Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

From Reality to Perception: Genre-Based Neural Image Style Transfer

Zhuoqi Ma[†], Nannan Wang[‡*], Xinbo Gao[†], Jie Li[†]
[†] State Key Laboratory of Integrated Services Networks,
School of Electronic Engineering, Xidian University, Xi'an 710071, China
[‡] State Key Laboratory of Integrated Services Networks,
School of Telecommunications, Xidian University, Xi'an 710071, China

Abstract

We introduce a novel thought for integrating artists perceptions on the real world into neural image style transfer process. Conventional approaches commonly migrate color or texture patterns from style image to content image, but the underlying design aspect of the artist always get overlooked. We want to address the in-depth genre style, that how artists perceive the real world and express their perceptions in the artwork. We collect a set of Van Goghs paintings and cubist artworks, and their semantically corresponding real world photos. We present a novel genre style transfer framework modeled after the mechanism of actual artwork production. The target style representation is reconstructed based on the semantic correspondence between real world photo and painting, which enable the perception guidance in style transfer. The experimental results demonstrate that our method can capture the overall style of a genre or an artist. We hope that this work provides new insight for including artists perceptions into neural style transfer process, and helps people to understand the underlying characters of the artist or the genre.

1 Introduction

While Vincent Van Gogh was looking out of the window of his asylum room at Saint-Rmy-de-Provence in a hot summer night, the glow of the stars and the moon were swirling against the dark blue sky in his vision. Van Gogh depicted his perception that view brought him with vigorous colors, swaying brushstrokes and distorted scenes, which made Starry Night one of the most famous art pieces of the world.

After years since Van Gogh passed away, the world start to explore and identify the unique beauty of his work. People start to study how to recreate images in his style. This problem, now being categorized under image artistic rendering field, has drawn much attention in recent years.

Development in convolutional neural networks for computer vision tasks has enlightened a neural style transfer paradigm shift in this field. Both academia and industry have witnessed the great success that neural style transfer has brought for the field. Typical neural style transfer approaches either train feed-forward networks [Johnson *et al.*, 2016; Zhu *et al.*, 2017a; Chen *et al.*, 2017; Dumoulin *et al.*, 2016], or apply iterative optimization methods [Gatys *et al.*, 2016] to migrate visual attribute, with filter pyramid of VGG network [Simonyan and Zisserman, 2015] as representations for content and style images. These methods yield impressive results, but feature maps at different layers are directly used to capture pixel correlations of the content and the style images, which lacks of structural constraint. In view of this problem, some research [Li and Wand, 2016; Liao *et al.*, 2017] regard this as texture or image synthesis, which enhanced the structure preservation in style migration and generated more credible results.

However, the style of an artist or a genre includes not only color or texture patterns of one single painting, but also the underlying creation ideas which were normally neglected in previous methods. In art production, the way that artists perceive the real world and express their perceptions in the artworks are critical design aspects to distinguish one style from the other. For example, the bright colors are often adopted to express the tension of life in Van Goghs paintings, while the broken up and reassembled geometry patches symbolize cubism artworks. Although current methods are capable of capturing visual appearance of single image, the overall style of the artist or the genre can not get fully expressed in the result.

In this paper, we present a framework for genre-based image style transfer. Driven by the well-known assumption that artists express their feelings of the real world in artworks, we introduce perception guidance in genre style transfer. We collect 309 pairs of Van Goghs painting and the corresponding photo that have similar scenery, some of them were shot from the site he used to live. We also collect cubist artworks and photos with similar semantic meanings to demonstrate the generalization of our method. We train an encoder-decoder network with separate genre paintings. The output of encoder forms an embedding representation space for each genre, from which the perception-guided style representation is reconstructed. Then the style representation is transformed

*Corresponding author: Nannan Wang (nnwang@xidian.edu.cn)

into neural artwork by the decoder.

