

Drafting and Revision: Laplacian Pyramid Network for Fast High-Quality Artistic Style Transfer

Tianwei Lin¹, Zhuoqi Ma^{1,2}, Fu Li¹, Dongliang He¹, Xin Li¹, Errui Ding¹, Nannan Wang², Jie Li², Xinbo Gao³

Department of Computer Vision Technology (VIS), Baidu Inc.¹
Xidian University.² Chongqing University of Posts and Telecommunications.³

lintianwei01@baidu.com, zhuoqi_ma@hotmail.com

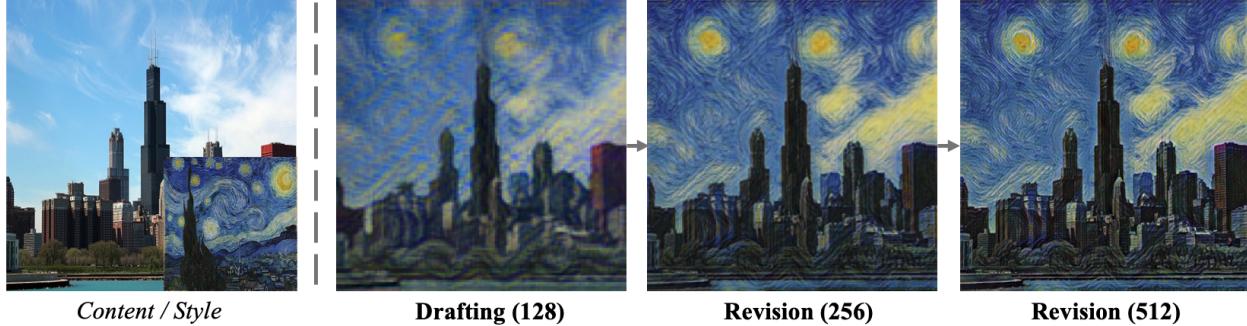


Figure 1. Illustration of our proposed style transfer process. First we transfer global patterns in low resolution, then revise local patterns in high resolution. For better visualization, we resize stylized images of different scale into the same size. Zoom in to have a better view.

Abstract

Artistic style transfer aims at migrating the style from an example image to a content image. Currently, optimization-based methods have achieved great stylization quality, but expensive time cost restricts their practical applications. Meanwhile, feed-forward methods still fail to synthesize complex style, especially when holistic global and local patterns exist. Inspired by the common painting process of drawing a draft and revising the details, we introduce a novel feed-forward method named **Laplacian Pyramid Network (LapStyle)**. LapStyle first transfers global style patterns in low-resolution via a Drafting Network. It then revises the local details in high-resolution via a Revision Network, which hallucinates a residual image according to the draft and the image textures extracted by Laplacian filtering. Higher resolution details can be easily generated by stacking Revision Networks with multiple Laplacian pyramid levels. The final stylized image is obtained by aggregating outputs of all pyramid levels. Experiments demonstrate that our method can synthesize high quality stylized images in real time, where holistic style patterns are properly transferred. Codes will

be released on [PaddleGAN](#).

1. Introduction

Artistic style transfer is an attractive technique which can create an art image with the structure of a content image and the style patterns of an example style image. It has been a prevalent research topic for both academy and industry. Recently, there have been a lot of methods proposed for neural style transfer, which can be roughly divided into two types: image-optimization and model-optimization methods.

Image-optimization methods iteratively optimize stylized image with fixed network. The seminal work of Gatys *et al.* [8] achieves style transfer in an iterative optimization process, where the style patterns are captured by correlation of features extracted from a pre-trained deep neural network. Following works improve [8] mainly in the form of different loss functions [15, 27]. Although superior stylization results are achieved, *e.g.*, STROTSS [15], widespread applications of these methods are still restricted by their slow online optimization process. On the contrary, model-optimization methods update neural networks by training and are feed-forward in testing. There are mainly three subdivided types:

(1) *Per-Style-Per-Model* methods [13, 18, 32, 33, 34] are trained to synthesize images with a single given style image; (2) *Multi-Style-Per-Model* methods [3, 7, 35, 20, 37] introduce various network architectures to simultaneously handle multiple styles; (3) *Arbitrary-Style-Per-Model* methods [12, 21, 30, 19, 25] further adopt diverse feature modification mechanisms to transfer arbitrary styles. Reviewing these methods, we find that although local style patterns can be transferred, complex style mixed with both global and local patterns is still not properly transferred. Meanwhile, artifacts and flaws appear in many cases. To this end, in this work, our main goal is to achieve superior high-quality artistic style transfer results with feed-forward network, where local and global patterns can be reserved aesthetically.

How human painters handle the complex style patterns while painting? A common process, especially for a beginner, is to first draw a draft to capture global structure and then revise the local details gradually, instead of directly finishing the final painting part-by-part. Inspired by this, we propose a novel neural network named **Laplacian Pyramid Network (LapStyle)** for style transfer. Firstly, in our framework, a *Drafting Network* is designed to transfer global style patterns in low-resolution, since we observe that global patterns can be transferred easier in low resolution due to larger receptive field and less local details. A *Revision Network* is then used to revise the local details in high-resolution via hallucinating a residual image according to the draft and the textures extracted by Laplacian filtering over the $2 \times$ resolution content image. Note that our Revision Network can be stacked in a pyramid manner to generate higher resolution details. The final stylized image is obtained by aggregating outputs of all pyramid levels. Further, we adopt shallow patch discriminators to adversarially learn local style patterns. As illustrated in Fig. 1, appealing stylization results are achieved by our “Drafting and Revision” process. To summarize, the main contributions are as follows:

- We introduce a novel framework “Drafting and Revision”, which simulates painting creation mechanism by splitting style transfer process into global style pattern drafting and local style pattern revision.
- We propose a novel feed-forward style transfer method named LapStyle. It uses a Drafting Network to transfer global style patterns in low-resolution, and adopts higher resolution Revision Networks to revise local style patterns in a pyramid manner according to outputs of multi-level Laplacian filtering of the content image.
- Experiments demonstrate that our method can generate high-resolution and high-quality stylization results, where global and local style patterns are both effectively synthesized. Besides, the proposed LapStyle is extremely efficient and can synthesize high resolution stylized image of 512 pix in 110 fps.

