

Studies Advanced in Image Style Transfer based on Deep Learning

Yifei Shen^{1, a, †}, Guo Tang^{2, b, †}, Qiaoyu Xu^{3, c, †, *}

¹ Beijing Normal University - Hong Kong Baptist Union International College, Zhuhai, China

² The Chinese University of Hong Kong, Shenzhen, Shenzhen, China

³ University of New Brunswick, New Brunswick, Canada

^a q030026123@mail.uic.edu.cn, ^b 119010277@link.cuhk.edu.cn, ^{c, *} Qiaoyu.Xu@unb.ca

* Corresponding Author Email: Qiaoyu.Xu@unb.ca

[†]These authors contributed equally

Abstract. Image style transfer (IST) is a hot topic in the computer vision community, which refers to learning the distribution of a given style image to convert any image into corresponding image style while the content of the original image is preserved as much as possible. Early style transfer mainly utilizes texture features. Thanks to the great improvement of deep learning technology, researches on IST based on convolutional neural networks (CNN) have achieved breakthroughs in accuracy and speed. Focusing on the topic of deep learning-based IST, we will introduce the latest algorithms in detail, including their basic ideas, key steps, advantages, and disadvantages. Also, we will give an analysis of the performance of representative methods. Furthermore, we discuss the problems to be solved in style transfer and summarize the challenges and development trends in the future.

Keywords: Image Style Transfer; Deep Learning; GAN.

1. Introduction

Image style transfer (ITS) has always been a crucial subfield in computer vision, which refers to the conversion of images in two different domains with two different styles while remaining the content of the original image as far as possible. IST has increasingly achieved success in application sectors such as gaming scene generation, movie animation production, and human-computer interaction in recent years, drawing a large number of academic and industrial researchers.

The conventional technique for transferring an image's style is more frequently referred to as texture transfer. In 2001, Efros et al. [1] proposed a straightforward texture algorithm to create new textures by cutting and pasting various sample textures. One of the most important techniques for photorealistic rendering was the image quilting approach that they employed for texture creation and conventional image transfer. Using visual analogies, Hertzmann et al. [2] proposed in the same year a method based on analogy to generate images with novel textures. The self-similarity is primarily used in Ashikhmin et al's [3] texture transfer algorithm to increase the resolution of the final image, and it is several times faster than other techniques available at the time. While the non-parametric IST methods have demonstrated promising results, they can only extract low-level image attributes and not higher-level abstract information. When working with photos with rich colors and textures, the final impact of image synthesis is too abrasive.

Thanks to the powerful feature representation capabilities of CNN, deep-learning technology has been used in various computer vision tasks. CNNs' strong feature representation skills have allowed deep learning to be used for a variety of computer vision tasks. People learning that the deep learning technology could be applied for training models for object identification was one of the key causes for the popularity. Some of the earlier object identification models identify objects by their geometry and various sections, while others do it using color, 3D modeling, or features. What is worth mentioning in the traditional object recognition algorithm is to recognize objects according to local features. The principle is as follows, for example, our goal is to find a person in a picture. To the

program, this person is just a bunch of pixels, and if you ask the program to find it directly, it will have to compare pixels one by one and return the closest one (nearest neighbor algorithm). However, the shape and color of the object will change in the real world. If you only have this one photo at hand, the speed and accuracy of directly looking for it is too low. One researcher came up with the idea of breaking up a person's picture into smaller pieces and comparing them piece by piece (a method called Bag of Features). And the last region that has the most similar blocks is the region. The benefit of this approach is that even if one patch goes wrong, there are other patches that can be used to identify it, and the risk of error is greatly reduced.

