Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU32985

Application of Generative Adversarial Networks in Object Synthesis in Complex Background

Yangyi Mou

College of Science, Mathematics and Technology, Wenzhou Kean University, Wenzhou, China
1306078@wku.edu.cn

Abstract. Due to rapid deep learning advances, Generative Adversarial Networks (GANs) are still the best at generating images using computers. When performing object synthesis in complex scenarios, the foreground object is placed among numerous different backgrounds, enabling it to seamlessly blend in while maintaining its shape, lighting, and all surrounding elements unchanged. However, there are still some issues, such as inconsistent perspectives, differences in lighting, occlusions, and background elements, which do indeed affect the realism of their appearance. Therefore, this paper provides a review of GAN methods for generating objects in complex scenarios. By studying the foundational GAN theory and its subsequent improvements such as conditional GANs, attention-based GANs, and composite GAN frameworks, this study investigates the current relevant methods in terms of spatial alignment, illumination coordination, and multi-object coordination. And it also discusses how the authenticity and fit of this situation are achieved through the use of the generative adversarial network architecture and training methods. From the above content, it is seen that aspects like spatial attention, semantic guidance, and separating the foreground from the background are indeed very important. However, there are still limitations in terms of its universality, computational capacity, and stability. Further studies are advised to combine both geometric and hybrid generative networks to enable more realism in the synthesis of complex scenes.

Keywords: Generative Adversarial Networks, Object Compositing, Complex Background, Image Synthesis, Illumination Consistency

1. Introduction

Image synthesis has become one of the most prominent research topics in the field of computer vision, especially since the generation adversarial networks (GANs) proposed by Goodfellow et al. in 2014 [1]. GANs create a competitive situation between two models, namely the discriminator and the generator, enabling them to generate very realistic images. Since then, other architectures like DCGAN, conditional GAN (cGAN), and Pix2pix have also made significant progress, driving the development of research in the fields of image-to-image transformation and object generation [2-4]. By adopting the aforementioned methods, object synthesis in complex backgrounds means placing the foreground objects on top of a scene with the same geometric shape, lighting, and background information. Although these are more traditional image editing techniques, they are all based on

DOI: 10.54254/2755-2721/2026.GU32985

manual editing. Nevertheless, through GANs, more realistic images can be generated, but there are some issues, such as geometric alignment problems, inconsistent lighting problems, and occlusion or interference with the background [5]. This paper examines how GAN-based methods perform in object synthesis under complex circumstances. In particular, a review of relevant literature is conducted, and some representative models based on GANs are compared. The challenges faced in object synthesis in the presence of complex backgrounds are mainly discussed, including how different types of GAN can improve the realism and consistency of their generated images, and whether there are any other limitations in current methods. This research is about object synthesis based on GAN and points out the direction for future work to make it more realistic and efficient.

2. Foundations and development of Generative Adversarial Networks

2.1. Fundamental principles and mechanisms of GAN

In 2014, Goodfellow et al. proposed GANs. In their theoretical framework, the problem of image generation was regarded as a zero-sum minimax game conducted by two neural networks [1]. The generator (G) is designed to capture the underlying data distribution and generate realistic fake samples, while the discriminator (D) is optimized to distinguish between real data and the generated fake samples. Within this framework, the explicit density estimation is circumvented by optimizing this adversarial loss. Although this advantage is quite obvious, the early GANs exhibited serious training instability and mode collapse characteristics. To overcome these bottlenecks, Radford et al. proposed DCGAN, which integrates deep convolutional neural networks, batch normalization, and LeakyReLU activation functions, stabilizing the dynamics of the training process and enhancing the ability to extract features from high-resolution images [2].

2.2. Major variants and innovations

With the maturation of foundational theories, GANs became more advanced with a strong control mechanism and enhanced generation fidelity. For instance, one of the earliest improvements is the cGAN, which incorporates auxiliary variables (labels or text descriptions) into both the generator and discriminator, and can generate targeted and controllable images [3]. Based on this situation, Pix2Pix introduces a general supervised image-to-image transformation system, which is achieved by using paired datasets [4]. However, in order to overcome the very strict limitations of paired data, CycleGAN proposed a new cycle consistency loss, enabling style transformation across different visual domains to be carried out without supervision [5].

In addition, StyleGAN will transform these higher-quality synthesis systems into specific styles through adaptive instance normalization methods. This approach separates the larger spaces and textures [6]. To further boost the overall consistency and solve the problem of small receptive fields in conventional convolutional layers, the Self-Attention GAN (SAGAN) was proposed, enabling the model to grasp the long-term dependencies and structures of distant areas in the image [7].

