in conjunction with the quality scores (CF, GE, LP) below.

We also try to optimize each factor/objective individually to an extreme to help better understand their effects. The results are shown in the bottom row of Fig. 8, we initialize them with white noise. As we can observe, the CF and GE scores can be extremely optimized by minimizing our $\mathcal{L}_{CF}$ (2^{nd} column) and $\mathcal{L}_{GE}$ (3^{rd} column) only. However, there are still small gaps between them and the optimal ones (0.077 for CF score and 0.024 for GE score), which can be attributed to the performance bottleneck of the optimization algorithm (we use L-BFGS (Zhu et al., 1997)), and the lack of additional regularizers for smoothing the result, e.g., total variation loss (Johnson et al., 2016; Chen and Schmidt, 2016; Aly and Dubois, 2005). Interestingly, for LP score (4^{th} column), we observe that the result is not as expected. This is because $\mathcal{L}_{LP}$ is calculated based on the correspondence between the content and style neural patches. While in this case the content information is only from white noise, which may match a large number of repeated style patches, so the result is poor, and the diversity score LP_2 (Eq. 7) in our LP score is relatively low.

The last column in the bottom row shows the result obtained by balancing the performance of these three aspects for white noise initialization. We manually fine-tune the parameters to achieve the best quality. Compared to content initialization (the last column in the top row), the result obtained by this way exhibits more characteristics from the style image, e.g., the details on the face, which, unfortunately, may change the inferred identity of the person in the portrait. Besides, another superiority of content initialization is that it could help our MO-Net achieve a balanced performance more easily as we only need to determine the parameters of $\mathcal{L}_{GE}$ and $\mathcal{L}_{LP}$. All of these reasons together make us decide to use the content image as the default initialization.

Importance Trade-offs between Our Factors in MO-Net. To demonstrate how our factors trade off with each other, we vary the importance weights of them in our MO-Net. For CF factor, we change its importance weight by multiplying $\mathcal{L}_{CF}$ by a constant. The results shown in the top row of Fig 9 indicate that increasing the importance of CF can exhibit more content information (higher CF score), while eliminating more style details (lower GE and LP scores). For GE and LP factors, we use the automatic parameter-tuning module to ensure that the $\mathcal{L}_{GE}$ and $\mathcal{L}_{LP}$ meet the varied importance ratio ($\frac{\mathcal{L}_{GE}}{\mathcal{L}_{LP}}$). The results are shown in the bottom row of Fig 9. Interestingly, we find that GE and LP factors are not always mutually exclusive, sometimes increasing the importance weight of GE factor can also improve the performance of LP factor, e.g., the 2^{nd} and 3^{rd} columns in the bottom row. Therefore, the trade-off between these two aspects should be specially considered, just like what we have done in our MO-Net.

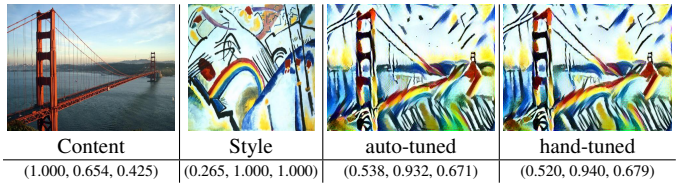

Fig. 10. Effect of automatic parameter-tuning. The result with auto-tuned parameters (3^{rd} column) is close to that with hand-tuned optimal parameters (4^{th} column).

Effect of Automatic Parameter-tuning. Fig. 10 demonstrates a comparison of our auto-tuned parameters with hand-tuned optimal parameters. As it shows, the two results are very close, both in terms of visual effects and quality scores (we test around 50 pairs of images). Note that our automatic parameter tuning module just provides an option for users to remove their manual adjustment or alleviate the burden of fine-tuning by offering near optimal parameters. Many other options based on metric learning or statistical normalization (Chen et al., 2018b) are also worth further exploration.

7. Conclusion and Future Directions

This paper proposes three quantifiable factors (i.e., the content fidelity/CF, global effects/GE and local patterns/LP) to evaluate the quality of style transfer and two novel approaches (i.e., the cascade style transfer/CST and multi-objective network/MO-Net) for exploiting these factors to improve the stylization quality. As we have discussed, there are lots of interesting directions worthy of further studies, such as adding other fine-grained factors to enrich the evaluation, manipulating the importance weights of our factors to best fit the user study results, extending our frameworks to other style transfer tasks, designing other effective automatic parameter-tuning strategy, accelerating the processes of CST and MO-Net, etc. In the future, our CST and MO-Net may provide some inspirations for the further employment or improvement of our quality factors, or more straightforwardly, these quality factors can be directly served as a general quality benchmark for style transfer.

