Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/57/20241310

Keywords-based conditional image transformation

Jincheng Liu

Faculty of Data Science, City University of Macau, Avenida Padre Tomás Pereira Taipa, Macau.

D21090103791@cityu.mo

Abstract. In recent years, Generative Adversarial Networks (GANs) and their variants, such as pix2pix, have occupied a significant position in the field of image generation. Despite the impressive performance of the pix2pix model in image-to-image transformation tasks, its reliance on a large amount of paired training data and computational resources has posed a crucial constraint to its broader application. To address these issues, this paper introduces a novel algorithm, Keywords-Based Conditional Image Transformation (KB-CIT). KB-CIT dynamically extracts keywords from the input grayscale images to acquire and generate training data, thus avoiding the need for a large amount of paired data and significantly improving the efficiency of image transformation. Experimental results demonstrate that KB-CIT performs remarkably well in image colorization tasks and can generate high-quality colored images even with limited training data. This algorithm not only simplifies the data collection process but also exhibits significant advantages in terms of computational resource requirements, data utilization efficiency, and personalized real-time training of the model, thereby providing new possibilities for the widespread application of the pix2pix model.

Keywords: Generative Adversarial Network (GAN), pix2pix, Image Generation, Image Colorization.

1. Introduction

With the rapid development of artificial intelligence, generative models have gradually emerged as a focal point of research. These models aim to comprehend the inherent patterns and distributions of data, intending to generate new samples that closely resemble real data. Among them, Generative Adversarial Networks (GANs), first proposed by Goodfellow et al. in 2014 [1], have become one of the most popular and effective generative models. GANs learn and enhance their capabilities through the interplay of two neural networks, the Generator and the Discriminator. The Generator samples from a latent space to create realistic data, while the Discriminator endeavors to differentiate whether the data it receives is real or generated by the Generator. They engage in mutual adversarial learning and optimization, ultimately rendering the Discriminator unable to discern the authenticity of the Generator's output, achieving a realistic rendering effect. The outstanding performance of GANs is not only manifested in image, sound, and video generation but also finds wide applications in multiple domains such as style transfer and super-resolution [2-4].

Although original GANs exhibit powerful generative capabilities, they encounter various issues in practical applications, including training instability, mode collapse, and insufficient optimization for specific tasks. To overcome these challenges, researchers have innovated and extended the original

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/57/20241310

GAN structure, giving rise to a series of variant GANs. These variants not only optimize the shortcomings of the original GANs but also provide more refined control for specific application scenarios. This paper primarily focuses on some milestone variants in the development history of GANs that are widely recognized by researchers.

DCGAN [5], introduced by Radford et al. in 2015, optimized the original GAN architecture, particularly in terms of network architecture, training stability, and generated sample quality. Its main innovation lies in the use of Convolutional Neural Networks (CNN) to construct the Generator and Discriminator, abandoning traditional pooling layers and employing strided convolutions and transpose convolutions to preserve spatial information, thus avoiding information loss. Furthermore, to enhance training stability, DCGAN introduced Batch Normalization in most layers and carefully selected activation functions to ensure a more stable network performance. These design improvements resulted in significant improvements in the quality and diversity of image generation. However, DCGAN still faces challenges in generating high-definition or high-resolution images, and its training process may encounter instability, requiring careful balance and adjustment.

Wasserstein Generative Adversarial Network (WGAN) [6], proposed by Martin Arjovsky and his team in 2017, is a significant milestone in the research field of Generative Adversarial Networks (GANs). It successfully addresses two major challenges faced by original GANs: training instability and mode collapse. The uniqueness of WGAN lies in the introduction of Wasserstein distance as a new metric, enhancing training stability and improving the quality and diversity of generated samples.

LAPGAN [7], introduced by Emily Denton et al. in 2015, is a model designed to overcome the limitations of traditional GANs in generating high-resolution images. Through a hierarchical pyramid structure, LAPGAN can effectively generate high-resolution, high-quality images and alleviate common mode collapse issues. Unlike a single Generator and Discriminator, LAPGAN adopts a multi-level Generator and Discriminator structure to finely capture the various detailed levels of the image.