2.3. Applications in object compositing

When applied to images, merely generating isolated realistic objects is not sufficient. The models should also be able to combine these components with the current scene. Adversarial learning is initially proposed on the understanding of structural context in image restoration through "context encoder" [8]. Subsequently, the Layered Recursive GAN (LR-GAN) proposes a specialized process

DOI: 10.54254/2755-2721/2026.GU32985

for separately generating the background and foreground parts, which is the first attempt towards structured scene combination [9]. Based on these concepts, later methods emphasize the interactions between object pairs. For instance, the compositional GAN explicitly points out the spatial and structural relationships of independent objects [10]. Moreover, realistic image combinations based on specific frameworks still mainly rely on adversarial learning. This approach ensures proper boundary fusion and global color coordination, enabling the generative adversarial network not only to perform pure image generation but also to execute complex and realistic editing tasks [11].

3. Difficulties in object compositing against a complex background

3.1. Geometric matching and perspective matching

When precisely placing the foreground objects in a complex scene, it is extremely challenging to maintain the continuity of the space and the accuracy of the geometry. And this is mainly because precise adjustments need to be made to their size, position and perspective. Even the slightest difference can quickly undermine the overall sense of realism, making the objects appear unstable. Especially in the background part, which is usually more structured, such as an image of a city street or an interior scene of a room. Due to reasons like the internal and external camera parameters of the source object and the target scene, spatial errors may occur. To properly handle these spatial relationships, one needs to have a deep understanding of three-dimensional geometry, including the orientation of objects, depth, and the layout of the scene. Purely two-dimensional convolutional networks mainly focus on nearby pixels, which makes them difficult to understand the geometry of distant objects. Thus, it is very difficult to correctly align objects when there is a large mixture of objects [10].

3.2. Lighting harmonization and texture mixing

In order to achieve realistic lighting and unified texturing, the hardest part of compositing an object is dealing with lighting differences. Different lighting directions, color temperatures, ambient light and other factors can all result in significant differences in illumination. This can be easily noticed. For example, placing a person photographed under controlled lighting into an outdoor dusk scene is not straightforward. Currently, this process typically begins by predicting the light conditions of the scenario, and then conducts adversarial training to update and optimize the model's understanding of the environment [12]. In addition, texture mixing around an object's edges must consider the object's details and how well it fits with the background's noisiness. If handled improperly, visible "cut-and-paste" artifacts can appear, making the composition look less realistic [11].

3.3. Background interference and occlusion issue

In the face of issues such as background interference and occlusion, there is another challenge in object synthesis in the real world. That is, in real scenes, there are usually overlapping objects and complex semantic information, which makes it difficult to determine whether the new object should be placed in front of or behind the existing elements. And if not properly arranged, the original topological layout may become disordered. In more complex situations, the result may even be unrecognizable. Good composite materials require a powerful foreground-background separation model, which can correctly distinguish depth and spatial relationships, and simultaneously preserve contextual information [13,14]. By directly modeling occlusion boundaries and object hierarchy, it is

DOI: 10.54254/2755-2721/2026.GU32985

possible to ensure that the inserted objects can interact naturally and realistically with the other parts of the scene.

4. Effective solutions for object compositing gans

4.1. Spatial integration and object embedding

In order to maintain the geometric shape and perspective effect of the object unchanged, the object must be placed in the correct position. To solve these problems, more and more research has begun to integrate spatial transformation networks (STN) and attention modules into GANs. This has led to improvements in spatial reasoning in attention-based GANs, as computing attention specifically focuses on the boundaries and overlapping areas of the object [7].

Furthermore, models such as compositional GANs can also calculate the parameters of bounding boxes and the relative spatial positions, enabling the object to interact physically and logically in the target scene. These mechanisms can take into account overlaps and spatial limitations, significantly improving the placement effect of the object and helping to maintain a reasonable scene structure [10].

4.2. Lighting and texture consistency optimization

For the combination of real objects, in order to maintain consistent lighting and texture, employing an encoder-decoder architecture with adversarial loss can enable the objects to match the color and lighting of the target scene, as achieved in the Depth Image Coordination framework [15]. This is an improved version of some multi-task generative adversarial networks, which includes multiple auxiliary streams to estimate the lighting and shadow trajectories. To consider environmental cues, the generator should be able to generate synthetic images with appropriate lighting and consistent texture, to ensure that the objects and their surrounding scenes do not appear significantly different from each other [16].