Researchers have recently started utilizing deep learning techniques to investigate the conventional picture Style Transfer subject and have proposed Neural Style Transfer (NST). In this field, Gatys et al. [4] first put up a CNN-based IST approach in 2016. Using CNN, they discovered that the content-abstract feature representation and the style-abstract feature representation of an image can be separated. By processing these high-level abstract feature representations separately, they successfully perform IST and achieve excellent results. The work of Gatys et al. is revolutionary in the area of picture style transfer and has garnered a great deal of interest from both academics and business. CNNs for visual style transfer were first proposed by Gatys et al., and several researchers have since conducted numerous investigations on the technique. The two main components of this research are image iteration and model iteration. Prior to the development of neural networks, algorithms for transferring picture styles often worked along the same lines: they would evaluate photographs of a certain style, develop a mathematical or statistical model for that style, and then change the transferred image to better match the model. The following three photographs demonstrate how well this works, but it has a significant flaw: a software is typically constrained to one style or one setting. As a result, typical style transfer research has relatively little practical relevance. The development of neural networks made it possible for software to automatically produce images from other images.

This paper carefully evaluates the key picture style transfer methods based on deep learning, assesses the latest research results and application possibilities, discusses the difficulties to be solved in depth, and makes some practical suggestions, laying the groundwork for further in-depth research. Finally, the future problems and development patterns are highlighted.

2. Method

Deep learning-based IST techniques primarily involve picture-based and model-based iteration. Image iteration optimizes and iterates directly on white noise images to achieve style transfer, with the image as its optimization objective. The model iteration's optimization objective is the Neural Network model. The style transfer is also implemented using a feedforward neural network. In the subsections that follow, we shall describe the representative techniques.

2.1 IST Method based on Image Iteration

The style transfer algorithms based on image iteration can mainly be divided into three categories as the three following subtitles. These three methods have the characteristics of high synthetic image quality, good controllability, easy parameter tuning, no need to train the network, long computation time, and large dependence on the pre-trained model.

2.1.1 IST with Maximum Mean Discrepancy

In 2015, Gatys initially used the VGG19 network for style transfer. Content and style are separated in CNNs, and style feature representations of arbitrary images can be retrieved by generating Gram matrices, according to the crucial finding of Gatys. Theoretically, Li et al. [5] demonstrated that the Gram matrix matches in a manner analogous to minimizing a particular maximum mean difference. The process of CNN is shown in the Figure 1.