Acknowledgments

This work was supported in part by the National Key R & D Plan Project (No: 2020YFC1523201, 2020YFC1523101, 2020YFC1522701), the Zhejiang Science and Technology Program (No: 2019C03137), Zhejiang Fund Project (No: LY19F020049), and the Cultural Relic Protection Science and Technology Project of Zhejiang Province (No: 2019008).

References

Aly, H.A., Dubois, E., 2005. Image up-sampling using total-variation regularization with a new observation model. IEEE Transactions on Image Processing (TIP) 14, 1647–1659.

Aydın, T.O., Smolic, A., Gross, M., 2014. Automated aesthetic analysis of photographic images. IEEE transactions on visualization and computer graphics 21, 31–42.

Champandard, A.J., 2016. Semantic style transfer and turning two-bit doodles into fine artworks. arXiv preprint arXiv:1603.01768 .

Chen, D., Liao, J., Yuan, L., Yu, N., Hua, G., 2017a. Coherent online video style transfer, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1105–1114.

Chen, D., Yuan, L., Liao, J., Yu, N., Hua, G., 2017b. Stylebank: An explicit representation for neural image style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1897–1906.

Chen, D., Yuan, L., Liao, J., Yu, N., Hua, G., 2018a. Stereoscopic neural style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6654–6663.

Chen, H., Zhao, L., Qiu, L., Wang, Z., Zhang, H., Xing, W., Lu, D., 2020. Creative and diverse artwork generation using adversarial networks. IET Computer Vision 14, 650–657.

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

Chen, Z., Badrinarayanan, V., Lee, C.Y., Rabinovich, A., 2018b. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks, in: International Conference on Machine Learning (ICML), pp. 794–803.

Chu, W.T., Wu, Y.L., 2018. Image style classification based on learnt deep correlation features. IEEE Transactions on Multimedia 20, 2491–2502.

Dumoulin, V., Shlens, J., Kudlur, M., 2017. A learned representation for artistic style, in: International Conference on Learning Representations (ICLR).

Gatys, L., Ecker, A.S., Bethge, M., 2015a. Texture synthesis using convolutional neural networks, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 262–270.

Gatys, L.A., Bethge, M., Hertzmann, A., Shechtman, E., 2016a. Preserving color in neural artistic style transfer. arXiv preprint arXiv:1606.05897 .

Gatys, L.A., Ecker, A.S., Bethge, M., 2015b. A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 .

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

Gatys, L.A., Ecker, A.S., Bethge, M., Hertzmann, A., Shechtman, E., 2017. Controlling perceptual factors in neural style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3985–3993.

Gu, S., Chen, C., Liao, J., Yuan, L., 2018. Arbitrary style transfer with deep feature reshuffle, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8222–8231.

Huang, X., Belongie, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1501–1510.

Jing, Y., Yang, Y., Feng, Z., Ye, J., Song, M., Jing, Y., Yang, Y., Feng, Z., Ye, J., Song, M., 2017. Neural style transfer: A review. arXiv preprint arXiv:1705.04058 .

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer. pp. 694–711.

Kolkin, N., Salavon, J., Shakhnarovich, G., 2019. Style transfer by relaxed optimal transport and self-similarity, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10051–10060.

Li, C., Wand, M., 2016a. Combining markov random fields and convolutional neural networks for image synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2479–2486.

Li, C., Wand, M., 2016b. Precomputed real-time texture synthesis with markovian generative adversarial networks, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer. pp. 702–716.

Li, X., Liu, S., Kautz, J., Yang, M.H., 2019. Learning linear transformations for fast image and video style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3809–3817.

Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H., 2017a. Diversified texture synthesis with feed-forward networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3920–3928.

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 (NeurIPS), pp. 386–396.

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), pp. 453–468.

Liao, J., Yao, Y., Yuan, L., Hua, G., Kang, S.B., 2017. Visual attribute transfer through deep image analogy. ACM Transactions on Graphics (TOG) .