2. Related Work

Style Transfer. Style transfer algorithms aim at migrating the style from an example image to a content image. With the initiation of the seminal work Gatys *et al.* [8], various methods have been developed thereafter to address different aspects of, including visual quality [17, 10], head portrait [28], semantic control [1, 9] and so on. Kolkin *et al.* propose STROTSS [15] and higher quality stylized images can be generated by adopting Earth Movers Distance (rEMD) loss for optimization, which deploys the style attributes with minimum distortion to content’s semantic layout. However, the expensive computational cost of optimization-based methods hinder their practical applications. In order to improve run-time efficiency, researchers have proposed to replace the iterative optimization procedure with feed-forward networks. *Per-Style-Per-Model* methods [13, 18, 32, 33, 34, 5] adopt auto-encoder as style transfer network trained with variants of content and style losses derived from [8]. *Multi-Style-Per-Model* methods [3, 7, 35, 20, 37] embed learnable affine transformation architectures in the middle of auto-encoder to incorporate multiple styles. Recently, *Arbitrary-Style-Per-Model* methods [12, 21, 30, 19, 25] achieve arbitrary style transfer via style feature embedding networks. Besides, video style transfer methods [2, 11, 4] exploit how to generate stable stylization video.

Model-optimization based methods greatly improve computation efficiency with visual quality compromises. AdaIN [12], WCT [21] and linear transformation [19] adjust holistic feature distributions so they all fail to preserve local style patterns. SANet [25] embed local style patterns in content feature map with the aid of style attention mechanism, but they cannot perform well with large-scale textures such as the swirls in *The Starry Night*. On the contrary, our proposed LapStyle can capture the style statistics at different scales, which greatly improves the visual quality over current model-optimization based methods.

Multi-scale Learning. In image manipulation area, working at multiple scales is a common technique to better capture a wide range of image statistics [6, 16, 29, 31, 23, 35, 15]. Lai *et al.* propose LapsRN [16] to progressively reconstruct the high-resolution images by predicting high-frequency residuals with cascaded convolutional networks. Shaham *et al.* propose SinGAN [29] to train the network with single image by capturing patch-level distribution at different image scales with a pyramid of adversarial networks. WCT [21] and PhotoWCT [22] also generate results coarse to fine gradually, but they work at the original RGB domain and not explicitly revise stylized details in the residual field as LapStyle does. WCT² [36] also exploits residual information via wavelet transform where the residual information is mainly used to keep spatial details of original image. Differently, LapStyle constructs the Revision Network in the residual field to better transfer and enhance local stylization details.

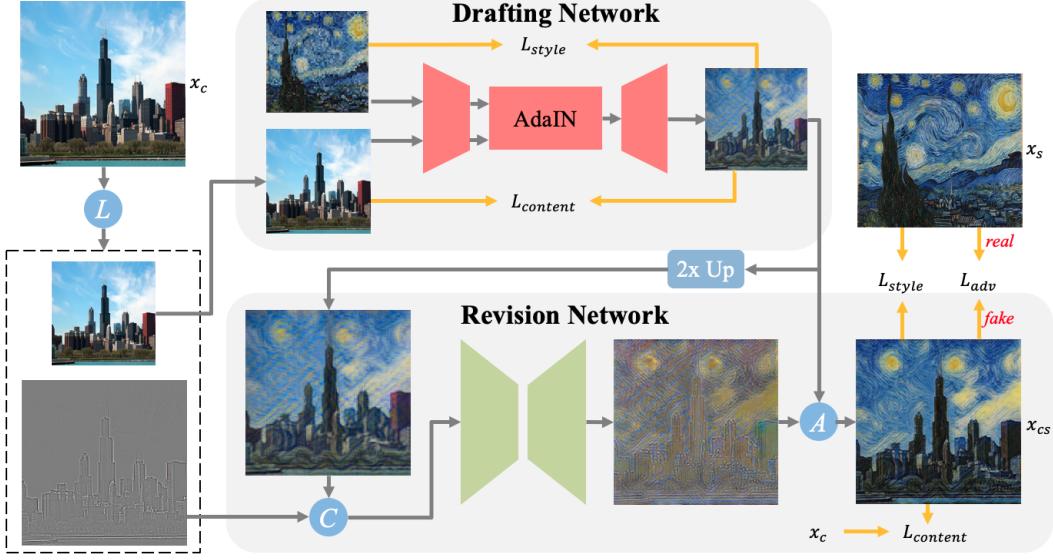


Figure 2. Overview of our framework. (a) We first generate image pyramid $\{\bar{x}_c, r_c\}$ from content image x_c with the help of Laplacian filter. (b) Rough low-resolution stylized image is generated by the Drafting Network. (c) Then the Revision Network generates stylized detail image in high resolution. (4) Final stylized image is generated by aggregating the outputs pyramid. L , C and A in image represent Laplacian, concatenate and aggregation operation separately.

STROTSS [15] also adopts a multi-scale scheme to apply style transfer by minimizing EMD loss at increasing resolution and exhibits high visual quality. However, the iterative optimization procedure suffers high computation cost and needs several minutes to synthesize one image. Our proposed LapStyle captures a wide range of style statistics from global distribution to local patterns by adopting a multi-scale network to better balance the trade-off between run-time efficiency and visual quality.

3. Approach

In this section, we will introduce the proposed feed-forward style transfer network LapStyle in detail. For ease of understanding, in this section, we only describe the framework with a 2-level pyramid. The base level is a Drafting Network and a Revision Network is used for the 2nd level of higher resolution, as shown in Fig. 2. It is quite straightforward to build more levels by stacking Revision Networks.

3.1. Network Architecture

Our proposed LapStyle takes a content image $x_c \in R^{H_c \times W_c}$ and a pre-defined style image x_s as inputs, and eventually synthesizes a stylized image x_{cs} . As shown in Fig. 2, for pre-processing, we construct a 2-level image pyramid $\{\bar{x}_c, r_c\}$. \bar{x}_c is simply a $2\times$ downsampled version of x_c . r_c is obtained with the help of Laplacian filter, i.e., $r_c = x_c - Up(\mathcal{L}(\bar{x}_c))$, where \mathcal{L} denotes Laplacian filtering and Up is $2\times$ upsample operation. The style image x_s is

also downsampled to a low-resolution version \bar{x}_s .

In the first stage, the *Drafting Network* first encodes context and style features from both \bar{x}_c and \bar{x}_s with a pre-trained neural network, then it modulates content feature using style feature in multiple granularities and finally generates the stylized image $\bar{x}_{cs} \in R^{H_c/2 \cdot W_c/2}$ using a decoder. In the second stage, the *Revision Network* first up-samples \bar{x}_{cs} to $x'_{cs} \in R^{H_c \cdot W_c}$, then it concatenates x'_{cs} and r_c as the input to generate stylized a detail image $r_{cs} \in R^{H_c \cdot W_c}$. Finally, we obtain stylized image $x_{cs} \in R^{H_c \cdot W_c}$ by aggregating the pyramid outputs:

$$x_{cs} = \mathcal{A}(\bar{x}_{cs}, r_{cs}), \quad (1)$$

where \mathcal{A} denotes the aggregation function. In the following, we will introduce the configuration of Drafting Network and Revision Network in detail.