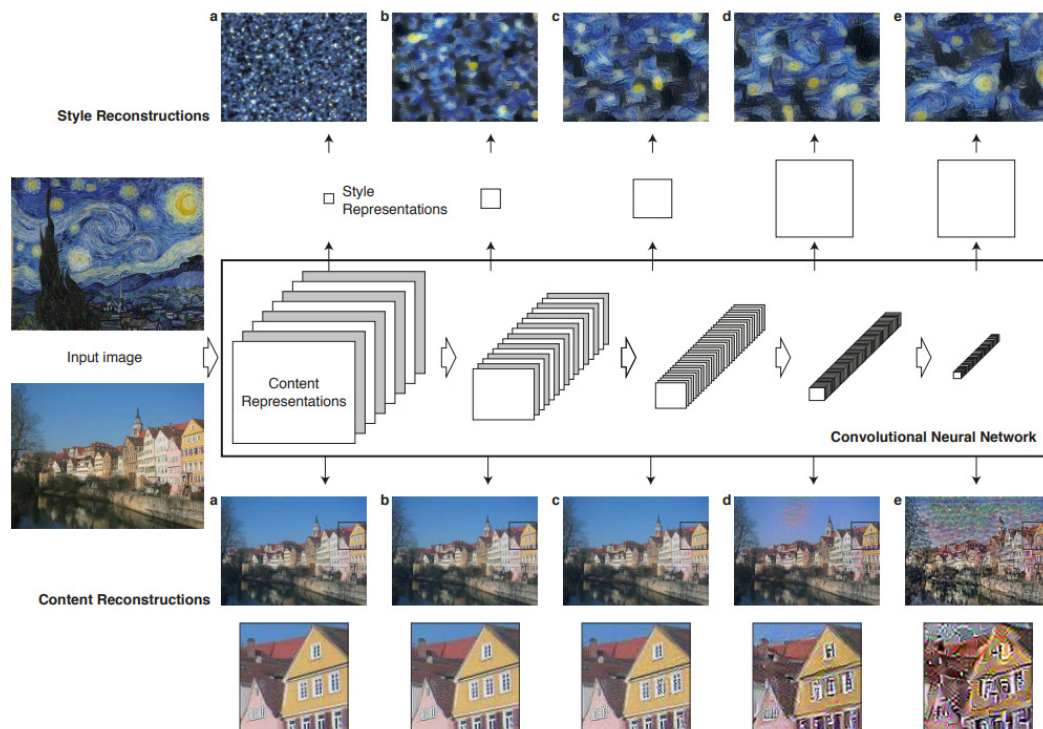


Fig 1. The Process of an Image Reconstructed in a CNN

Then, Nikulin et al. [6] deeply explored the neural algorithm of artistic and discussed the influence of different hyperparameters and stylized attributes on the effect of IST. Novak et al. suggest some avenues to improve the literature approach. Li et al. soon [7] conducted an in-depth discussion on the Gram matrix based on the literature and proposed the use of multiple kernel functions, including the linear kernel function and the Gaussian kernel function, to improve the style loss function. In addition, the literature approach is unstable during iterative optimization, making texture generation susceptible to disruption. Risser et al. [8] created the histogram loss function to solve the issue of image texture disorder resulting from the instability of the iterative optimization technique. Yin suggested a content-aware approach that efficiently controls the composition of image content and image texture, thereby enhancing the resolution of the synthetic image.

2.1.2 Style Transfer with Markov Random Field

When CNN's pooling operation compresses picture data, a large amount of data is lost. Gatys et al. employed a filter pyramid in the VGG network to represent the image at a high level. However, separate convolutional layers can only record the connection between every pixel in an image, not its spatial distribution. Li et al. [9] suggested a model integrating Markov random field and CNN, replacing Gram matrix matching in the Gatys model with the Markov regularization model, in response to the aforementioned issues. The Markov model defines a collection of features belonging to the same class; hence, partitioning the CNN image feature map into several regional blocks and matching can improve the visual rationality of the synthesized picture. This technique offers an intriguing expansion to image-based iterative style transfer, but it has significant limitations: (1) The input image is subject to a stringent restriction: the style image must be reconstructible by Markov. Images having strong transmission structures, for example, are incompatible with the proposed technique. In comparison to the original image, the resultant image has blurred edges. This is due to the photos lost during network training. In conclusion, the style transfer method based on Markov Random Field will give good outputs only when the form of the input image and the style image are comparable and there is no significant change in angle and size.

2.1.3 Style Transfer with Deep Image Analogy

Hertzmann et al. were the first to propose the concept of picture analogy. It focuses mostly on the topic of image mapping relationship deep mining. Liao et al. [10] coupled the concept of picture analogy with deep learning to present a depth image analogy approach that is iteratively optimized by area block matching in order to uncover semantically relevant dense correspondences between two input images. This method can apply the concept of image analogy to the feature space of a deep neural network to discover dense semantically significant correspondences, hence enhancing the efficiency of IST.

Liao et al. generated impressive results for style transfer by integrating deep learning (VGG19) with picture analogy (PatchMatch). PatchMatch is a classical algorithm for image analogy. Given a set of images A and B, A provides semantic and content information, and B provides appearance (e.g., color, lighting, style, etc.) and detail (e.g., texture) information. Both images are divided into fixed size pixel blocks. When the pixel block is small enough and all the pixels are replaced and smoothed, the style transfer result map of A with the style of B 'is obtained immediately. As depicted in Figure 2, Liao et al. first trained a VGG19 network to get the picture's features using deep learning for preparation. Then, they expanded PatchMatch from the picture domain to the feature domain, directing the transfer of visual attributes at the semantic level. The algorithm is relatively accurate. However only when transferring semantically similar content the result is acceptable and it lacks generalization ability.