CGANs [8], proposed by Mehdi Mirza and Simon Osindero in 2014, are a variant of GANs aimed at generating images with specific attributes. Unlike traditional GANs, CGANs incorporate conditional information, such as class labels or text descriptions, in addition to random noise, thereby providing more targeted control over generated content. The primary advantage of this structure is its higher generative control capability, giving it a significant advantage in tasks that require the generation of images satisfying specific conditions. The introduction of CGANs not only improves the quality and controllability of images but also enhances the potential of GANs in specific generative tasks.

Pix2pix [9] is a variant based on Conditional GAN, a typical supervised image translation model. In the original GAN, the Generator begins generating images from a random noise vector, while in pix2pix, the Generator receives an actual image as input. This design enables pix2pix to perform well in various image translation tasks, such as converting semantic label maps to real photos, sketches to real images, and image colorization. However, pix2pix's drawback lies in its requirement for a large number of paired training data, which may be difficult to obtain or costly in certain scenarios. Additionally, for large or complex datasets, it may require more computational resources. The supervised nature also implies that the model may not perform optimally on unseen data, limiting its generalization capability. [10-12]

To overcome these limitations, this paper proposes the Keywords-Based Conditional Image Transformation algorithm (KB-CIT). The core of KB-CIT lies in its dynamic acquisition and generation of training data through keywords from input images, thereby circumventing the need for a large amount of paired data and greatly simplifying the data collection process. This method of dynamic data acquisition enables KB-CIT to construct its training set in real-time, significantly improving the efficiency of image transformation. Compared to traditional methods, KB-CIT achieves efficient image transformation without relying on a large amount of data. Most importantly, by continuously obtaining new image data from the internet, the model's training process is strengthened, enabling it to better respond to various image scenarios and styles and enhancing its generalization capability.

DOI: 10.54254/2755-2721/57/20241310

2. Pix2pix

2.1. Pix2pix Network Architecture

The generator in Pix2pix typically adopts the U-Net [13] structure, which is an encoder-decoder architecture with hierarchical feature extraction capability. Its core feature lies in the use of skip connections, enabling the U-Net to preserve structural information at lower levels while retaining high-level semantic information.

The discriminator usually employs the PatchGAN [9] structure, also known as the Markovian discriminator. Unlike traditional GAN discriminators, it does not output a single evaluation value for the entire image but divides the image into "patches" of size NxN and independently evaluates the authenticity of each patch. This design allows the discriminator to effectively capture high-frequency details and textures of the image while having fewer parameters, running faster, and being suitable for images of any size.

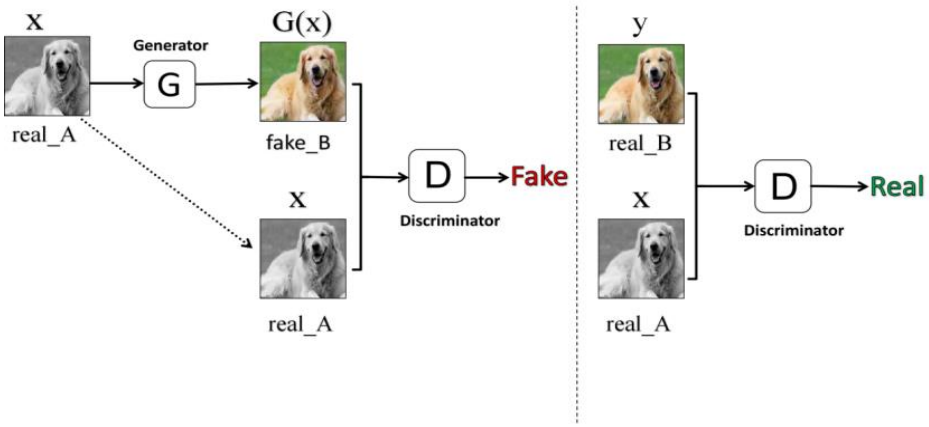

Figure 1. Pix2Pix Architecture

2.2. Pix2pix Training Process

Pix2pix is designed based on Conditional Generative Adversarial Networks (CGAN) specifically tailored for image-to-image translation tasks, which holds particular significance in tasks such as image colorization. As shown in Figure 1., the input is a grayscale image X, with the goal of transforming it into a colored version.