3.2. Drafting Network

The Drafting Network aims at synthesizing global style patterns in low resolution. Why low resolution? As demonstrated in Section 4.3, we observe that global patterns can be transferred easier in low resolution, due to large receptive field and less local details. To achieve single style transfer, earlier work [13] directly trains an encoder-decoder module, where only content image is used as input. To better combine the style feature and the content feature, we adopt AdaIN module from recent arbitrary style transfer method [12].

The architecture of Drafting Network is shown in Fig

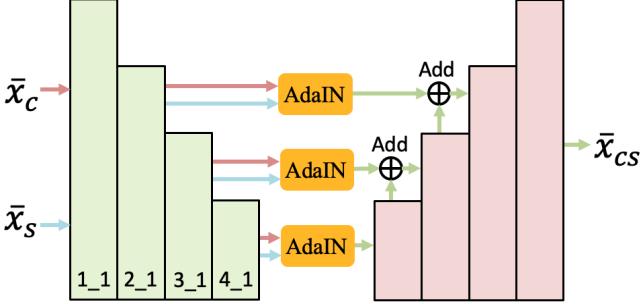


Figure 3. Illustration of the proposed Drafting Network.

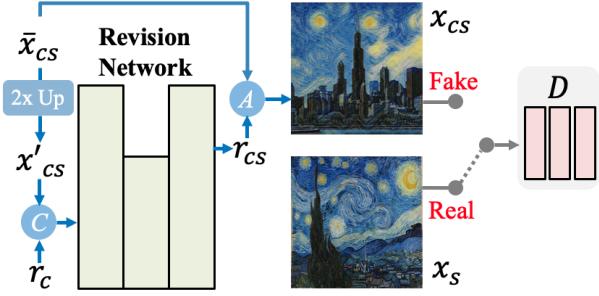


Figure 4. Illustration of the proposed Revision Network. C and A here represent concatenate and aggregation operation separately.

3, which includes an encoder, several AdaIN modules and a decoder. **(1)** The encoder is a pre-trained VGG-19 network, which is fixed during training. Given \bar{x}_c and \bar{x}_s , the VGG encoder extracts features in multiple granularity at 2_1, 3_1 and 4_1 layers. **(2)** Then, we apply feature modulation between the content and style feature using AdaIN modules after 2_1, 3_1 and 4_1 layers, respectively. **(3)** Finally, in each granularity of decoder, the corresponding feature from the AdaIN module is merged via a skip-connection. Here, skip-connections after AdaIN modules in both low and high levels are leveraged to help to reserve content structure, especially for low-resolution image.

3.3. Revision Network

The Revision Network aims to revise the rough stylized image via generating residual details image r_{cs} , while the final stylized image is generated by combining r_{cs} and rough stylized image \bar{x}_{cs} . This procedure ensures that the distribution of global style pattern in \bar{x}_{cs} is properly kept. Meanwhile, learning to revise local style patterns with residual details image is easier for the Revision Network.

As shown in Fig. 4, the Revision Network is designed as a simple yet effective encoder-decoder architecture, with only one down-sampling and one up-sampling layer. Further, we introduce a patch discriminator to help Revision Network to capture fine patch textures under adversarial learning setting.

We define the patch discriminator D following SinGAN [29], where D owns 5 convolution layers and 32 hidden channels. We choose to define a relatively shallow D to (1) avoid overfitting since we only have one style image and (2) control the receptive field to ensure D can only capture local patterns.

3.4. Training

During training, the Drafting Network and the Revision Network are both optimized with content and style loss, while the Revision Network further adopts adversarial loss. Thus, we first describe style and content losses, then introduce the full objective of two networks separately. Since our LapStyle is “Per-Style-Per-Model”, during training, we keep a single x_s , and a set of x_c from content dataset X_c .

Style Loss. Following recent optimization-based method STROTSS [15], we combine the relaxed Earth Mover Distance (rEMD) loss and the commonly used mean-variance loss as style loss. To begin with, given an image, we can use pre-trained VGG-19 encoder to extract a set of feature vectors as $\mathcal{F} = \{F^{1,1}, F^{2,1}, F^{3,1}, F^{4,1}, F^{5,1}\}$. The rEMD loss aims at measuring the distance between the feature distributions of style image x_s and stylized image x_{cs} . For simplicity, we omit the superscripts which indicate layer index in the following. Supposing $F_s \in R^{h_s w_s \times c}$, $F_{cs} \in R^{h_{cs} w_{cs} \times c}$ are the features of x_s and x_{cs} , their rEMD loss can be calculated as:

$$l_r = \max \left(\frac{1}{h_s w_s} \sum_{i=1}^{h_s w_s} \min_j C_{ij}, \frac{1}{h_{cs} w_{cs}} \sum_{j=1}^{h_{cs} w_{cs}} \min_i C_{ij} \right), \quad (2)$$

where the cosine distance term C_{ij} is defined as:

$$C_{ij} = 1 - \frac{F_{s,i} \cdot F_{cs,j}}{\|F_{s,i}\| \|F_{cs,j}\|} \quad (3)$$

To keep the magnitude of the feature vectors, we also adopt the commonly used mean-variance loss as:

$$l_m = \|\mu(F_s) - \mu(F_{cs})\|_2 + \|\sigma(F_s) - \sigma(F_{cs})\|_2, \quad (4)$$

where μ and σ calculate the mean and co-variance of the feature vectors separately.

Content Loss. For content loss, we adopt the commonly used normalized perceptual loss and the self similarity loss between $F_c \in R^{h_c w_c \times c}$ and $F_{cs} \in R^{h_{cs} w_{cs} \times c}$ proposed in [15]. Note that h_{cs} equals h_c and w_c equals w_{cs} because of x_c and x_{cs} are of the same resolution. The perceptual loss is defined as:

$$l_p = \|norm(F_c) - norm(F_{cs})\|_2, \quad (5)$$

where $norm$ denotes the channel-wise normalization for F . The self-similarity loss aims to retain the relative relation in content image to stylized image, which is defined as:

$$l_{ss} = \frac{1}{(h_c w_c)^2} \sum_{i,j} \left| \frac{D_{ij}^c}{\sum_i D_{ij}^c} - \frac{D_{ij}^{cs}}{\sum_i D_{ij}^{cs}} \right|, \quad (6)$$

here D_{ij}^c and D_{ij}^{cs} are the $(i,j)^{th}$ entry of self-similarity matrices D^c and D^{cs} , respectively. Here D_{ij} is pairwise cosine similarity $\langle F_i, F_j \rangle$.