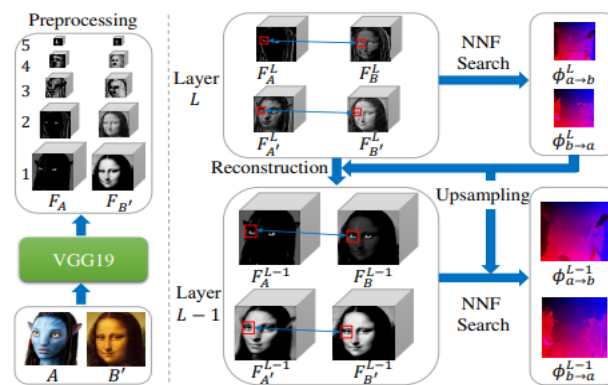


Fig 2. The pipline of the system

2.1.4 Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

Due to its ingenious loss function design method, the effect of GAN network is leading the way in the field of computer vision, and researches on GAN network also receives great popularity and attention. It also has been established in recent years.

The difference between GAN network and traditional network lies in the design of loss. The loss function of the Gatys' model employs the VGG network to extract the characteristics of style photos and target images as labels. The features of the output image are extracted to make the result, and the difference between the two is compared as the loss. The loss design of GAN network is ingenious. GAN network consists of two networks: generator and discriminator. The generator is responsible for reconstructing the input image. Together with the real data set, the reconstruction results are sent to the discrimination network for judgment. The discrimination network is responsible for determining whether the output results of the generator are real images from the real data set or generated images from its own. At this point, a tricky thing happens. The generator will try to fool the discriminator with its generated images. The more realistic the generated images, the easier it is to mislead the discriminator. The discriminator then determines whether the input is real or generated, and its feedback guides the generator to generate more realistic images, resulting in a dynamic game where the output images are similar in content and style to the real dataset.

One thing that needs to be considered to implement style transfer with GAN is how to add content and style constraints to the GAN network [11]. The famous CycleGan variant achieves this with its

own loss design. CycleGan uses two datasets with different styles and optimizes the adversarial loss to achieve the purpose of style constraint. This process requires two networks consisting of a generator and a discriminator. CycleGan achieves the purpose of content constraint through a reverse change. Similarly, this process also requires two networks of generator and discriminator. Consequently, the network components of CycleGan consist of two generators and two discriminators, whereas the loss component consists of adversarial loss and content distance loss between the original input and output after inverse transformation. Due to the GAN network's great performance in the field of picture production, it is also commonly utilized in style transfer, and many GAN-based variations are capable of achieving the desired style transfer effect.

2.2 Style Transfer Algorithm based on Model Iteration

Although methods based on image iteration can produce images with good style synthesis effects, they are inefficient. The model iteration-based picture style transfer method trains the generative model using vast amounts of data. Per-Style-Per-Model Methods and Multi-Style-Per-Model Methods constitute the majority of the model-iterative style transfer algorithm.

2.2.1 Per-Style-Per-Model Methods

Due to the great number of iterations and the long iteration time of IST methods based on picture iteration, rapid style transfer methods have been proposed to address the speed issue. Perceptual Losses for Real-Time Style Transfer and by Justin Johnson et al. propose the most representative and innovative model-based iterative approach to IST. The real-time style transfer process suggested by Super-Resolution employs the IST network as a model. This method is also called as the Fast Style Transfer method due to its rapid generation speed. The IST network is also known as the fast style network[12].

The main idea of this approach is that for each style image, the researcher trains a feedforward network at a specific training point, so that when we test the network, we just need to input an original image into the model and perform a single feedforward calculation to obtain the output. Compared with the model proposed by Gatys, the input is no longer the noise background map but the target image, and an autoencoder-shaped forward network is added to fit the process of IST. Other ideas are same as the former one: The loss function is still content loss+style loss extracted by another neural network (usually VGG), and collectively referred to as perceptual loss. The texture modeling is expressed using Gram matrix.