At the beginning of training, this grayscale image X is fed into the generator, which is a U-Net structured network. Its task is to 'infer' color information as much as possible from this grayscale image and output a colored image G(X). The design of the generator allows it to retain many details of the input image, making the generated colored image highly similar in structure to the input grayscale image.

Subsequently, it is necessary for the discriminator to evaluate the quality of the generator's work. The role of the discriminator is to identify whether an image is "real" or "fake" generated by the generator. The generated colored image G(X) and the original grayscale image X are fed to the discriminator together, attempting to identify whether this image pair is "fake." Similarly, to provide a real reference for the discriminator, the real colored image y is also fed to the discriminator along with the grayscale image X, where the discriminator needs to confirm that they are a "real" pair.

Regarding the calculation of the loss, in addition to the CGAN loss, the L1 loss is also introduced. This ensures that the generated colored image G(X) is not only similar to the real colored image y in terms of content but also closely resembles it at the pixel level. Based on these losses, we backpropagate the error and update the network's parameters, continuously improving the generator and discriminator.

DOI: 10.54254/2755-2721/57/20241310

After many iterations, the generator eventually learns how to accurately colorize the grayscale image, while the discriminator becomes more adept at distinguishing between real and generated images.

2.3. Loss Function

In the image-to-image translation task, the pix2pix model proposes a composite loss function that combines the conditional generative adversarial network (cGAN) loss and the L1 loss to drive its training process. The core objective of this composite loss strategy is to ensure that the generated image is not only similar to the target image in overall structure but also highly similar in details to the real image.

$$\mathcal{L}_{\text{cGAN}}(G, D) = E_{x,y}[\log D\,(x, y)] + E_{x,z}\left[\log\left(1 - D\big(x, G(x, z)\big)\right)\right] \tag{1}$$

Ex,y[log D(x, y)] represents the probability evaluated by the discriminator D that the real data pair (x, y) is real. Here, x is the conditional image, and y is the target image.

logD(x,y) is the natural logarithm of the probability output by the discriminator when it sees the real image pair (x, y). The goal of the discriminator D is to maximize this probability, even as its output gets closer to 1. This is because the discriminator aims to confirm that the input image pair is real.

Ex,z[log(1 − D(x, G(x, z))] represents the probability evaluated by the discriminator D that the generated data pair (x, G(x, z)) is fake.

log(1−D(x,G(x,z)) is the natural logarithm of the probability output by the discriminator when it sees the generated image pair (x, G(x, z)) as fake. The generator hopes to minimize this value, while the discriminator aims to maximize it.

In the pix2pix model's conditional CGAN loss function, there exists an adversarial relationship between the discriminator and the generator. The generator's task is to generate as realistic images as possible, while the discriminator's task is to differentiate between real and generated image pairs. This adversarial nature ensures the continuous evolution of the model during the training process, gradually making the generated images closer to the target images.

In the pix2pix model, in addition to using the cGAN loss function, the L1 loss function is also used to measure the pixel-wise difference between the generated image and the real target image. The L1 loss function is represented as:

$$\mathcal{L}_{\mathcal{L}1}(G) = E_{x,y,z}[||y - G(x, z)||_1] \tag{2}$$

where the ||·|| denotes the L1 norm, which is the sum of the absolute values of the elements in the vector. It is used to calculate the total pixel difference between the generated image G(x, z) and the real target image y. The L1 loss function focuses on the difference between each pixel of the generated image and the real target image.

During the optimization process, the goal of the generator G is to minimize this L1 loss, meaning that it aims to generate an image that is as pixel-wise similar to the real target image y as possible. Meanwhile, the cGAN loss function ensures that the generated image is structurally realistic. To ensure that the generated image is both structurally realistic and pixel-wise accurate, pix2pix combines the cGAN loss and the L1 loss. The final loss function of the model is:

$$G^* = \arg\min_G \max_D \mathcal{L}cGAN\,(G, D) + \lambda\mathcal{L}_{\mathcal{L}1}(G) \tag{3}$$

The combination of these two losses ensures that the generated image is not only similar to the real image in macro structure but also highly faithful in micro details.