Loss of Drafting Network. In training phase of Drafting Network, low resolution images \bar{x}_c and \bar{x}_s are used as the network input, and also are used to calculate the content loss and style loss separately. The overall training objective function of the Drafting Network is defined as:

$$L_{Draft} = (l_p + \lambda_1 \cdot l_{ss}) + \alpha \cdot (l_m + \lambda_2 \cdot l_r) \quad (7)$$

where λ_1 , λ_2 and α are weight terms. We control the balance of content and style loss via adjusting α . Specifically, l_r and l_{ss} are defined on 3_1 and 4_1 layers, meanwhile l_m and l_p are defined from 1_1 to 5_1 layers.

Loss of Revision Network. In training phase of Revision Network, the parameters of Drafting Network are fixed and the training loss is built upon x_{cs} . To better learn local fine-grain textures, except for base content and style loss $L_{base} = L_{Draft}$, we introduce a discriminator and train Revision Network with an adversarial loss term. The overall optimization objective is defined as:

$$\min_{Rev} L_{base} + \beta \cdot \min_{Rev} \max_D L_{adv}(Rev, D), \quad (8)$$

where Rev denotes the Revision Network, D denotes the discriminator, and β controls the balance between base style transfer losses and adversarial loss. L_{adv} is standard adversarial training loss.

4. Results

4.1. Dataset and Setup

Dataset. To train our model, we need a single style image and a collection of content images. In this work, following conventions, we use MS-COCO [24] as content images and select style images from WikiArt [26]. Some copyright free images from *Pexels.com* is also used as style images.

Implementation Details. In LapStyle, Drafting and Revision Networks are trained in sequence. The former one is first trained with resolution of 128×128 , then the later one is trained with 256×256 . To achieve higher resolution, we can consecutively train more Revision Networks using resolution of 512 and 1024. For both networks, we use the Adam optimizer [14] with a learning rate of 1e-4 and a batch size of 5 content images. For both networks, the training process consists of 30,000 iterations. The loss weight term, λ_1 , λ_2 , α and β are set to 16, 3, 3 and 1, respectively. More detailed network configurations of LapStyle is presented in our supplementary material.

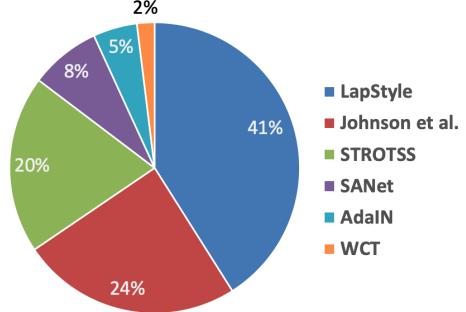


Figure 5. User preference results of five SOTA methods.

Method	Time (256pix)	Time(512pix)
Gatys et al. [8]	15.863	50.804
STROTSS [15]	163.052	177.485
Johnson et al. [13]	0.132	0.149
WCT [21]	0.689	0.997
AdaIN [12]	0.011	0.039
Linear [19]	0.007	0.039
SANet [25]	0.017	0.055
Ours	0.008	0.009

Table 1. Execution time comparison (in seconds).

4.2. Comparison with Prior Works

Qualitative Comparison. As shown in Fig. 6, we compare our method with state-of-the-art feed-forward methods. Like our LapStyle, Johnson et al. [13] is also a single style transfer method. [13] can synthesize some local style patterns with clear structure (e.g. row 8), however, the color distributions and texture structures of content image are often maintained (e.g. rows 2 and 8), resulting in insufficient stylization. AdaIN [12], WCT [21] and SANet [25] are arbitrary style transfer models, which have some common features: (1) they mainly transfer the color distribution and simple local patterns of style image; (2) the complex style patterns with bigger size is basically not transferred (e.g. rows 2, 5 and 6); (3) the local patterns are usually not accurately transferred, resulting in messy local textures (e.g. rows 2, 8). To be specific, the problem of retaining color distribution of content is severe in AdaIN [12] (e.g. rows 2, 7 and 8). WCT [21] discards too much context structure, resulting in disordered and chaotic image. In contrast to these methods, our method can simultaneously transfer simple local style patterns and complex global style patterns, retaining clear and clean structure of style patterns. The color distribution is also completely transferred. Although LapStyle can not transfer arbitrary style, we believe that improving the stylization quality is most important for feed-forward method. We leave arbitrary LapStyle for future work.

