Cross-domain Image Generation: Style Transfer and Image-to-image Translation

Pengfei Ran

{Rpf983405436@stu.cqut.edu.cn}

School of Artificial Intelligence, Chongqing University of Technology, Chongqing, China

Abstract. With the development of deep learning, cross-domain image generation has become an important research direction in computer vision, with broad application prospects in artistic creation and image editing. This paper summarizes the research results, principles, and applications of cross-domain image generation from two aspects: style transfer and image-to-image translation. Through comparative analysis of core technologies such as conditional generative adversarial networks (GANs), recurrent consistency GANs, and attention mechanisms, this paper outlines the development process of cross-domain image generation technology. From this can conclude that style transfer focuses on the art and image translation on the semantic maps. These two algorithms differ in loss functions, networks, and training but both suffer from generation quality, robustness, and computational cost. Through comparative analysis, this paper finds that style transfer emphasizes artistic rendering through content–style decoupling, whereas image-to-image translation focuses on semantic consistency across domains using paired or unpaired data. Although significant progress has been achieved in generation quality, diversity, and controllability, challenges remain in computational efficiency, training stability, generalization ability, and evaluation metrics. Recent advances in diffusion models further demonstrate improved generation fidelity and stability, suggesting a shift from adversarial learning to probabilistic generative modeling. Overall, this work provides a comprehensive overview of current techniques and identifies future research directions, including lightweight model design, multimodal integration, and more robust and interpretable evaluation frameworks.

Keywords: Cross-domain image generation; Style transfer; Image-to-image translation; Generative adversarial networks; Deep learning

1 Introduction

Over the past decade, the rapid development of deep learning technology has greatly propelled progress in the field of computer vision. Many tasks that used humans to perceive the images can now be tackled with high-performing neural network models. In this field, cross-domain image generation is an active research field. It aims to generate new images that preserve the semantic content of the image and translate it into another image domain. In terms of style transfer and image-to-image translation, style transfer attempts to apply the artistic style of one image to the content structure of another image, and image-to-image translation tries to translate

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365351

an image from the source domain to the target domain, maintaining semantic consistency. Image generation has become a trend due to the power of representation learning in deep neural networks, especially convolutional neural networks and generative models. With the growing need for visual data processing, image generation has also become an important area in the computer vision industry. Several technical advances have been made in cross-domain generation. The first major contribution of GANs was a 2014 novel generative model by two neural networks that train against each other in a minimax game[1]. This adversarial training provides the model with a challenging distribution and an efficient generation from random noise. GAN became a standard framework in image synthesis and motivated a lot of future work. Another significant breakthrough came in 2015 when Leon Gatys et al. proposed the Neural Style Transfer algorithm[2]. They showed that a convolutionAL neural network can separate content and style information of an image. A method that optimizes the generated image to ensure its high-level content features are consistent with the content image and its style statistical features are synchronized with the style image can generate images with an artistic style. Although the method can produce impressive results, the iteration optimization process is expensive. Despite the great advancements in cross-domain image generation, several challenges remain. For example, computational efficiency, evaluation measures, and generalization ability are still to be investigated. In addition, accurate and controllable image generation with high diversity is an open problem. Future works include fusing different generation frameworks, improving training and building a common generation model that is capable of multiple tasks, such as style transfer and style matching.

2 Style Transfer Technology Analysis and Implementation

2.1 Neural style transfer

The Gatis algorithm, proposed in 2015 by Gatis and his research team, is one of the pioneering algorithms in the field of neural style transfer and has since gained widespread recognition in subsequent research[2]. A pre-trained convolutional neural network uses features from content and style images to form a hierarchical feature representation. The novelty is a mixture of loss to re-use style and content representations. The content loss is the difference between the generated image and the content image, such that the generated result preserves the structure of the original image. In optimization, the algorithm continuously updates the pixel values of the generated images with gradient descent and gradually converges the model by weighting the content and style losses. This approach essentially optimizes the generated images to maintain content consistency with a high semantic level, but consistent with the target style in terms of statistical features. Fig. 1 shows some Images combining the content of a photograph with the style of several classical artworks

