AI6103 Team Project

Kozhimadam Shahin Shah (G2303660L)[1], Gong Lehan (G2205030D)[2], Uwineza joseph (G2303477F)[3], Ntambara Etienne (G2304253K)[4], Mukanyandwi joselyne (G2304244J)[5]

[1] SHAHINSH001@e.ntu.edu.sg , [2] GONG0072@e.ntu.edu.sg , [3] JOSEPH005@e.ntu.edu.sg , [4] NTAM0001@e.ntu.edu.sg , [5] JOSELYNE001@e.ntu.edu.sg

Abstract

This project presents an enhanced approach to paired image translation using Generative Adversarial Networks, with a focus on reimagining the Pix2Pix framework. We introduce a novel architecture, integrating a uniquely modified U-Net generator and a dual-critic system, specifically designed for translating edge maps to photographic images. Diverging from standard implementations, our model is built from the ground up without leveraging existing source codes, and features advanced loss functions like the Wasserstein loss for stability and the L1 loss for fidelity. The model's efficacy is quantitatively assessed using the Frechet Inception Distance, highlighting its capability to generate realistic and accurate translations.

Introduction

The field of image processing has seen remarkable advances in recent years, especially in the area of image translation - the task of transforming an image from one domain into another while retaining its inherent characteristics. This field is crucial in various applications, ranging from artistic image generation to practical uses in medical imaging and computer vision. Among the numerous approaches to image translation, Generative Adversarial Networks (GANs) have emerged as a groundbreaking method, offering unprecedented capabilities in generating realistic images. This study focuses on a specific application of GANs: paired image translation, where the objective is to translate images between two closely related domains, such as sketches to photographs. We present a novel approach to this problem by reimagining the Pix2Pix GAN framework, incorporating significant enhancements in the network architecture and training process.

Background

In the realm of image processing, paired image translation stands as a vital field, wherein the goal is to translate an image from one domain to another while retaining core aspects of the original image. This process is deeply rooted in understanding and manipulating complex image features, requiring sophisticated models capable of intricate feature extraction and transformation. Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) have revolutionized this domain by their ability to generate highly realistic images. The core principle of GANs involves two networks, a generator and a discriminator (or critic), working in opposition to each other. The generator aims to create images indistinguishable from real ones, while the discriminator evaluates their authenticity. This adversarial dynamic is pivotal in driving the model towards producing more refined and realistic outputs.

The Pix2Pix GAN (Isola et al. 2017), an extension of the basic GAN framework, is specifically tailored for paired image translation tasks. It employs a modified U-Net architecture (Ronneberger, Fischer, and Brox 2015) for the generator, known for its effectiveness in detailed feature extraction and image reconstruction. The U-Net's design, characterized by its symmetric down-sampling and up-sampling paths connected by skip connections, is adept at preserving both high-level and low-level image details. However, training GANs, including Pix2Pix, often encounters challenges such as vanishing gradients and mode collapse, where the model fails to capture the full diversity of the target domain. To address these issues, the Wasserstein loss (W loss) (Arjovsky, Chintala, and Bottou 2017) is introduced. This loss function, unlike traditional ones, provides a more stable and meaningful gradient for the generator, effectively preventing mode collapse and encouraging the generation of diverse and realistic images.

Evaluating GAN performance in image translation is challenging due to the subjective nature of image quality and the varying criteria for successful translation based on task and context. The Frechet Inception Distance (FID) (Heusel et al. 2017) emerges as a critical metric in this scenario. FID measures the distance between feature vectors of real and generated images, extracted using a pre-trained Inception network. By comparing the distributions of these high-level features, FID provides a quantitative assessment of the quality and diversity of the generated images. A lower FID score indicates a closer resemblance to real images, signifying a more successful image translation. This metric is particularly valuable in paired image translation tasks, where the objective is not just to create visually appealing images but to accurately replicate specific attributes of the source domain, ensuring fidelity and realism in the translated images.

Figure 1: Sample pairs from the Edges2Shoes dataset showcasing the edge map of a shoe (top) and its corresponding real photograph (bottom).

Dataset

Our project utilizes the Edges2Shoes dataset, a specialized collection of image pairs designed for image-to-image translation tasks, particularly in the domain of translating edge maps to colored images. Each pair in the Edges2Shoes dataset consists of two images: one representing a sketch or edge map of a shoe and the other a corresponding real photograph of the shoe. This pairing is essential for our work, as it provides a direct correlation between the input (edge map) and the desired output (photographic image), facilitating an accurate and focused translation task. The Edges2Shoes dataset is well-suited for our study due to its clear delineation between the source and target domains and its consistent pairing of related images. See Figure 1 for sample pairs from the dataset.

The dataset offers several advantages for our study. Firstly, it provides a focused context (shoes) for the translation task, allowing for a detailed examination of the model's effectiveness in handling various aspects such as color, texture, and shape. The variety of styles, colors, and shapes within the shoe images in the dataset ensures that our model is exposed to a wide range of image characteristics, which is critical for robust learning. This diversity is instrumental in training the model to adapt to different styles and details in the images. Secondly, the real-world nature of the photographs in the dataset means that our model's outputs can be evaluated for their practical applicability and realism. Training with real-world images ensures that the model learns to produce outputs that are not only visually appealing but also realistically aligned with what would be expected in actual photographs. The Edges2Shoes dataset, therefore, serves as both a training platform and a benchmark for assessing the performance of our model, providing a comprehensive foundation for evaluating the success of our image translation approach.