Lu, M., Zhao, H., Yao, A., Chen, Y., Xu, F., Zhang, L., 2019. A closed-form solution to universal style transfer, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 5952–5961.

Lu, M., Zhao, H., Yao, A., Xu, F., Chen, Y., Zhang, L., 2017. Decoder network over lightweight reconstructed feature for fast semantic style transfer, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2469–2477.

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

Mechrez, R., Talmi, I., Zelnik-Manor, L., 2018. The contextual loss for image transformation with non-aligned data, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 768–783.

Nikulin, Y., Novak, R., 2016. Exploring the neural algorithm of artistic style. arXiv preprint arXiv:1602.07188 .

Park, D.Y., Lee, K.H., 2019. Arbitrary style transfer with style-attentional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5880–5888.

Risser, E., Wilmot, P., Barnes, C., 2017. Stable and controllable neural texture synthesis and style transfer using histogram losses. arXiv preprint arXiv:1701.08893 .

Rosin, P.L., Mould, D., Berger, I., Collomosse, J., Lai, Y.K., Li, C., Li, H., Shamir, A., Wand, M., Wang, T., et al., 2017. Benchmarking non-photorealistic rendering of portraits, in: Proceedings of the Symposium on Non-Photorealistic Animation and Rendering, pp. 1–12.

Santurkar, S., Ilyas, A., Tsipras, D., Engstrom, L., Tran, B., Madry, A., 2019. Image synthesis with a single (robust) classifier, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 1262–1273.

Shen, F., Yan, S., Zeng, G., 2018. Neural style transfer via meta networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8061–8069.

Sheng, L., Lin, Z., Shao, J., Wang, X., 2018. Avatar-net: Multi-scale zero-shot style transfer by feature decoration, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8242–8250.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Song, C., Wu, Z., Zhou, Y., Gong, M., Huang, H., 2019. Etnet: Error transition network for arbitrary style transfer, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 668–677.

Ulyanov, D., Lebedev, V., Vedaldi, A., Lempitsky, V.S., 2016. Texture networks: Feed-forward synthesis of textures and stylized images., in: International Conference on Machine Learning (ICML), pp. 1349–1357.

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

Wang, H., Li, Y., Wang, Y., Hu, H., Yang, M.H., 2020a. Collaborative distillation for ultra-resolution universal style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1860–1869.

Wang, X., Oxholm, G., Zhang, D., Wang, Y.F., 2017. Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5239–5247.

Wang, Z., Zhao, L., Chen, H., Qiu, L., Mo, Q., Lin, S., Xing, W., Lu, D., 2020b. Diversified arbitrary style transfer via deep feature perturbation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7789–7798.

Wang, Z., Zhao, L., Chen, H., Zuo, Z., Li, A., Xing, W., Lu, D., 2021. Diversified patch-based style transfer with shifted style normalization. arXiv preprint arXiv:2101.06381 .

Wang, Z., Zhao, L., Lin, S., Mo, Q., Zhang, H., Xing, W., Lu, D., 2020c.

Glstylenet: exquisite style transfer combining global and local pyramid features. IET Computer Vision 14, 575–586.

Yao, Y., Ren, J., Xie, X., Liu, W., Liu, Y.J., Wang, J., 2019. Attention-aware multi-stroke style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1467–1475.

Yeh, M.C., Tang, S., Bhattad, A., Forsyth, D.A., 2018. Quantitative evaluation of style transfer. arXiv preprint arXiv:1804.00118 .

Yoo, J., Uh, Y., Chun, S., Kang, B., Ha, J.W., 2019. Photorealistic style transfer via wavelet transforms, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 9036–9045.

Zhang, H., Dana, K., 2018. Multi-style generative network for real-time transfer, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer. pp. 349–365.

Zhang, Y., Fang, C., Wang, Y., Wang, Z., Lin, Z., Fu, Y., Yang, J., 2019. Multimodal style transfer via graph cuts, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 5943–5951.

Zhao, L., Mo, Q., Lin, S., Wang, Z., Zuo, Z., Chen, H., Xing, W., Lu, D., 2020. Uctgan: Diverse image inpainting based on unsupervised cross-space translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5741–5750.

Zhu, C., Byrd, R.H., Lu, P., Nocedal, J., 1997. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. ACM Transactions on Mathematical Software (TOMS) 23, 550–560.