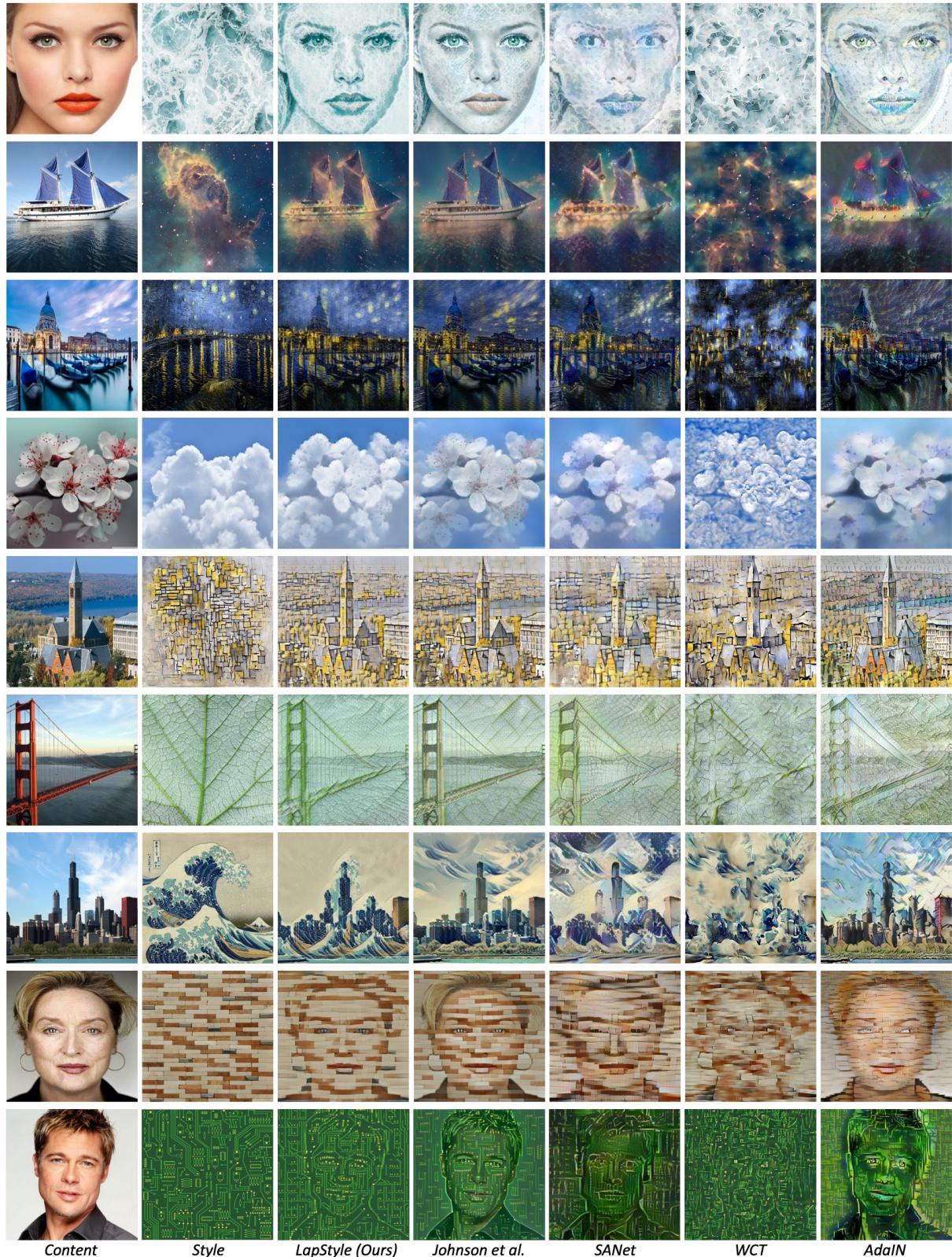


Figure 6. Qualitative comparisons between our method and state-of-the-art feed forward methods.

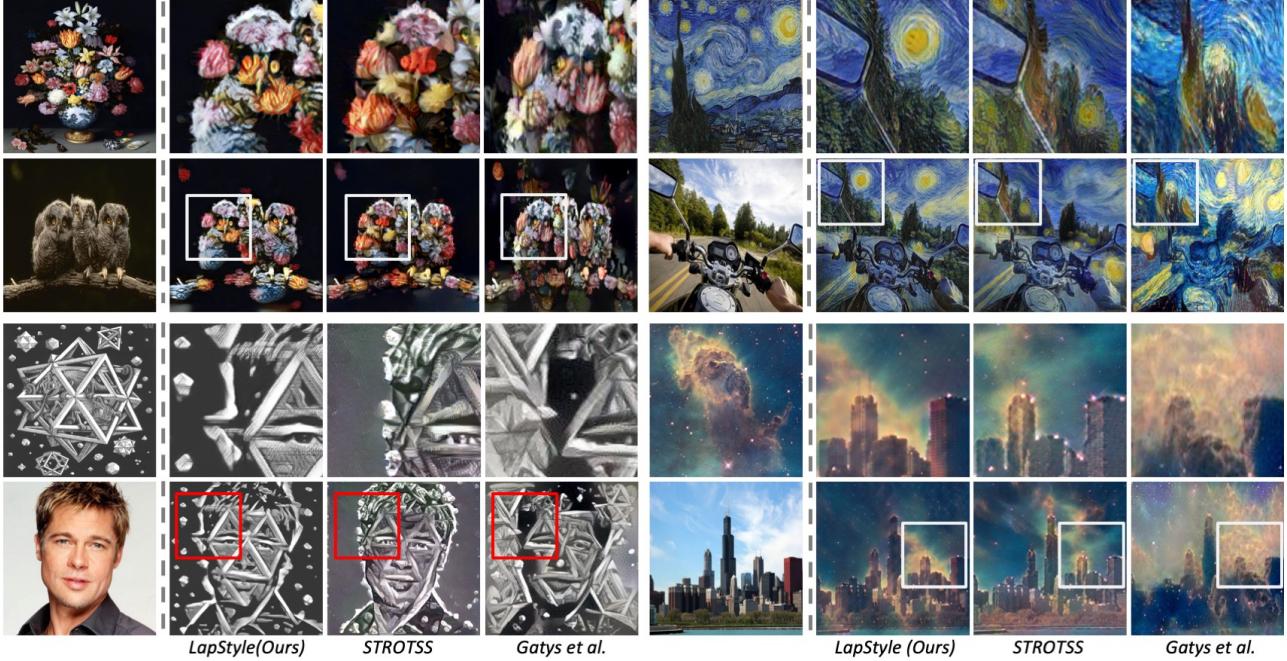


Figure 7. Qualitative comparisons between our method and optimization-based methods STROTSS [15] and Gatys et al. [8].

In Fig. 7, we show some stylization examples synthesized by our method and two optimization-based style transfer methods [15, 8], where zoom-in views are also demonstrated for better comparison. Gatys et al. [8] synthesis stylized image in single scale via optimizing gram matrix. As shown in Fig. 7, although holistic style patterns are transferred, the distribution of style patterns are often inappropriate (e.g. left-down and right-down). Meanwhile, the color distribution of stylized image is not accurate enough. STROTSS [15] is the state-of-the-art optimization-based method, which synthesizes stylized image in multiple scales sequentially (from 32 pix to 512 pix) with EMD loss. As shown in Fig. 7, the stylized images have delicate texture and clear style pattern. As a feed-forward method, our method achieves comparable stylization results with STROTSS. In some cases (e.g. top-right and bottom-left in Fig. 7), large scale patterns are better synthesized by our method. The comparative advantage of STROTSS is that style patterns and context structures are combined better in some cases (e.g. bottom-right in Fig. 7), due to its optimization process.

User Study. We choose 15 style images and 15 content images to synthesize 225 images in total using our method and 5 competitive SOTA methods. Then, we randomly sample 20 content-style pairs. For each pair, we display stylized images side-by-side in a random order to subjects and ask them to select their most favorite one. As shown in Fig 5, we collect 2000 votes from 100 users and show the percentage of votes for each method in the form of pie chart. The

comparison results demonstrate that our stylized results are significantly more appealing than competitors.

Efficiency Analysis. We compare the efficiency of our proposed method and other optimization methods and feed-forward methods. Two image scales are used during evaluation: 256 and 512 resolution. For 512-pixel inference, two Revision Networks are used. All experiments are conduct using a single Nvidia Titan X GPU. The comparison results are shown in Table. 1, where we can find our method runs in real-time and achieves 120 fps and 110 fps with 256-pix and 512-pix, respectively. There are three reasons for the fast inference speed: (1) The Drafting Network is build upon low-resolution; (2) AdaIN module is efficient and (3) the Revision Network is shallow. As shown in Fig. 6 and Fig. 7, the quality of stylized image generated by our methods is comparable with optimization-based method, and is significantly better than feed-forward methods. To conclude, Table. 1 demonstrates that our method achieves the SOTA inference speed among feed-forward methods, and is much more faster than optimization methods.