2.2.2 Mutil-Style-Per-Model Methods

The single model fast style transfer algorithm trains a model for each style. This works for small classes but having thousands of classes required would require thousands of models, which is not acceptable for industrial implementations. In fact, the need to integrate multiple styles into a single model is important. Training a model for each style is a redundant, unjustified waste of resources. The solution to this concept is to identify the common components between multiple style networks and then modify just the unique components for the new style while leaving the shared component unchanged. This is the fundamental concept behind the MSPM algorithm developed by Google Brain. A Scholarly Illustration of Artistic Style. [13] They found that, given a trained stylization network, by simply applying an affine transformation to the Instance Normalization layer (which they called Conditional Instance Normalization, or CIN), a result with a completely different style can be obtained. But the reason why CIN layers work is also not rigorously derived now. One possible explanation is that CIN is capable of performing one style normalization, which directly normalizes the style of an image to another style. Another MSPM method StyleBank proposed by Dongdong Chen: An Explicit Representation for Neural Image Style Transfer is similar to the Google Brain paper [14]. The main idea is to separate out the middle layers of the style network (called StyleBank in the paper), bind them to each style, and train only the middle layers for each new style, leaving the rest unchanged.

3. Experiment

As style is not a well-defined term to have a cognitive definition to analyze the changes from A style to B style. Analyzing and evaluating the results of style transfer is hindered by limited quantitative analysis methods that are massively used in other computer visual fields. Therefore, this section does the common and reliable way to do the evaluation is by human evaluation of the visual results [15].

3.1 Comparison between Algorithms based on Maximum Mean Discrepancy

By comparing the performance of the 5 algorithms based on image iteration, CNN, and VGG16 network as mentioned in the 2.1.1 section, the results are transferred from realistic pictures to some oil paintings to explore the difference in the performance of this application, which can be seen in Figure 3.



Fig 3. Comparison between Style Transfer with Maximum Mean Discrepancy

Overall, every algorithm has good preservation of the content. There is hardly a loss of a semantic part, for instance missing the ear of the cat's ear or the windows of a building, which could be caught by human eyes. On the other hand, those algorithms change the content input according to the texture of the style reference to various degrees. Nikulin's algorithm and Risser's algorithm would change the colour of the content picture based on the style picture massively. For Gatys, Novak and Yin's work, it is easy to tell that the colour of results preserved more from the original content input. Besides colour, there are preferences on how the five algorithms imitate how the oil painting artist would layer each stroke to further change the style of the content image. Overall, all result seems to successfully capture the stroke pattern of the style image. Nikulin, Risser and Yin's method stands out to have a more natural touch of human painting. It is convincing to tell someone that those three results are painted by humans rather than synthesized by a computer program. Partially the reason that some results are looked more similar to human painting is the match of the stroke pattern with the semantic parts in the image. This idea is heavily focused on Yin's work where the image would be looking separate into some region of semantic parts. The algorithm would match each semantic parts from the content input to the style input. In the idea that Yin's team proposed, it is to treat each semantic part of the content input and style input as a pair of content picture and style picture, then have the result from each pair of semantic parts combine into the result. [16] To have this focus on matching semantic meaning and treating each semantic part accordingly would avoid the texture of the sky leaking to the region of the building to satisfy the expectation of someone on each semantic part. For example, one could expect the texture of the sky from the style image, *The Starry Night*, to appear in the sky of the Gatys' result, but the texture of the sky appeared in the building of the Gatys' result as well, which is unexpected that the building and the sky of *The Starry Night* have a different texture. In the way that how Gatys and Novak apply the texture of the style image, it works more like a filter in photo editing software that works on a global degree of the whole image. The advantage of Nikulin's and Yin's work can do well on each semantic part close to how the artist would draw in style image as expected. For example, for Risser's result, the stroke pattern for the hair and face of the result matches close to the hair and face of the style image accordingly. The face would have a wider stroke which appeared on the forehead of the result and style image. Meanwhile, the hair would have smoother, thinner, and longer strokes. Apart from the stroke pattern, the colour matches mostly in each part of the semantic meaning as well. For example, the darker red appears on the girl's cheek, and the lighter red appears on the girl's forehead, which both match the style image.