The reconstruction is based on the reality-perception correspondence between semantically-related photo and painting. Solving the correspondence problem is equivalent to matching the heterogeneous domains, in our case is the photo domain and artwork domain. Recently, deep convolutional neural networks [Simonyan and Zisserman, 2015] trained for discriminative tasks have brought up striking success. An excellent property of deep CNNs is that they encode image information into hierarchical representations, with spatial consistency between adjacent layers. This property facilitates a coarse-to-fine Nearest-Neighbor Field (NNF) computing, an essential link in establishing the correspondence from the reality to perception. Our method adopts pyramid of feature maps from pre-trained CNNs for image classification as representations for both domains. We use correlation coefficient of the deep representations as similarity metric. Heterogeneous images such as photos and oil paintings, have different visual appearances even when they are semantically alike. Using correlation coefficient could eliminate the affect of the amplitude change brought by different visual appearances, and in the meanwhile reflect the semantic similarity over the variance of deep representations in local region.

The major contributions of this work are:

1) We present a novel framework for incorporating reality-perception correspondence into neural image style transfer, which we show to be effective for generating artworks with distinctive genre characters.

2) We extend PatchMatch with a new similarity metric in deep feature space for heterogeneous domain matching, which proved to provide better semantic discriminability.

Neural artworks could be more than just migrating color distributions or texture patterns from one style image. It can also express the underlying design ideas as we model after real human artists by integrating their perceptions into style transfer process. We show that our method can generate artworks with distinct genre characters through extensive experiment results. We believe that the proposed framework could provide new insights for considering peoples role in image style transfer problems.

2 Related Work

Artistic Image Style Transfer. The target of image artistic rendering [Semmo *et al.*, 2017] is to simulate the appearance of reference image while maintaining the content of the source image. Traditional methods [Hertzmann *et al.*, 2001; Semmo *et al.*, 2015] employs machine learning or statistical models to emulate the visual appearance from example pairs. However, these methods only perform low-level texture transfer, which limited its control over the design aspects of artwork creation.

Representing the style and content images had always been technical limitations in image artistic rendering problems before, but recent advancements in deep learning and convolutional neural networks (CNNs) has enlightened a paradigm shift in this field. The first attempt to generate neural artwork may be DeepDream [Mordvintsev *et al.*, 2015]. Inspired by their work, Gayts *et al*. [2016] produce impressive results for transferring styles of famous paintings onto images. Their method applies different layers of pre-trained VGG network as filters to extract content and style representations. Then, the output is stylized based on pixel correlations between the representation of the style image and the content image. This idea has been further developed to painting style transfer on head portraits [Selim *et al.*, 2016] and style image synthesis [Li and Wand, 2016]. Semantic segmentation were introduced in CNNs to turn two-bit doodles into fine art [Champandard, 2016].

Aside from the aforementioned iterative optimization approaches, more and more work start to train feed-forward networks for faster or even real-time stylization. Zhu *et al*. [2017a] introduced cycle-consistent adversarial networks for unpaired images and applied their method on style transfer. Johnson *et al* [2016] designed a transform net which is trained with custom perceptual loss functions for real-time image style transfer. Chen *et al* [2017] and Dumoulin *et al*. [2016] designed network architectures to learn and capture the representation of artistic style, and further enabled fusion of different styles.

However, among all of these methods, the style they transferred from one or more reference images are merely colors and texture patterns, but the underlying design aspects involved in actual artwork creation have not been taken into consideration. In contrast, our framework incorporated artists perceptions of the real world into style transfer process, which complements the artwork production mechanisms, and produce results with distinctive genre character. We reconstruct the perception-guided style representation in an embedding space created by the auto-encoder trained with genre dataset, as it has more genre-specific characteristic than original pixel intensities or dCNN features.

Heterogeneous Domain Transformation. The core of our method is to reconstruct the style representation of the real world photo based on the correspondence from reality to perception, which could be regarded as heterogeneous domains. Methods have been proposed for heterogeneous image transformation problems, such as photos to sketches [Wang *et al.*, 2017], and photos to paintings [Semmo *et al.*, 2015]. However, these methods are very target-specific and can not be easily extended to other image styles.