4.3. Ablation Study

Loss Function. We conduct ablation experiments to verify the effectiveness of each loss term used for training LapStyle, the results are shown in Fig. 8. (1) Without rEMD loss l_r , the style patterns of yellow circle disappear and the overall stylization degree is decreased. This result demonstrates the effectiveness of rEMD loss, and we are the first to train

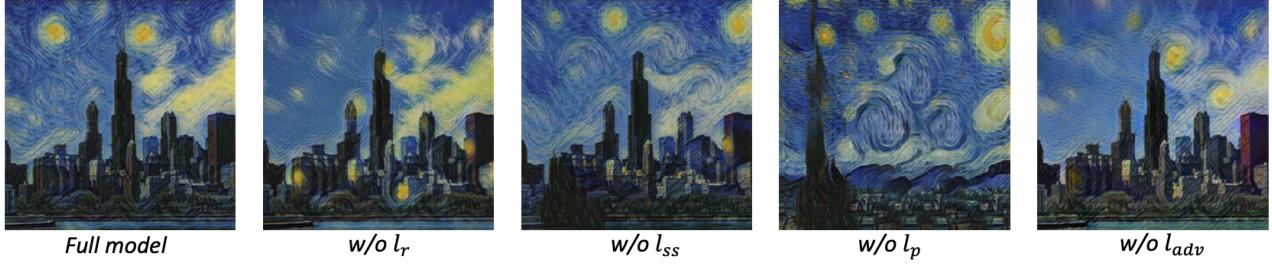


Figure 8. Ablation study of effects of loss function used during training. Here, l_r , l_{ss} and l_p are used in both networks, while l_{adv} only used in the Revision Network. l_m is used in all ablation settings to keep style transferred. Best viewed zoomed-in on screen.

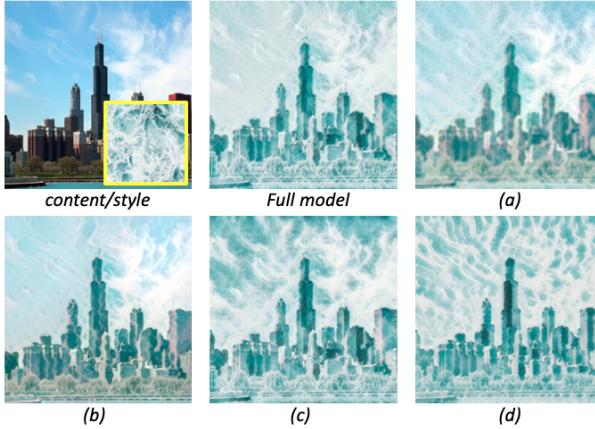


Figure 9. Ablation study of effectiveness of the Revision Network. (a) Drafting Network trained on 128 pix (result image is up-sampled to 256 pix). (b) Drafting Network directly trained on 256 pix. In (a) and (b), Revision Network is not used. (c) Revision Network is directly built upon RGB image instead of difference image. (d) Revision Network is trained without Drafting Network.

feed-forward network with rEMD loss; (2) Without self-similarity loss l_{ss} , some inappropriate black style patterns appear in the bottom-left corner. (3) Without perceptual loss l_p , the structure of content image is totally discarded and the LapStyle directly re-builds the style image. These results suggest that l_p is necessary for our method, meanwhile l_{ss} can further constrain the content consistency to achieve better style distribution. (4) Without adversarial loss l_{adv} , the texture quality and color distribution become worse than full model. This comparison demonstrates that the adversarial learning in revision phase can effectively improve the stylization quality, especially local texture and color distribution.

Effectiveness of Revision Network. The results of ablation experiments are shown in Fig. 9. Before revision, the result of Drafting Network is blur in Fig. 9 (a), due to low resolution. If we directly train Drafting Network on 256 pix, as Fig. 9 (b) shows, the result is clear but its stylization degree is limited. These results demonstrate the effectiveness and

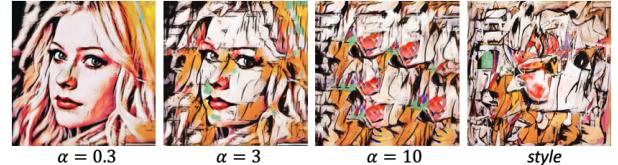


Figure 10. Trade-off of content-style losses.

necessary of our “Drafting and Revision” framework. Another question is whether it is necessary to revise the rough stylized image with the help of Laplacian difference image? The image of Fig. 9 (c) is directly generated by the Revision network in RGB space. We can see the style distribution of revision result is divorced from the drafting image (a) and it seems less harmony than the original result. This observation suggests revising stylized image in a residual form is more controllable and can generate better results.

Effectiveness of Drafting Network. As shown in Fig. 9, without DraNet, the RevNet can still capture style patterns to some extend, but significantly worse than full model.

Content-style Tradeoff. In the training phase, we can control the stylization degree by adjusting the weight term α . As shown in Fig. 10, the network tends to preserve more details and structures of the content image with low style loss, and synthesize excess style patterns with high style loss.

5. Conclusion

In conclusion, we propose a new feed-forward style transfer algorithm LapStyle which synthesizes stylized image in a progressive procedure. In LapStyle, we propose the novel framework “Drafting and Revision”, which first synthesizes a rough drafting with global pattern and then revises local style patterns according to residual image generated with the help of Laplacian filtering. Experiments demonstrate that our method is effective and efficient. It can synthesize images that are preferred over other state-of-the-art feed-forward style transfer algorithms and can run in real-time. Currently, our LapStyle is designed following the *Per-Style-Per-Model* fashion, arbitrary style transfer is left to be our future work.

References

- [1] Alex J Champandard. Semantic style transfer and turning two-bit doodles into fine artworks. *arXiv preprint arXiv:1603.01768*, 2016. 2
- [2] Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. Coherent online video style transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1105–1114, 2017. 2
- [3] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stylebank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1897–1906, 2017. 2
- [4] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stereoscopic neural style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6654–6663, 2018. 2
- [5] Tian Qi Chen and Mark Schmidt. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337*, 2016. 2
- [6] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015. 2
- [7] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. 2
- [8] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 1, 2, 5, 7
- [9] Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3985–3993, 2017. 2
- [10] Shuyang Gu, Congliang Chen, Jing Liao, and Lu Yuan. Arbitrary style transfer with deep feature reshuffle. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8222–8231, 2018. 2
- [11] Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. Real-time neural style transfer for videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 783–791, 2017. 2
- [12] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 2, 3, 5
- [13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 2, 3, 5
- [14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [15] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019. 1, 2, 3, 4, 5, 7
- [16] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017. 2
- [17] Chuan Li and Michael Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2479–2486, 2016. 2
- [18] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European conference on computer vision*, pages 702–716. Springer, 2016. 2
- [19] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3809–3817, 2019. 2, 5
- [20] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Diversified texture synthesis with feed-forward networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3920–3928, 2017. 2
- [21] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in neural information processing systems*, pages 386–396, 2017. 2, 5
- [22] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 453–468, 2018. 2
- [23] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088*, 2017. 2
- [24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5
- [25] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5880–5888, 2019. 2, 5
- [26] Fred Phillips and Brandy Mackintosh. Wiki art gallery, inc.: A case for critical thinking. *Issues in Accounting Education*, 26(3):593–608, 2011. 5
- [27] Eric Risser, Pierre Wilmot, and Connelly Barnes. Stable and controllable neural texture synthesis and style transfer using histogram losses. *arXiv preprint arXiv:1701.08893*, 2017. 1
- [28] Ahmed Selim, Mohamed Elgharib, and Linda Doyle. Painting style transfer for head portraits using convolutional neural networks. *ACM Transactions on Graphics (ToG)*, 35(4):1–18, 2016. 2