4.3. Background coordination and multi-object compositing

To achieve semantic and spatial coherence, the background elements must be coordinated with each other, and multiple objects need to be managed in complex scene combinations. This method has been proven effective in handling occlusion and highly structured backgrounds. SPADE is a notable approach that employs spatial adaptive normalization and is also semantic normalization, ensuring that foreground texture patterns are not generated in inappropriate locations [14]. The separated foreground-background GAN first separates the potential objects, backgrounds, and synthetic masks, and then recombines them to control the interaction between the objects and the background.

On the other hand, frameworks such as FBC-GAN can generate various foreground-background combinations and maintain semantic accuracy [17]. Although GANs dominate this field, recently, diffusion-based generative models have demonstrated strong capabilities in achieving global scene coherence, opening up a promising direction for improved object synthesis solutions [18].

5. Conclusion

This paper explores the application methods of GANs in complex scenarios for object combination. And the framework that combines adversarial learning, conditional control, spectral attention, and precise foreground-background segmentation makes automated image editing appear more realistic.

DOI: 10.54254/2755-2721/2026.GU32985

It also improves geometric calibration, uniformity of lighting, and reduces background interference, which remain bottleneck issues in practical scenarios. The results demonstrate that the improved GAN model has made progress in addressing these challenges. The spatial attention mechanism and semantic-guided normalization like SPADE perform exceptionally well, solving perspective issues and context anomalies. The deep image coordination system using adversarial loss also reduces differences in lighting, color, and texture, generating more coherent synthetic images in perception. However, generative adversarial networks inherently have adversarial issues during training, which can lead to insufficient diversity in the generated results. High-resolution synthesis requires precise fusion and complex attention models, which are computationally demanding and make real-time implementation difficult. Numerous models are trained based on specific distributions and perform poorly in zero-sample synthesis of novel or highly non-structured scenarios. Future improvements cannot solely depend on two-dimensional pixel operations. In particular, incorporating explicit 3D geometric reasoning and physically based rendering (PBR) into the generation process is crucial for achieving consistent lighting and perspectives. Relying on stable and consistent diffusion-based models, the hybrid GAN-diffusion method also shows potential. In practical applications such as film production, augmented reality (AR), and automated digital art generation, it is necessary to improve efficiency, robustness, and zero-sample generalization capabilities.

References

[1] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.
[2] Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv: 1511.06434.
[3] Gauthier, J. (2014). Conditional generative adversarial nets for convolutional face generation. Class project for Stanford CS231N: convolutional neural networks for visual recognition, Winter semester, 2014(5), 2.
[4] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 1125-1134.
[5] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, 2223-2232.
[6] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 8110-8119.
[7] Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019, May). Self-attention generative adversarial networks. In International conference on machine learning, 7354-7363.
[8] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2536-2544.
[9] Yang, J., Kannan, A., Batra, D., & Parikh, D. (2017). Lr-gan: Layered recursive generative adversarial networks for image generation. arXiv preprint arXiv: 1703.01560.
[10] Azadi, S., Pathak, D., Ebrahimi, S., & Darrell, T. (2020). Compositional gan: Learning image-conditional binary composition. International Journal of Computer Vision, 128(10), 2570-2585.
[11] Chen, B. C., & Kae, A. (2019). Toward realistic image compositing with adversarial learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 8415-8424.
[12] Gardner, M. A., Sunkavalli, K., Yumer, E., Shen, X., Gambaretto, E., Gagné, C., & Lalonde, J. F. (2017). Learning to predict indoor illumination from a single image. arXiv preprint arXiv: 1704.00090.
[13] Ni, J., Zhang, S., Zhou, Z., Hou, L., Hou, J., & Gao, F. (2021). Background and foreground disentangled generative adversarial network for scene image synthesis. Computers & Graphics, 97, 54-66.
[14] Park, T., Liu, M. Y., Wang, T. C., & Zhu, J. Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2337-2346).
[15] Tsai, Y. H., Shen, X., Lin, Z., Sunkavalli, K., Lu, X., & Yang, M. H. (2017). Deep image harmonization. In Proceedings of the IEEE conference on computer vision and pattern recognition, 3789-3797.

DOI: 10.54254/2755-2721/2026.GU32985

[16] Li, X., Teng, G., An, P., & Yao, H. Y. (2023). MT-GAN: toward realistic image composition based on spatial features. EURASIP Journal on Advances in Signal Processing, 2023(1), 46.
[17] Cui, K., Zhang, G., Zhan, F., Huang, J., & Lu, S. (2021). FBC-GAN: Diverse and flexible image synthesis via foreground-background composition. arXiv preprint arXiv: 2107.03166.
[18] Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 34, 8780-8794.