3. Keywords-Based Conditional Image Transformation

This paper proposes a Keywords-Based Conditional Image Transformation model based on Pix2pix image translation. This model aims to address the limitations of traditional pix2pix in data collection and computational resources and provide users with a more efficient and personalized image translation solution.

DOI: 10.54254/2755-2721/57/20241310

The core of this algorithm lies in the use of cutting-edge image understanding techniques to automatically extract keywords from the input uncolored image. These keywords are subsequently used to automatically retrieve related colored images from the internet. Upon obtaining these colored images, the algorithm converts them into grayscale versions and pairs them with their original colored versions, thus constructing a new training dataset. Subsequently, the pix2pix model is trained using this dataset. Upon completion of the training, the model can automatically colorize the input uncolored image and output its colored version.

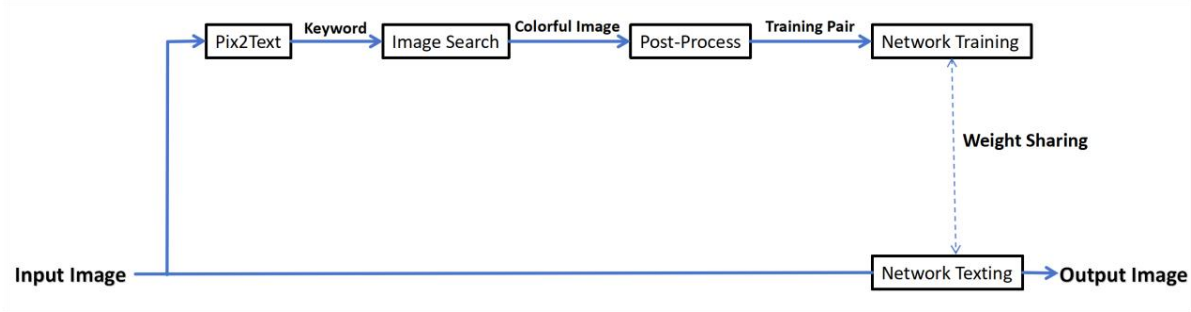

Figure 2. KB-CIT Algorithm Architecture

3.1. Pix2Text

Upon successfully acquiring the input image, the subsequent task is to comprehend its content deeply and transform this information into descriptive text, namely keywords. For this purpose, a series of advanced image understanding techniques are employed in this study. This step is crucial as the accurate and appropriate extraction of keywords from the image directly determines the success rate of subsequent searches for colored images.

During this process, in addition to identifying the core objects in the image, it is also necessary to delve into the details of the entire scene. Taking an example of an image of a golden retriever on a lawn, it is essential not only to accurately capture the features of the "golden retriever" but also to pay attention to the "lawn" background environment in which it is situated. This ensures that the keywords not only describe the features of the object but also reveal the background characteristics of the overall scene.

To achieve this step, a variety of cutting-edge image-to-text conversion technologies have been adopted in this study. Among them, Image Captioning [14-16] is a technique that can generate descriptive sentences for images, usually based on deep learning, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [17-18]. Additionally, Visual Transformers [19] and Attention-based Models have also demonstrated outstanding performance in the field of image understanding as they can perform a detailed analysis of specific regions in the image. The combination of these technologies enables us to understand the content of the image from multiple perspectives and generate high-quality keywords matching it. Currently, there are various readily available tools and platforms in the market that support these functions, such as Microsoft's Azure AI Vision, IBM's Watson Visual Recognition, and some prompt generation tools such as Google's Image to Prompt Generator plugin and Midjourney's Image-to-Text function, which directly generates multiple different prompts based on the content of the image. They provide developers with powerful image understanding and conversion tools, enabling the rapid fulfillment of the aforementioned requirements.

3.2. Image Search Using Keywords

After obtaining the keywords, the immediate task is to search for corresponding colored images on the internet using these keywords. For this purpose, automated web crawling technology has been employed in this research.

DOI: 10.54254/2755-2721/57/20241310

Google Image Search has been selected as the primary source of image data due to its rich image resources and excellent search functionality. The crawler program used in the study simulates the actual behavior of users in Google Image Search, such as entering queries, scrolling to load more images, and so on.