- [29] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4570–4580, 2019. [2](#), [4](#)
- [30] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8242–8250, 2018. [2](#)
- [31] YiChang Shih, Sylvain Paris, Connelly Barnes, William T Freeman, and Frédo Durand. Style transfer for headshot portraits. *ACM Transactions on Graphics (TOG)*, 33(4):148, 2014. [2](#)
- [32] Dmitry Ulyanov, Vadim Lebedev, Victor Lempitsky, et al. Texture networks: Feed-forward synthesis of textures and stylized images. In *International Conference on Machine Learning*, pages 1349–1357, 2016. [2](#)
- [33] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. [2](#)
- [34] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4105–4113. IEEE, 2017. [2](#)
- [35] Xin Wang, Geoffrey Oxholm, Da Zhang, and Yuan-Fang Wang. 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*, pages 5239–5247, 2017. [2](#)
- [36] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9036–9045, 2019. [2](#)
- [37] Hang Zhang and Kristin Dana. Multi-style generative network for real-time transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. [2](#)

Table 2. The encoder architecture of the Drafting Network. The shape of an input sample is $3 \times 128 \times 128$. We use “same padding” for each convolution layer, and ReLU is added to the end of each convolution layer.

stage	output	architecture
$F_{1.1}$	$64 \times 128 \times 128$	1×1 conv, 3 3×3 conv, 64
	$64 \times 128 \times 128$	3×3 conv, 64
$F_{1.2}$	$64 \times 64 \times 64$	2×2 max pooling, stride 2
	$128 \times 64 \times 64$	3×3 conv, 128
$F_{2.1}$	$128 \times 64 \times 64$	3×3 conv, 128
	$128 \times 32 \times 32$	2×2 max pooling, stride 2
$F_{2.2}$	$256 \times 32 \times 32$	3×3 conv, 256
	$256 \times 16 \times 16$	2×2 max pooling, stride 2
$F_{3.1}$	$256 \times 16 \times 16$	3×3 conv, 512
$F_{3.2}$	$512 \times 16 \times 16$	3×3 conv, 512
$F_{4.1}$	$512 \times 16 \times 16$	3×3 conv, 512

6. Supplementary

The outline of this supplementary material is as follows:

- The architecture details of our proposed LapStyle.
- A video stylization demo generated by LapStyle.
- Qualitative results of high resolution stylization.

6.1. Network Structure

In this supplementary material, we introduce the detailed configuration of our proposed LapStyle, which is comprised of a Drafting Network and a Revision Network.

The Drafting Network contains an encoder and a decoder. The network configuration of encoder is shown in Table. 2, which is a part of VGG-16 network and pre-trained on ImageNet dataset. We do not optimize the encoder during training. The architecture of decoder is shown in Table. 3, where AdaIN modules are added in multiple granularity to ensure style patterns of different granularity are properly transferred. The Revision Network (Table. 4) has a simple encoder-decoder architecture of only 6 convolution layers. The light architecture ensures fast speed of the Revision Network, meanwhile restricts the potential of Revision Network so that it can only handle local style patterns.

6.2. Video Stylization

As stated in the part of efficiency analysis in our main draft, our LapStyle can synthesis stylized image in real-time. Thus, it is suitable for video stylization application. In Fig. 11, we demonstrate the style image and one frame of video stylization demo. As shown in the video, the synthesized style patterns are generally stable with jitter to some extent. These jitter can be further removed by optical-flow based smoothness.

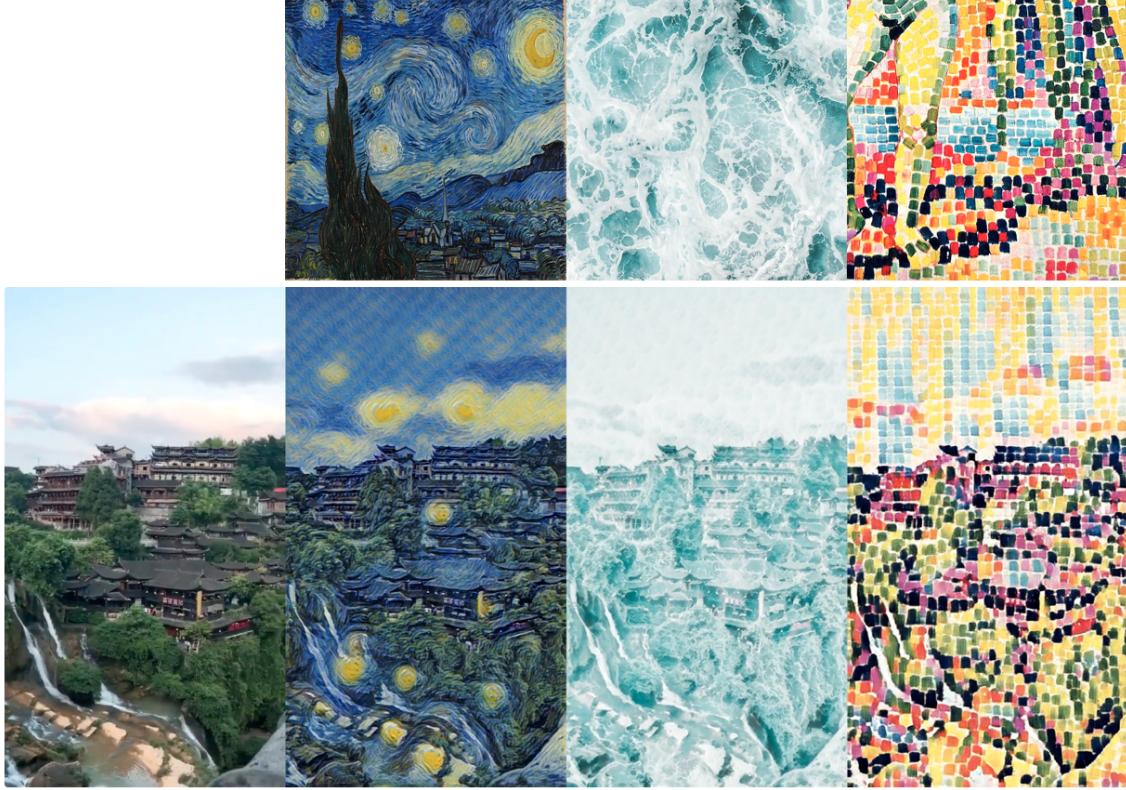


Figure 11. Video stylization demo. This figure illustrates a single frame from video, which can be found in supplementary material.