Early matching methods, such as optical flow [Brox *et al.*, 2004] and PatchMatch [Barnes *et al.*, 2009], were designed in image space to match local image statistics like brightness consistency or pixel intensities, which would easily fail under appearance variations. Then, the development of various handcrafted local region descriptors [Lowe, 2004] invoked noteworthy progress. These descriptors are robust to visual variations and local transformations. Some methods [Liu *et al.*, 2016] combine these descriptors with dense matching to perform more reliable matching. However, these approaches are still based on low-level features and hence would fail to match semantically-related images. While other methods [Zhu *et al.*, 2017b; Liao *et al.*, 2017] apply dCNN representations in semantic matching, they directly use Euclidean distance between deep features as similarity measurement. However, images from heterogeneous domains are commonly different in visual appearances, even when they

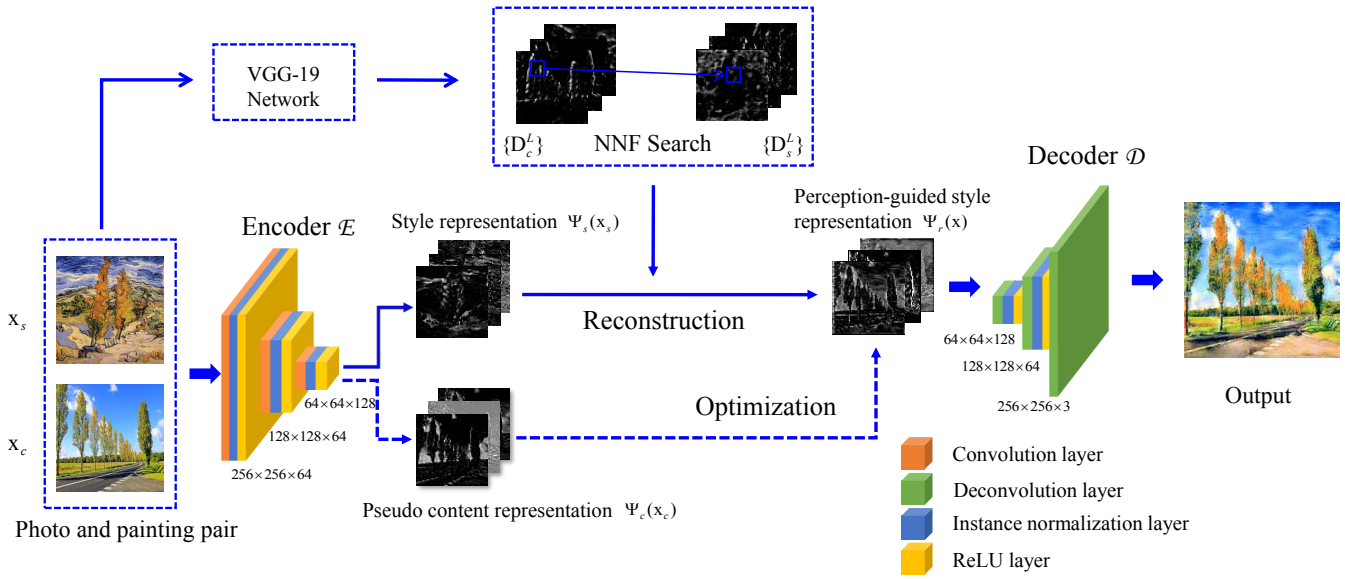

Figure 1: Our framework consists of three parts: encoder E, representation reconstruction module, decoder D.

have similar semantic contents. As a result, the Euclidean distance between their deep representations does not serve the best as semantic similarity measurement.

3 Model

As shown in Figure 1, our framework consists of three parts: encoder E, representation reconstruction module, and decoder D. Encoder E and decoder D are combined during training and separated to perform encoding and decoding functions in stylization. The framework is designed after the process of artwork creation, where artists first perceive the real world scenes, and then create artworks with distinctive styles based on their personal perceptions.

The encoder-decoder network creates an embedding representation space, in which to reconstruct perception-guided style representation $\Psi_r(\mathrm{x})$. To include perception guidance, we compute a NNF that establishes correspondences between dCNN features of the photo and the painting. In stylization phase, the encoder subnetwork E transforms the real world photo into pseudo content representation $\Psi_c(\mathrm{x})$. The reconstructed style representation $\Psi_r(\mathrm{x})$ would be optimized by minimizing its difference from $\Psi_c(\mathrm{x})$ for preserving better structure with the real world photo. Then the decoder subnetwork D would transform the optimized style representation $\Psi_r{}'(\mathrm{x})$ into neural artwork.