To ensure the accuracy of the search results, it is necessary not only to use the extracted keywords from the image for the queries but also to combine them with some search optimization strategies. For instance, specific filters are used to restrict the time range or source of search results. The content of the search result pages is subsequently parsed to extract the image URLs. Using these URLs, the program automatically downloads the relevant colored images and stores them in the designated directory, preparing them for the next processing step.

150 images were obtained in the experiment as the dataset, with 120 images used as the training set and the remaining 30 images used as the test set.

Throughout the entire search and download process, respect for data privacy and copyright has been maintained, ensuring that all data used are from publicly accessible web pages and complying with all relevant terms of use and agreements. Additionally, necessary precautions have been taken for images that may contain personal information or are protected by copyright, ensuring that they are not illegally downloaded or used.

3.3. Post-Process

After successfully downloading the colored images, grayscale conversion and pairing operations have been performed for the needs of model training, both of which have been automated through Python scripts. Firstly, the colored images are batch-converted to their corresponding grayscale versions, retaining only brightness information while simplifying color information. Subsequently, in the pairing operation, each original colored image is combined with its corresponding grayscale image to form a training pair. Here, the grayscale image serves as the input during model training, while the colored image serves as the expected output. This automated batch processing method enhances work efficiency and ensures the consistency of the entire dataset in processing, laying a solid foundation for the subsequent model training.

3.4. Training and Testing the Network Using Pix2pix

After processing the dataset, the model training phase begins. Considering the outstanding performance of pix2pix and its expertise in image-to-image transformation tasks, it was decided to use this model for subsequent training and testing.

The core of the training is an adversarial process that aims to make the colored images generated by the generator as close as possible to the real colored images while making it difficult for the discriminator to distinguish between the generated images and the real images. This process aims to ensure that the generated colored images have high quality and realism.

4. Experimental Analysis

4.1. Keywords-Based Search vs. Image-Based Search: Reasons and Advantages

In the design of the algorithm flow, the strategy of keyword extraction followed by image searching using these keywords has been adopted. This method has significant advantages compared to the direct image-to-image search:

(1) Providing richer semantic information: The pix2text step can extract more detailed semantic descriptions from the image. Compared to using pixel information or features directly from the image, the textual description can provide the search algorithm with more in-depth and human-contextual context information, thus more accurately pinpointing the desired image content.

(2) Accuracy and relevance of the search: Through keyword-based searching, ensuring that the retrieved images highly match the original description. Compared to image-to-image search, this method

DOI: 10.54254/2755-2721/57/20241310

can better control the relevance of search results, ensuring that the obtained images accurately reflect the context and content of the original image.

(3) Improved search efficiency: Text-based search technology is relatively mature, and compared to direct image-to-image search, keyword-based searching on existing search engines is often more efficient, quickly returning high-quality results.

(4) Avoiding redundancy and repetition: Image-to-image searching may return images that are very similar to the original image, leading to redundancy in the dataset. Through text descriptions and keyword searches, it is easier to obtain images that are similar in content but different in form, thereby increasing the diversity of the data.

(5) Better avoidance of copyright risks: Direct image-to-image search may involve the use and dissemination of image content, whereas the search method based on textual descriptions and keywords is better able to avoid direct copying and dissemination of the original image, reducing copyright-related risks.

(6) Providing greater flexibility: Text-based description and search methods provide researchers with more operational space. The extraction strategy of keywords can be adjusted according to the requirements, text descriptions can be optimized, or switching between different search engines can be done flexibly.

Overall, through the approach of pix2text and keyword extraction, not only the accuracy and efficiency of image search have been improved, but also the quality and diversity of the data have been ensured, creating favorable conditions for subsequent image processing tasks.

4.2. Experimental Environment and Results

This study conducted experiments in a computational environment equipped with an Intel Core i9-12900HX CPU, NVIDIA GeForce RTX 3080 Ti GPU, and 32.0 GB RAM. The operating system was 64-bit Windows 11, and the programming language chosen was Python 3.9.18. The deep learning framework used was PyTorch, combined with CUDA version 11.3 for GPU accelerated computing.