3.2 Comparison between Style Transfer with Image Iteration and Model Iteration

In this section, the comparison across image-iteration based algorithm and model-iteration based will be compared on the application from a realistic photographic to a drawing. The results from Risser's team, Li's team and Liao's team are categorized as image-iteration-based algorithms in 2.1. The results from Johnson's team and Dumoulin's team are categorized as model-iteration-based algorithms in 2.2. As shown in Figure 4, all algorithms capture the colour of the style image well, and the results are all towards the colour of the style image except Dumoulin's result. Dumoulin's work is categorized as Mutil-Style-Per-Model Fast Neural Method in 2.2.2 to have the results closest to the colour of the content image. Liao's work stands out to have the best colour accuracy to have the same light and shadow effect from the left face to the right face. Johnson's work seems to have very less colour adapted from the style image with the result having all the yellow and brown colours on a global degree, but the style image clearly has more colour distribution than the results. On texture-wise, an interesting trend happened that image-iteration-based algorithms have results with a texture more toward style image, however, model-iteration-based algorithms have results that preserved more texture from the content image. For instance, by looking at Johnson's result, it is obvious that the girl's hair in result has a more realistic hair texture as the content image rather than the hand painting texture as the style image compared to Li's and Risser's result. Liao's result has a very pleasant imitation of a detailed drawing pattern. For example, Liao's method has long strokes on the shirt collar and face which matches how the artist of style image how draws on those semantic parts.

3.3 The Difference between the Classic Gatys' Method and Methods based on GAN

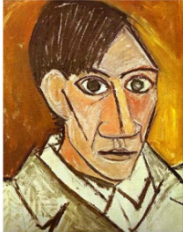
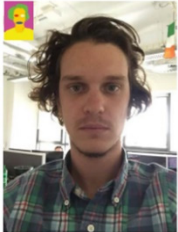
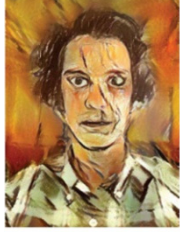
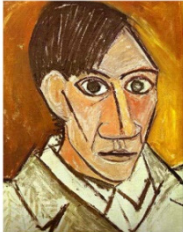
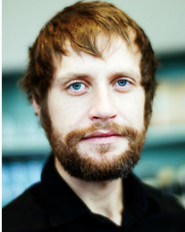
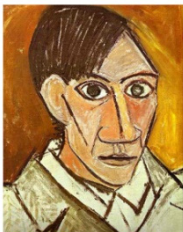
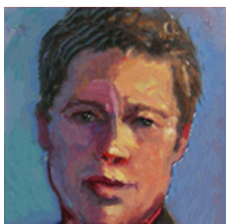
Method	Style image	Context image	transfer results
Risser (2017)			
Li (2016)			
Johnson (2016)			
Liao (2017)			
Dumoulin (2017)			

Fig 4. Comparison between algorithms based on image iteration and model iteration

In this section, results from GAN algorithms (2.1.4) and the classic Gatys' CNN style transfer algorithms as the picture shows. Since Zhu's method does not use a style image as a reference compared to other methods mentioned in 2.1 and 2.2, the result would not be highly focused on style transferring, but rather on the tasks. As shown in Figure 5, Gatys' method works on a more global scale to create a more dramatic effect on changing into art style, at the same time, Zhu's project changes the style of the objects that, the task focuses on such as changing horses into zebras and leave the rest of the image as backgrounds, to get a more realistic result. In comparison to Gatys' method, the task performed poorly on changing the colour of the horses partially on mostly legs and the colour of the sky unreally. In the task of changing apples to oranges, both methods have similar results, since the image is full of apples to have almost no background to avoid the drawback of Gatys' method that is just mentioned.