3.1 Encoder-Decoder Network

Network Architecture. Inspired by the architecture used in [Johnson *et al.*, 2016], the encoder E contains 3 convolutional layers: one stride-1 convolution layer followed by two stride-2 convolution layers to downsample the input image. Symmetrically, the decoder D starts with two convolutional layers with stride $1/2$ to upsample the style representation and concatenate with one stride 1 convolution layer. Each convolution layer is followed by an instance normalization

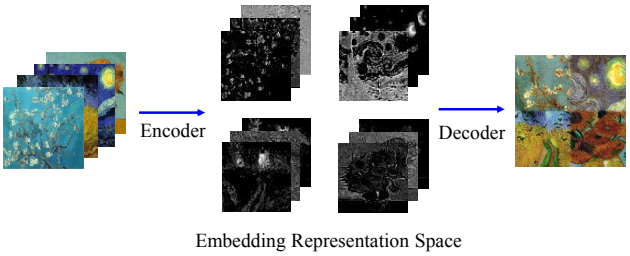

Figure 2: The style representations in embedding space can be pieced out and recombined to form a new artwork.

[Ulyanov *et al.*, 2016] and ReLU nonlinearity layer to form a block, except for the output layer. The first and last convolutional layers use 9×9 kernels, other layers use 3×3 kernels. Jonhson *et al.* [2016] added residual blocks for its capability of learning identity function. However, as the target of our network is to create a stable representation space for each genre, we discard residual blocks because its mechanism of adding residuals to the input would introduce noise into embedding space.

Training. We train the network with separate art collections to produce an output painting as same as possible to the input painting. Each painting is resized to 256×256. The identity loss $L_{identity}$ is the squared Euclidean error between input image I and output image O:

$$L_{identity} = \|O - I\|^2 \tag{1}$$

Genre-specific Representation Space. The output of the encoder forms an embedding representation space for each genre. The 256×256 paintings are encoded into $128 \times 64 \times 64$ style representations $\Psi_s(\mathrm{x})$, a 3D tensor with $channel \times width \times height$. From Figure 2 we can see that, the artworks are encoded into multi-layer feature maps, and when we piece out the fragments of the feature maps from different

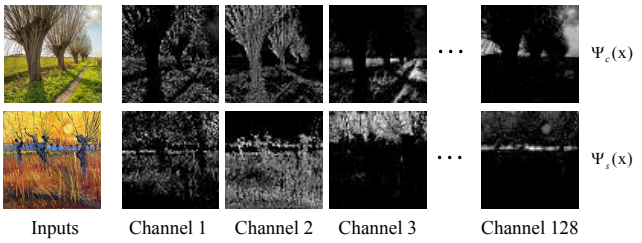

Figure 3: Illustration of corresponding channels of $\Psi_c(\mathrm{x})$ and $\Psi_s(\mathrm{x})$ in embedding space.

paintings, the decoder can still transform it into a plausible artwork.

In stylization phase, the encoder E and decoder D operate independently. We feed real world photo to E to generate pseudo content representation $\Psi_c(\mathrm{x})$, the layout of which is same with the input photo while the response amplitude is identical with style representation $\Psi_s(\mathrm{x})$. Figure 3 demonstrates some corresponding channels of $\Psi_s(\mathrm{x})$ and $\Psi_c(\mathrm{x})$. Each channel of the feature maps generated by a convolution kernel could be considered as one aspect of genre's style characteristic.

3.2 Style Representation Reconstruction

Given an real world photo and artwork pair x_c and x_s, which is different considerably in appearance but have similar semantic contents, we reconstruct a perception-guided style representation $\Psi_r(\mathrm{x})$ for real world photo by finding the mappings from the deep feature maps of x_c to x_s.