Regarding the training parameters, the Adam optimizer [20] was selected, with a learning rate set to 0.0002, and the batch size set to 1. The generator adopted a U-Net structure, while the discriminator used a PatchGAN. To monitor the training progress in real-time, visdom was used as the real-time visualization tool.

The experimental results show that after 200 epochs of training, the KB-CIT algorithm performed exceptionally well in the image coloring task, despite using a limited amount of training data. As shown in Figure 3, the model not only successfully colored the core object "dog" in the image but also demonstrated a deep understanding of the details of the entire scene, such as the "lawn" background, and performed well in the coloring effect. However, despite the overall impressive performance, some minor shortcomings were noticed in some images. For example, in the first set of images, there was a slight disparity between the color of the lawn and the real image; in the second set of images, the dog's tongue was colored green, which is an area for further improvement.

Overall, the KB-CIT algorithm still performed remarkably well with a small amount of training data, which holds significant application potential for the image coloring task.

DOI: 10.54254/2755-2721/57/20241310

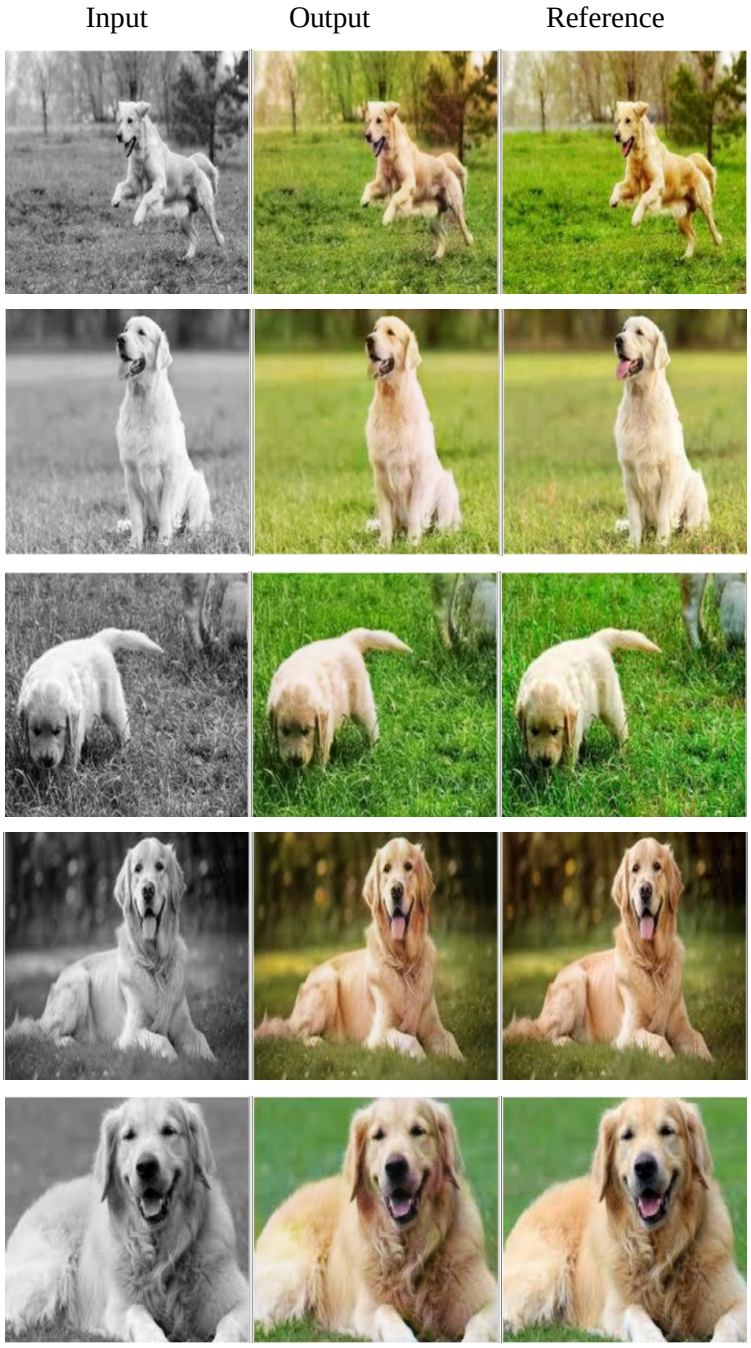

Figure 3. Image Colorization Results

4.3. Algorithm Advantages

(1) Real-time training of personalized data: The model can be customized for training based on specific requirements and data characteristics, significantly increasing the flexibility of the model.