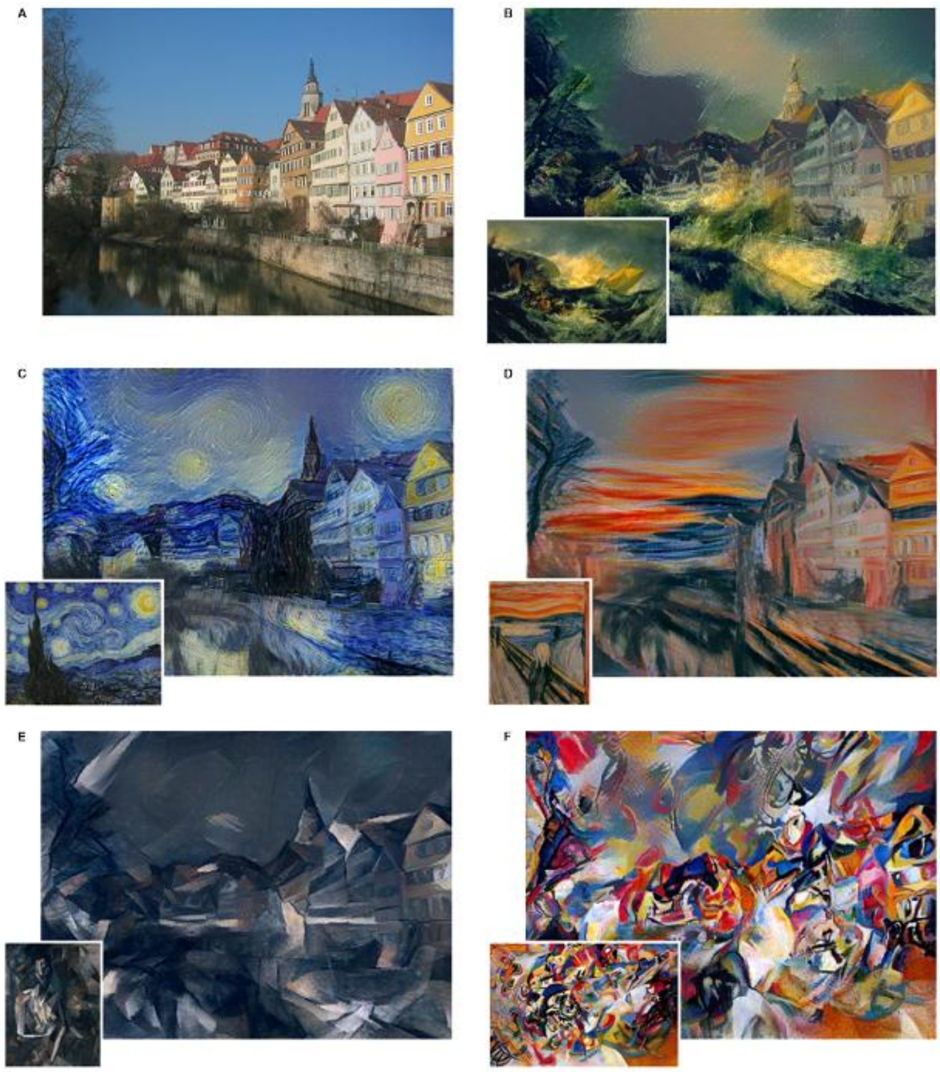

Fig. 1. Examples of photographic content rendered in classical artistic styles [2].

2.2 Classic style transfer algorithm

However, since this algorithm still relies on a large number of iterative calculations, it still has certain limitations in real-time applications. To address this issue, researchers have proposed an alternative method based on forward propagation, which can, to some extent, compensate for its shortcomings in real-time processing.

In order to solve the problem of the slow computation speed of the original Gatys algorithm, researchers have proposed a fast style transfer method based on neural networks, so that the style transfer process no longer depends on the complex iterative optimization process. In 2016, Justin Johnson et al. proposed a transfer network structure [3], which can complete style transfer in one forward propagation, and its speed is much faster than the traditional iterative optimization method. This method uses a generator network to realize style transfer. The network structure usually consists of an upsampling layer, a downsampling layer and a residual module, which can directly map the content image to the output image with the target style and learn this mapping relationship during training. The model training adopts a perceptual loss framework. Features are extracted through the pre-trained VGG network proposed by Karen Simonyan and Andrew Zisserman[4], and the content loss and style loss are calculated, so that the transfer network can learn a stable and effective mapping relationship.

AdaIN achieves an arbitrary style transfer method that does not require style-by-style training by aligning the statistical distribution of content and style features. It achieves a good balance between efficiency and flexibility and is an important milestone in fast style transfer [5].

3 Image-to-image translation technology

3.1 Pair translation problem

In the process of cross-domain image generation, image-to-image conversion is a typical paradigm that can learn the mapping function between different visual domains. Paired and unpaired conversion problems are key to building a conversion framework. Paired conversion requires the use of datasets that can correspond one-to-one with source domain images to target domain images, forming a one-to-one relationship. Pix2Pix is a typical supervised learning method, first proposed in 2017. It uses conditional generative adversarial networks to learn the mapping function [6]. With the help of paired training data, the model can accurately grasp the correspondence between pixels and produce excellent conversion results with strong structural consistency.