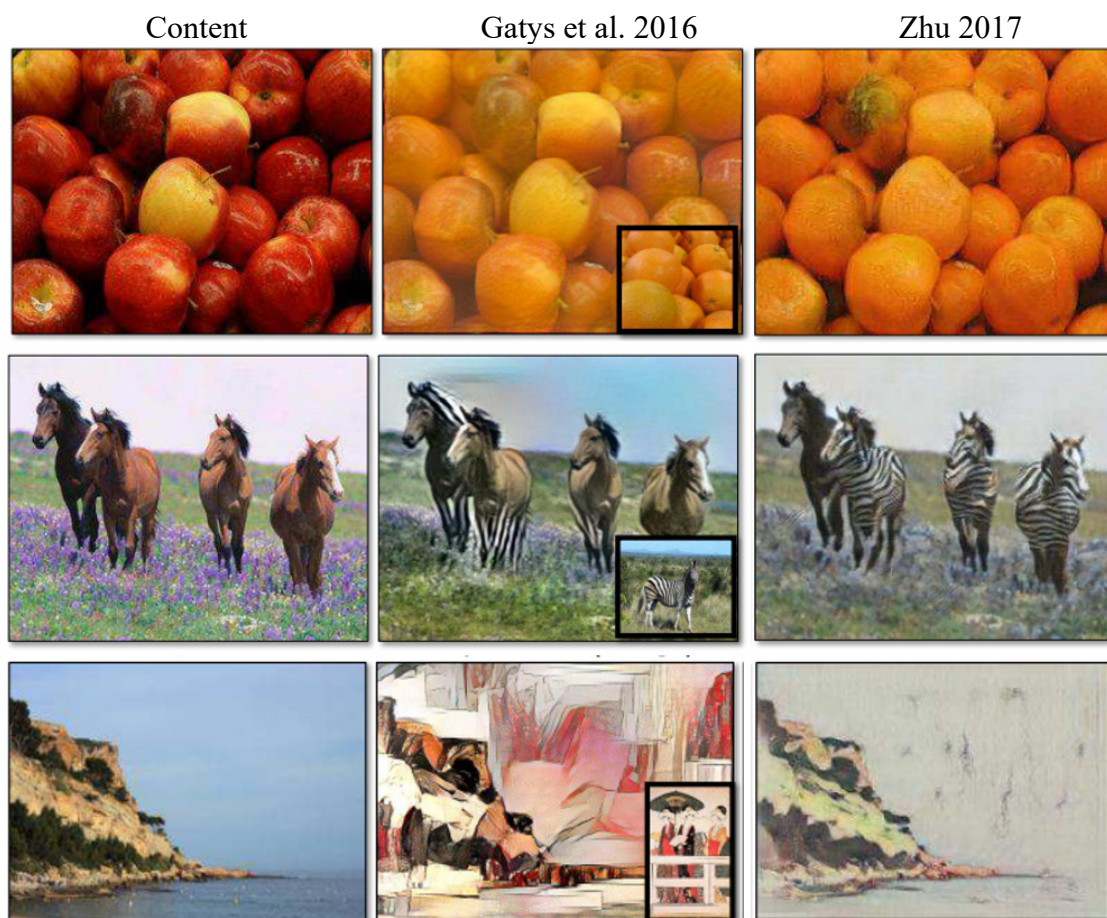


Fig 5. Results from Gatys 2015 and Zhu 2017

4. Discussion

Progress in the area of IST based on deep learning is encouraging, and some algorithms have been used in industrial applications. Although the current algorithm performs well in general, there are still some challenges and outstanding problems.