(2) Reduced computational resource requirements: Compared to the traditional pix2pix model, the new algorithm appears to be more lightweight in terms of computational resources.

(3) Efficient data utilization: Despite using only a small amount of training data, the model still achieves satisfactory generation results.

DOI: 10.54254/2755-2721/57/20241310

(4) Simplified data collection process: Traditional data collection methods, such as manual annotation and filtering, are both expensive and time-consuming. The new algorithm greatly simplifies this process through automation, saving time and costs.

The proposed improvement method not only simplifies the use and training process of pix2pix but also enhances its performance in multiple aspects. The results of this study pave the way for the widespread practical application of deep learning, demonstrating its broad practical value.

5. Conclusion

This paper proposed the Keywords-Based Conditional Image Transformation (KB-CIT) algorithm, aiming to address the issues of the pix2pix model in terms of large-scale paired training data and computational resource requirements. The KB-CIT algorithm dynamically acquires and generates training data through keywords, greatly simplifying the data collection process and achieving the ability to generate high-quality color images under conditions of limited training data. Furthermore, KB-CIT demonstrates its unique superiority in data utilization efficiency and real-time training of personalized models.

However, the algorithm also has certain limitations. Inaccurate image descriptions may lead to the generation of suboptimal training data. Moreover, the availability of online images that match the input keywords could restrict its application in less-represented or highly specialized domains.

The development prospects for the KB-CIT algorithm are promising, with significant opportunities for both enhancements and expanded applications. Employing more advanced natural language processing techniques could refine keyword extraction and interpretation, thus improving the accuracy and relevance of the training data. Furthermore, the algorithm's effectiveness in various types of images and genres warrants exploration.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems, 27, 2672–2680

[2] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative Adversarial Networks: An Overview. IEEE Signal Processing Magazine, 35(1), 53–65

[3] Wang, Z., She, Q., & Ward, T. E. (2020). Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy (arXiv:1906.01529)

[4] Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., & Wang, F.-Y. (2017). Generative adversarial networks: Introduction and outlook. IEEE/CAA Journal of Automatica Sinica, 4(4), 588–598

[5] Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks (arXiv:1511.06434)

[6] Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein Generative Adversarial Networks. Proceedings of the 34th International Conference on Machine Learning, 214–223

[7] Denton, E. L., Chintala, S., szlam, arthur, & Fergus, R. (2015). Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. Advances in Neural Information Processing Systems, 28, 1486–1494

[8] Mirza, M., & Osindero, S. (2014). Conditional Generative Adversarial Nets (arXiv:1411.1784)

[9] Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5967–5976

[10] Bousmalis, K., Irpan, A., Wohlhart, P., Bai, Y., Kelcey, M., Kalakrishnan, M., Downs, L., Ibarz, J., Pastor, P., Konolige, K., Levine, S., & Vanhoucke, V. (2018). Using Simulation and Domain Adaptation to Improve Efficiency of Deep Robotic Grasping. 2018 IEEE International Conference on Robotics and Automation (ICRA), 4243–4250

DOI: 10.54254/2755-2721/57/20241310

[11] Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. 2017 IEEE International Conference on Computer Vision (ICCV), 2242–2251

[12] Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q., & Shen, D. (2017). Medical Image Synthesis with Context-Aware Generative Adversarial Networks. Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention, 10435, 417–425

[13] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (pp. 234–241). Springer International Publishing

[14] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and Tell: A Neural Image Caption Generator (arXiv:1411.4555)

[15] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6077–6086

[16] Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-Memory Transformer for Image Captioning. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10575–10584

[17] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Proceedings of the 32nd International Conference on Machine Learning, 2048–2057

[18] You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image Captioning with Semantic Attention (arXiv:1603.03925)

[19] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020, October 2). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations

[20] Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization (arXiv:1412.6980)