Table 3. The decoder architecture of the Drafting Network. *AdaIN* is used in 4_1, 3_1, 2_1. A resblock consists of a 3×3 convolution layer with *ReLU* activation, an 1×1 convolution layer and a residual connection.

stage	output	architecture
F_4^d	$256 \times 16 \times 16$	$\text{AdaIN}(F_{4,1}^c, F_{4,1}^s)$ ResBlock(512) 3×3 conv, 256, <i>Relu</i>
F_3^d	$128 \times 32 \times 32$	upsample, scale 2 add $\text{AdaIN}(F_{3,1}^c, F_{3,1}^s)$ ResBlock(256) 3×3 conv, 128, <i>Relu</i>
F_2^d	$64 \times 64 \times 64$	upsample, scale 2 add $\text{AdaIN}(F_{2,1}^c, F_{2,1}^s)$ 3×3 conv, 128, <i>Relu</i> 3×3 conv, 64, <i>Relu</i>
F_1^d	$3 \times 128 \times 128$	upsample, scale 2 3×3 conv, 64, <i>Relu</i> 3×3 conv, 3

Table 4. The architecture of the Revision Network. The shape of input sample is $6 \times H \times W$, which is the concatenated tensor combined by different image and stylized image generated by the Drafting Network (or the Revision Network at the previous scale).

stage	output	architecture
R_1	$64 \times \frac{H}{2} \times \frac{W}{2}$	3×3 conv, stride 1, 64, <i>Relu</i> 3×3 conv, stride 2, 64, <i>Relu</i>
R_2	$64 \times \frac{H}{2} \times \frac{W}{2}$	ResBlock(64)
R_3	$3 \times H \times W$	upsample, scale 2 3×3 conv, stride 1, 64, <i>Relu</i> 3×3 conv, stride 1, 3

6.3. High Resolution Stylization

In supplementary material, we further demonstrate some stylization results on high resolution. Two Revision Networks are adopted at 256 px and 512 px successively to generate stylized images with 512 px. As shown in Fig. 12, style patterns on multiple granularity are properly transferred by our method.

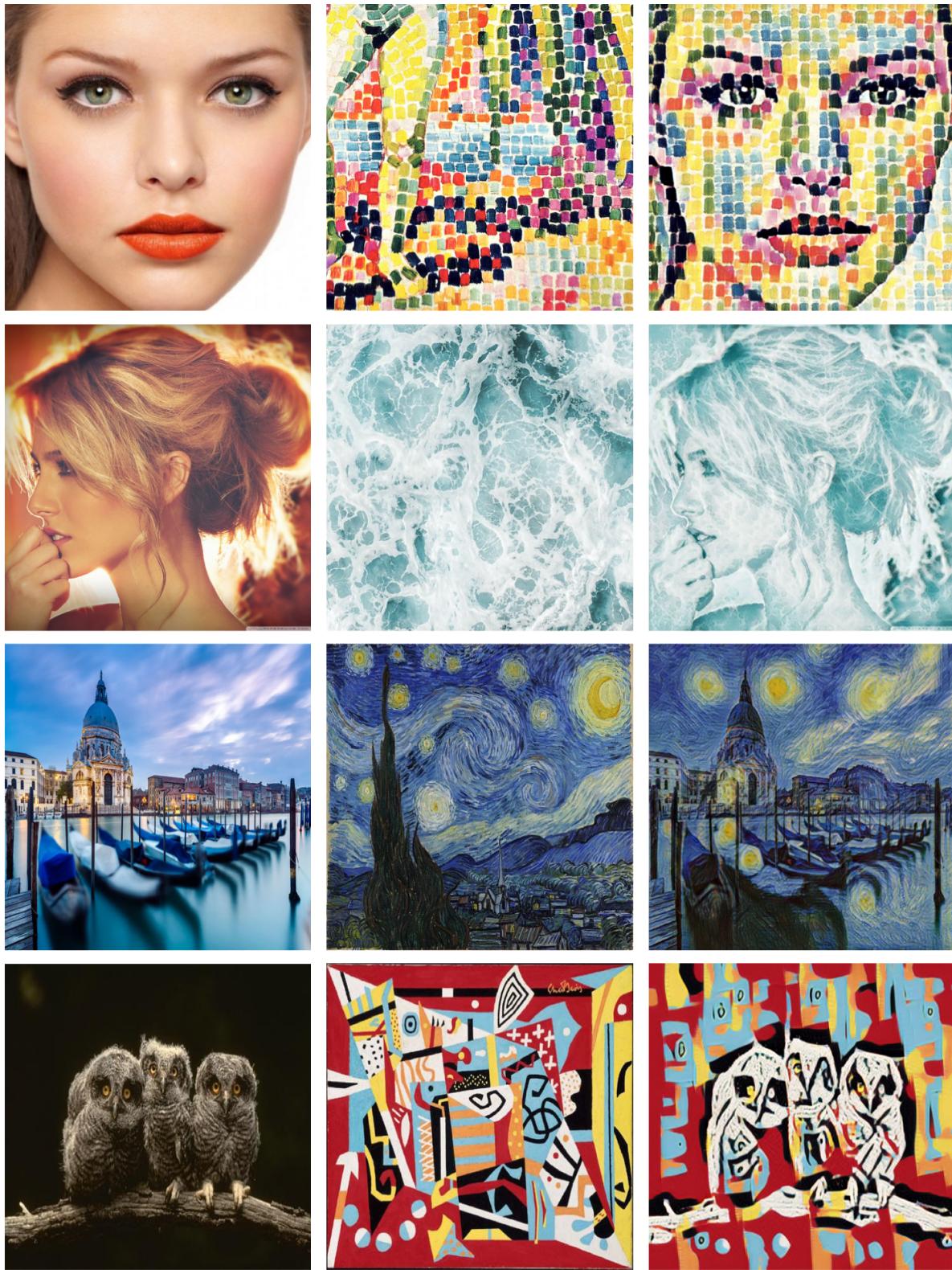


Figure 12. Qualitative results of high resolution stylization. Here, left images are content images, middle images are style images and right images are stylized images at 512 px.