Preprocessing

Our method starts by feeding the photo and the painting into VGG-19 network [Simonyan and Zisserman, 2015], which was pre-trained on ImageNet database for image classification tasks, and perform forward propagation. We obtain the pyramid of feature maps $\{\mathrm{D}_c^L\}$ and $\{\mathrm{D}_s^L\}$ $(L = 1, 2, 3)$ for the input photo x_c and painting x_s. The feature maps of each scale derive from the $reluL\_2$ layer of the Lth convolution blocks in VGG-19. The deep feature maps $\{\mathrm{D}_c^L\}$ and $\{\mathrm{D}_s^L\}$ are resized to 64×64, the same size of the style representation $\Psi_s(\mathrm{x})$. Then we divide all the representations and deep feature maps into 3×3 patches, with stride 1. The patches are very densely sampled to achieve the best reconstruction quality.

Coarse-to-fine Nearest-Neighbor Field Search

We adopt a coarse-to-fine searching strategy for computing NNF: we estimate a quick initial NNF guess by taking PatchMatch [Barnes *et al.*, 2009] into deep feature space, then a more precise NNF is obtained by searching the neighborhood around the initial guess. The extracted NNF together with representations from the embedding space are used to reconstruct perception-guided style representation of the real world photo.

For deep feature maps at layer L, we estimate an initial NNF represented by mapping functions $f_{c\to s}^L$. $f_{c\to s}^L$ maps a point in deep feature map D_c^L to another in D_s^L. $f_{c\to s}^L$ is computed by minimizing the following function:

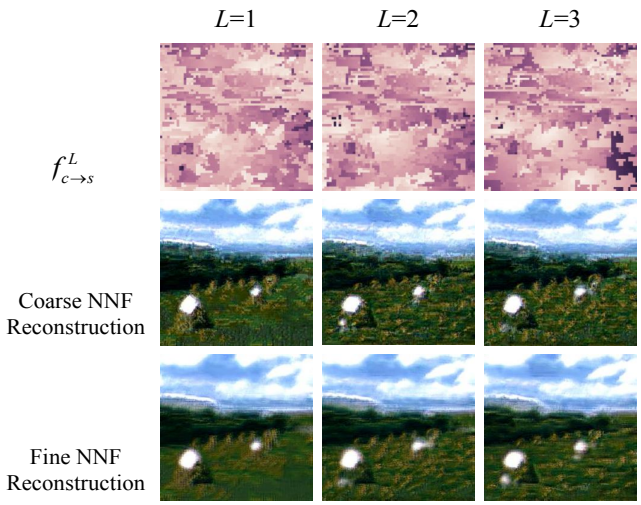

Figure 4: Visualization of NNFs (top row), reconstructed results with coarse NNFs (middle row) and reconstructed results with fine NNFs (bottom row).

$$f_{c\to s}^L(p) = \arg\min_q \sum_{x\in M(p), y\in M(q)} -corr(\mathrm{D}_c^L(x), \mathrm{D}_s^L(y)), \tag{2}$$

where $M(p)$ is the patch around pixel p. The patch size is set to be 3×3. $corr(\mathrm{D}_c^L(x), \mathrm{D}_s^L(y))$ is the correlation coefficient between $\mathrm{D}_c^L(x)$ and $\mathrm{D}_s^L(x)$. For each patch around pixel p in D_c^L, we find its nearest neighbor position $q = f_{c\to s}^L(p)$ in D_s^L.

Equation (2) can be efficiently optimized with PatchMatch method [Barnes *et al.*, 2009]. Here, we adopt PatchMatch in deep feature domain for better representation capability, and we use correlation coefficient between deep feature maps as similarity metric to eliminate the affect of visual appearances variation over semantic contents.

$f_{c\to s}^L$ obtained from optimizing Equation (2) is considered as an initial coarse NNF, and is further transformed into nearest indexes $\phi_{c\to s}^L(x_i)$ for patches.

For the ith patch $\mathrm{D}_c^L(x_i)$ in deep feature maps D_c^L, we find its best matching patch $\mathrm{D}_s^L(x_{NN(i)})$ in deep feature maps D_s^L within the search region around $\phi_{c\to s}^L(x_i)$ for patches:

$$NN(i) = \arg\min_j -corr(\mathrm{D}_c^L(x_i), \mathrm{D}_s^L(x_j)) \tag{3}$$

Then the obtained nearest patch indexes is transformed into mapping functions $f_{c\to s}^{L+1}$, and serve as guidance for coarse NNF search to limit the random search space in layer $L+1$. The coarse-to-fine NNF search is repeated for layer $1-3$ (as shown in Figure 4), updating correspondences with multi-scale deep feature maps.