In the process of data pairing, there are many constraints that cannot be solved by technical means and the cost is relatively high. This challenge has prompted scholars to explore unpaired translation methods, which can obtain domain mapping without corresponding information.

3.2 Non-paired translation

CycleGAN has made a breakthrough in the field of image transformation, eliminating the need for paired data for training and solving the most challenging problem in the development of this technology. Zhu et al. pointed out that this technology introduces cycle consistency loss in the regularization process, so that the mapping between two domains can be achieved without direct correspondence. The most fundamental innovation of this technology is that it can train two generator networks at the same time, one mapping the source domain X to the target domain Y, and the other transforming Y into X in the reverse [7]. Cyclic consistency constraint requires that after the image is transformed to the target domain, it can be transformed back to the original image, which is a reconstruction requirement in mathematics. This statement is relatively reasonable, enabling the model to obtain cross-domain relationships from the unpaired image set, and also enabling image transformation technology to be more widely used in various scenarios.

StarGAN proposes a unified multi-domain image translation framework [8]. By introducing domain labels and combining adversarial loss, classification loss and cycle consistency loss, it achieves flexible conversion between multiple domains for a single model compared to methods such as CycleGAN, significantly improving efficiency and scalability. StarGAN v2 introduces a style encoder and mapping network on this basis, realizes multimodal generation by modeling the latent style space, and improves the realism and diversity of the generated results by using style reconstruction and diversity constraints [9]. MUNIT starts from the perspective of content and style decoupling, represents images as shared content and domain-specific styles, and realizes diversified outputs by randomly sampling style vectors, breaking through the one-to-one mapping limitation [10]. These methods have jointly promoted the development of multi-domain and multimodal image translation.

3.3 Diffusion: A New Era of Image Generation

Denoising Diffusion Probability Model (DDPM): a forward progressive noise addition and reverse denoising that transform a complex data distribution into a set of conditional probability modeling problems, which do not train GANs and achieve good performances in terms of generation quality and diversity [11]. Latent Diffusion Model (LDM): which moves the diffusion process from pixel space to low-dimensional latent space, allows a low-resolution and conditional generation that enables high-resolution and conditional generation, and is also used to derive Stable Diffusion [12]. Based on this, Nichol and Dhariwal improved DDPM by learning the variance of the reverse process, introducing a mixed loss, and optimizing the sampling strategy with much greater log likelihood and sampling efficiency, as well as verifying the advantages of the diffusion model in terms of distribution quality in terms of precision and recall [13]. These works lead diffusion models towards efficiency and practicality. Diffusion models offer a new concept of image translation, which is different from GANs.

Overall, this approach drives image translation from adversarial learning to stable, high-quality, and controllable generative modeling for complex scenes, which can yield a more reliable solution for cross-domain image generation in complex scenes.

4 Conclusion

This paper explores the development, core mechanisms and applications of cross-domain image generation technology. First, from a neural style transfer perspective, the usefulness of convolutional neural networks in content and style feature extraction and fusion, and Second, images-to-image conversion based on generative adversarial networks (GANs) can achieve high-quality mapping between image domains while still retaining semantic consistency. And also summarize the main open issues of this technology and also present high computational complexity, high training cost, low inference efficiency and lack of single/objective image quality evaluation standards. Moreover, model generalization ability is also a bottleneck in practice, and when faced with unseen data or new domains, the model is not guaranteed to maintain stable performance and generative diversity. In order to address these issues, future work can include: first, develop lightweight models and train with low computational overhead and real-time performance; second, introduce attention mechanisms to facilitate model generalizing abilities to model long-distance dependency, thus improving generation quality; third, explore zero-shot or few-shot learning to improve model generalisation ability in new domains; fourth, utilize multimodal information from text and images to achieve a more controllable and intelligent generation process. Evaluation issues, including: whether there is a more comprehensive evaluation system which includes multiple dimensions such as semantic consistency, perceived quality and diversity, but also how detected and ethical standards of generated content, like watermarking technology to detect generated images, and whether there are reasonable use and supervision mechanisms to promote healthy development of cross-domain image generation technologies.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 2672-2680 (2014)

[2] Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015)

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

[4]Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations (ICLR) (2015)

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

[6] Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1125-1134 (2017)

[7] Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2223-2232 (2017)

[8] Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., Choo, J.: StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8789-8797 (2018)

[9] Choi, Y., Uh, Y., Yoo, J., Lee, J.Y.: StarGAN v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8188-8197 (2020)

[10] Huang, X., Liu, M.-Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 172-189 (2018)

[11] Ho, J., et al.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)

[12] Rombach, R., et al.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

[13] Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 8162-8171 (2021)