All kinds of style transfer algorithms have their own shortcomings. For the current image style conversion technology, it can realize the style image with high color contrast or clear lines, or the image style information is obvious, such as the painting with a variety of saturated color lines and clear lines, better combined in the image. However, there will also be problems due to image contrast, clarity, and texture quality, resulting in the training of the style of the effect is not very obvious. For high resolution images, the detail is not obvious. For low resolution images, the result is a lot of blurs and drag.

Another notable disadvantage of neural networks is their black-box nature. In other words, from receiving input to producing output, we only know what the input is and what the output is. We have no way of knowing why and how this output is arrived at, and it is hard to see his intermediate steps. Therefore, for such an algorithm, its interpretability is low, so it is usually difficult to understand the cause of looking for errors.

We believe that the image style conversion based on deep learning in the future can try to combine different machine learning methods, such as using the reward mechanism of reinforcement learning to strengthen the feature extraction of the style and combining transfer learning to use the network parameters of large data tasks to small data tasks, to improve the efficiency and accuracy of training.

5. Conclusion

In this paper, focusing on IST based on deep learning, we introduce the latest research progress in this field in detail from image iteration method and model iteration method, including the basic design ideas, key steps, advantages, and disadvantages of representative algorithms. We also qualitatively compare the transfer visualization results of different algorithms to discuss the application boundaries of different algorithms. Finally, our discussion presents some shortcomings of the topic of IST and how to potentially improve them.

References

- [1] Efros A A, Freeman W T. Image quilting for texture synthesis and transfer[C]//Proceedings of the 28th annual conference on Computer graphics and interactive techniques. 2001: 341-346.
- [2] Hertzmann A, Jacobs C E, Oliver N, et al. Image analogies[C]//Proceedings of the 28th annual conference on Computer graphics and interactive techniques. 2001: 327-340.
- [3] Ashikhmin, N. (2003). Fast texture transfer. *IEEE Computer Graphics and Applications*, 23(4), 38-43.
- [4] Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2414-2423).
- [5] Li Chuan, Wand M. Combining Markov random fields and convolutional neural networks for image synthesis [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ: IEEE Press, 2016: 2479-2486.
- [6] Nikulin Y, Novak R. Exploring the neural algorithm of artistic style[J]. *arXiv preprint arXiv:1602.07188*, 2016.
- [7] Li Y, Wang N, Liu J, et al. Demystifying neural style transfer[J]. *arXiv preprint arXiv:1701.01036*, 2017.
- [8] Risser E, Wilmot P, Barnes C. Stable and controllable neural texture synthesis and style transfer using histogram losses[J]. *arXiv preprint arXiv:1701.08893*, 2017.
- [9] Li Chuan, Wand M. Combining Markov random fields and convolutional neural networks for image synthesis [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ: IEEE Press, 2016: 2479-2486.
- [10] Liao J, Yao Y, Yuan L, et al. Visual attribute transfer through deep image analogy[J]. *arXiv preprint arXiv:1705.01088*, 2017.
- [11] Bao F, Neumann M, Vu N T. CycleGAN-Based Emotion Style Transfer as Data Augmentation for Speech Emotion Recognition[C]//INTERSPEECH. 2019: 2828-2832.
- [12] Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution [C]// European conference on computer vision. Springer, Cham, 2016: 694-711.
- [13] Dumoulin V, Shlens J, Kudlur M. A learned representation for artistic style[J]. *arXiv preprint arXiv:1610.07629*, 2016.
- [14] Chen D, Yuan L, Liao J, et al. Stylebank: An explicit representation for neural image style transfer [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1897-1906.
- [15] Briakou E, Agrawal S, Zhang K, et al. A Review of Human Evaluation for Style Transfer, 10.18653/v1/2021.gem-1.6[P]. 2021.
- [16] Yin R . Content Aware Neural Style Transfer:, 10.48550/arXiv.1601.04568[P]. 2016.