Perception-Guided Representation Reconstruction

After we obtain the final nearest indexes for patches, the perception-guided representation $\Psi_r(\mathrm{x_c})$ of the real world photo x_c could be reconstructed from the embedding representation $\Psi_s(\mathrm{x}_s)$ with the reality-perception correspondence. The best matches are stitched together into a complete style representation with overlapping area averaged.

(a) Van Gogh's style

(b) Cubism

Figure 5: Genre-based style transfer results. The top row is real world photo, the bottom row is the stylized neural artworks based on reality-perception correspondence in Van Gogh's style and cubism respectively.

3.3 Optimization

Let $\Psi_r{}'(\mathrm{x})$ denote the modified version of $\Psi_r\,(\mathrm{x_c})$. Our objective is to make the reconstructed style representation be structurally similar to the real world photo. Hence $\Psi_r{}'(\mathrm{x})$ can be obtained by minimizing the following loss function :

$$L_{\Psi_r{}'(\mathrm{x})} = \left\|\Psi_r{}'(\mathrm{x}) - \Psi_r(\mathrm{x_c})\right\| + \alpha \left\|\Psi_r{}'(\mathrm{x}) - \Psi_c(\mathrm{x_c})\right\| \quad (4)$$

, where $\Psi_c(\mathrm{x_c})$ is the pseudo content representation from the encoder with real world photo as input, α is set to 0.2 in our experiment. We minimize Equation (4) using back-propagation with L-BFGS [Zhu *et al.*, 1997]. Such an optimization is similar to updating the stylization result by pre-trained VGG network described in [Gatys *et al.*, 2016], except in our case, the target representation $\Psi_r{}'(\mathrm{x})$ is updated through encoder E. Further, the altered representation would be transformed into neural artworks through pre-trained decoder:

$$O = D(\Psi_r{}'(\mathrm{x})) \quad (5)$$

4 Experimental Results and Analysis

4.1 Experimental Settings

Dataset

We conducted our experiments on our Van Gogh-photo dataset and the cubism-photo dataset. The Van Gogh-photo dataset consists of 309 Van Goghs painting and the semantically corresponding real world photo pairs. The cubism-photo dataset has 34 cubist painting and real world photo pairs. We choose these two genres on account of two reasons: Firstly, Van Goghs paintings and cubism artworks are commonly used as reference images in style transfer methods. Secondly, their paintings have distinctive genre characters, such as the bright colors in Van Goghs paintings and the broken geometry patches in cubism artworks.

Encoder-Decoder Network Training Details

The network is trained on the paintings of separate genres. All the paintings and the photos is scaled to the size of 256×256. We train the network with a batch size of 5 for 300 epochs. And we adopt the Adam optimization method [Kingma and Ba, 2015] with the initial learning rate of 0.001. We do not apply weight decay as the model would not easily overfit.

All experiments are conducted with Python on Ubuntu 16.04 system, with i7-4790 3.6G CPU and 12G NVIDIA Titan X GPU. GPU was used in the network training and deep feature extraction phase. The two most time-consuming parts of our proposed method lie in style reconstruction phase: (1) nearest neighbor field computation (about 120 seconds), which need to compute correlation coefficient over hundreds of feature channels, and (2) optimizing the reconstructed style representation (about 100 seconds), which require about 400 iterations to converge for Equation (4).

4.2 Result and Comparison

Result

The main purpose of our method is to imitate the art production process and transfer the artists and genre's overall style to photo. Figure 5 shows some results of the generated neural artworks and the original real world photos in both genres. For Van Goghs style, our results possess oil painting traits: the strokes is more clear, and there are small brush-like textures resembling the actual oil painting. The colors in the generated artworks are brighter than the original photo, which is a distinct personal painting style of Van Gogh. As for cubism, of which the objects are analyzed, broken up and reassembled in an abstracted form, our results can also express this character rather than merely migrating the color patterns to the photo.

Comparison on Matching

We evaluate the matching quality of our approach and most commonly used matching methods on heterogeneous image pairs with semantically-related scenes. From Figure 6 we can see that existing matching methods ([Barnes *et al.*, 2009; Liu *et al.*, 2016]) is incapable of establishing semantic correspondences between images when they are vastly different in appearances. On the contrary, our matching method can produce more acceptable result.

Comparison on Style Transfer

Recent state-of-the-art techniques in neural style transfer yield very impressive results, we compare our method with

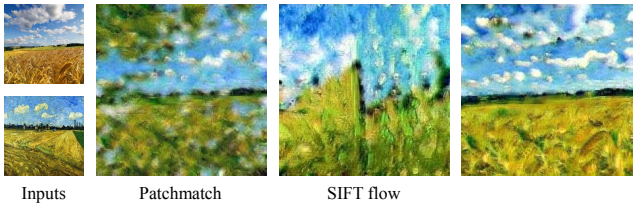

Figure 6: Comparison of different matching methods on semantically-related heterogeneous image pairs but with vastly different appearance.

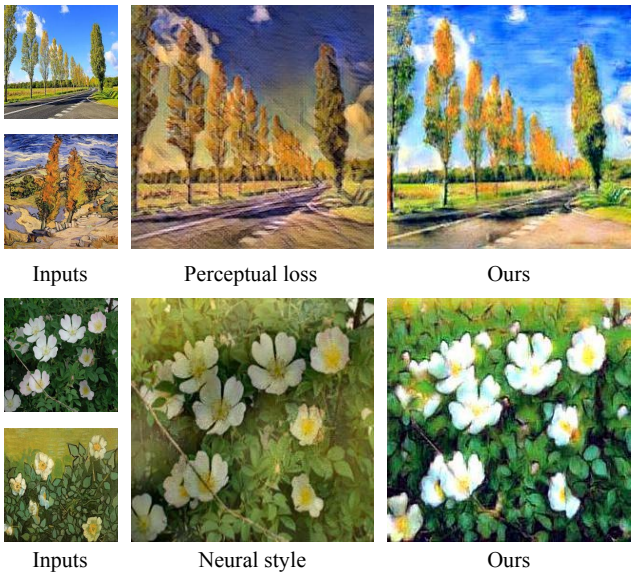

Figure 7: Comparison with Neural Style and Perceptual Loss.

these approaches on two different categories of data: (1) single painting as style image, and (2) using a collection of paintings. It is difficult to evaluate the quality of the stylized image quantitatively as there is no acknowledged quality assessment criteria in neural image artistic rendering. For comparable visual quality, we ran author-released implementations with default settings.

Single Image. Typical CNN-based style transfer methods use only one style image and one content image. As reality-perception correspondence is the core of our idea, we choose one real world photo as content image, and one semantically-related painting from Van Gogh's collections as style image for comparison with other methods.

In Figure 7, we compare our results with two most representative methods under iterative optimization (*Neural Style* [Gatys *et al.*, 2016]) and feed-forward network (*Perceptual Loss* [Johnson *et al.*, 2016]) categories. When the style image and content image are semantically corresponding, CNN-based style transfer methods can cast better color pattern of the original style image than ours, but the result lost semantic meaning to some extent. For example, in the result of Perceptual loss, the color of sky can not remain blue around the tree. And the yellow color pattern in Neural Style result reduced the contrast between flowers and leaves. In our method, stylization is an overall transformation tendency, like brighter colors. Our method mainly emphasize the reality-perception correspondence in genre style transfer, which makes the results more plausible and hand-painted like.

Figure 8: Comparison with CycleGAN.

Using Collection of Paintings. In Figure 8, we compare our result with CycleGAN [Zhu *et al.*, 2017a]. CycleGAN used unpaired images to train the network by minimizing cycle consistency loss. We use author-released Van Gogh-photo datasets to train CycleGAN for Van Gogh style transfer. However, due to the absence of reality-perception correspondence in the unpaired training data and the property of generative network, their result is more similar to the input real world photo and lack of painting characteristic. In comparison, our result has distinct hand-painted feature in strokes, and the colors are brighter.

5 Discussion and Conclusion

In this paper, we proposed a novel framework for genre-based style transfer. The key insight of the framework is that we redefine the style representation of the input photo based on the reality-perception correspondence from the genre. We modeled after the actual artists' creation process, trying to incorporate the specific artists or genres overall character into style transfer. The artworks generated in Van Goghs style is much brighter in color, and the ones in cubism have similar broken and ensemble feature. We believe this method may provide a new idea in neural style transfer. It may also be proven useful for other genre-based applications, such as genre classification, and art piece identification.

There are still some interesting issues for further investigation. For example, finding correspondence from the reality to perception is not completely reliable. There are chances that an object in the real world photo has no match in the whole genre, then the semantic correspondence could not be well-established. Moreover, for paintings that are highly abstractions of the real world, the correspondence from reality to perception is hard to establish, as dCNN feature maps lack discrimination for abstracted paintings. A possible solution would be to train a regression model on real world element and its abstraction pairs to learn the correspondence rather than simply matching the dCNN features.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (under Grants 61501339, 61772402, 61671339, 61432014, U1605252), in part by Young Elite Scientists Sponsorship Program by CAST (under Grant 2016QNRC001), in part by Natural Science Basic Research Plan in Shaanxi Province of China (under Grant 2017JM6085), in part by Young Talent fund of University Association for Science and Technology in Shaanxi, China, in part by CCF-Tencent Open Fund (under Grant IAGR 20170103), in part by the Leading Talent of Technological Innovation of Ten-Thousands Talents Program under Grant CS31117200001.

References

[Barnes *et al.*, 2009] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics-TOG*, 28(3):24, 2009.

[Brox *et al.*, 2004] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *European conference on computer vision*, pages 25–36. Springer, 2004.

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

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

[Dumoulin *et al.*, 2016] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *International Conference on Learning Representations*, 2016.

[Gatys *et al.*, 2016] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2414–2423. IEEE, 2016.

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

[Johnson *et al.*, 2016] 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.

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[Li and Wand, 2016] 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.

[Liao *et al.*, 2017] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *ACM Transactions on Graphics-TOG*, 36(4), 2017.

[Liu *et al.*, 2016] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. In *Dense Image Correspondences for Computer Vision*, pages 15–49. Springer, 2016.

[Lowe, 2004] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[Mordvintsev *et al.*, 2015] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. *Google Research Blog. Retrieved June*, 20(14):5, 2015.

[Selim *et al.*, 2016] Ahmed Selim, Mohamed Elgharib, and Linda Doyle. Painting style transfer for head portraits using convolutional neural networks. *ACM Transactions on Graphics (ToG)*, 35(4):129, 2016.

[Semmo *et al.*, 2015] Amir Semmo, Daniel Limberger, Jan Eric Kyprianidis, and Jürgen Döllner. Image stylization by oil paint filtering using color palettes. In *Proceedings of the workshop on Computational Aesthetics*, pages 149–158. Eurographics Association, 2015.

[Semmo *et al.*, 2017] Amir Semmo, Tobias Isenberg, and Jürgen Döllner. Neural style transfer: a paradigm shift for image-based artistic rendering? In *Proceedings of the Symposium on Non-Photorealistic Animation and Rendering*, page 5. ACM, 2017.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[Ulyanov *et al.*, 2016] Dmitry Ulyanov, Andrea Vedaldi, and Victor S Lempitsky. Instance normalization: The missing ingredient for fast stylization. *CoRR*, 2016.

[Wang *et al.*, 2017] Nannan Wang, Xinbo Gao, Leiyu Sun, and Jie Li. Bayesian face sketch synthesis. *IEEE Transactions on Image Processing*, 26(3):1264–1274, 2017.

[Zhu *et al.*, 1997] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

[Zhu *et al.*, 2017a] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*. ICCV, 2017.

[Zhu *et al.*, 2017b] Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Li. Deep graphical feature learning for face sketch synthesis. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3574–3580. AAAI Press, 2017.