Methodology

Our methodology section details the comprehensive framework and strategies employed in our image-to-image translation project. This includes an in-depth discussion of the network architecture, which comprises a modified generator and a dual-critic system designed for precise image evaluation. We elaborate on the specific modifications made to the standard U-Net architecture in our generator and the distinctive roles of the global and local critics. The section also covers the selection and application of key loss functions, including the rationale behind their use, and outlines our approach to the optimization process. This includes strategies for training stability and achieving a balance between realism and fidelity in the generated images. Additionally, the methodology encompasses our approach to evaluating the performance of the model, highlighting the metrics used for assessing the quality and accuracy of the image translation.

Network Architecture

The architecture of our model consists of a modified generator and two critics, each integral to the image translation process.

Generator (Modified U-Net): The generator is a reinterpretation of the standard U-Net architecture (Ronneberger, Fischer, and Brox 2015) shown in Figure 2, customized for the demands of image-to-image translation. At the heart of this architecture are three types of blocks: DownSampling, UpSampling, and FeatureMap blocks. Each type plays a crucial role in the image processing pipeline. The DownSampling blocks, equipped with a double convolution mechanism followed by max pooling, are designed to progressively reduce the spatial dimensions of the input while simultaneously increasing the depth of feature maps. This process is critical for distilling important features from the input images. Conversely, the UpSampling blocks utilize transposed convolutions to gradually reconstruct the image in the target domain. These blocks work to reverse the dimensional reduction carried out by the DownSampling blocks, restoring the image to its original size with new features representative of the target domain. Complementing these two is the FeatureMap block, which serves to adjust the channel dimensions, ensuring that the network can adaptively manage the depth of feature maps at different stages of the architecture.

In our model, these blocks are arranged in a carefully designed sequence to optimize feature extraction and image reconstruction. The generator begins with a sequence of DownSampling blocks, each diving deeper into the feature space of the input image. This is followed by a corresponding series of UpSampling blocks, which work to gradually rebuild the image, integrating features learned during the downsampling phase. Skip connections are employed between each DownSampling and corresponding UpSampling block, allowing the network to preserve and utilize fine-grained details from earlier layers. This design ensures a harmonious blend of low-level details and high-level semantic information in the final output.

However, our implementation diverges significantly from the typical U-Net architecture. While standard U-Nets slightly reduce spatial dimensions in each DownSampling and UpSampling block due to convolution operations, our

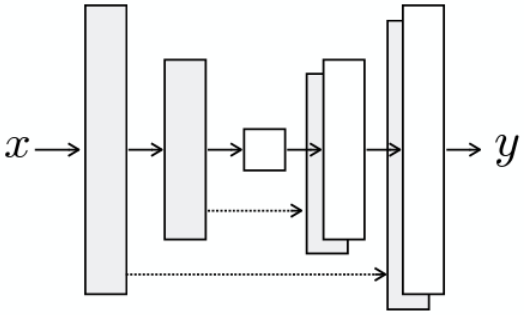

Figure 2: Standard U-Net Architecture. Adapted from (Isola et al. 2017).

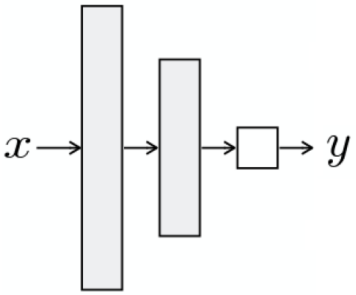

Figure 3: Critic Architecture. Adapted and Modified from (Isola et al. 2017).

model incorporates padding in these blocks to maintain the spatial dimensions throughout the network. This approach ensures a more detailed feature representation by avoiding the loss of spatial information. Additionally, in a few of the DownSampling blocks, dropout is applied to introduce stochasticity, enhancing the robustness and variability of the output images. This feature is particularly valuable in preventing overfitting and ensuring that the generated images are not just high in fidelity but also diverse in representation.Furthermore, our generator features an increased depth with 7 DownSampling and 7 UpSampling blocks, as opposed to the usual 4, enabling the extraction and processing of more complex features. This depth enhancement is crucial for capturing and translating intricate details necessary in high-fidelity image translation tasks. Additionally, unlike the traditional U-Net, our model does not require cropping the outputs of DownSampling blocks before passing them through skip connections to the UpSampling path. The spatial dimensions of the outputs are preserved, facilitating seamless integration and propagation of features across the network. Finally, the generator's output undergoes sigmoid activation to normalize pixel values, a vital step in ensuring the visual quality of the translated images.

Critics: The critics in our model are innovatively structured to evaluate the authenticity of generated images with a high degree of precision. A general architecture of the critic is shown in Figure 3. Each critic begins with a Feature Map block, identical to that found in the generator, which functions to initially process the input image. This is followed by a number of downsampling blocks, mirroring those in the generator, which progressively reduce the image resolution while deepening the feature analysis. The culmination of this process is a final convolution layer that outputs a grid of numbers. This grid represents a patch-level assessment of the image, with each number indicating whether the corresponding image patch is real or fake. Unlike traditional approaches that yield a single scalar output for the entire image, this grid output allows for a more granular evaluation, enabling the critic to scrutinize specific regions of the image for authenticity.

Incorporating two distinct critics, global critic and local critic, each targeting different resolution levels, significantly bolsters the model's evaluative capacity. The global critic, equipped with 4 downsampling blocks, is adept at analyzing the image on a macro scale. It evaluates the overall structure and composition, ensuring that the generated images adhere to the broader aesthetic and structural norms of the target domain. In contrast, the local critic, with just 2 downsampling blocks, hones in on the finer, high-resolution details of the image. It scrutinizes textures, edges, and other minute details, playing a crucial role in enhancing the realism of the image. This dual-critic approach not only enriches the feedback provided to the generator but also fortifies the training process against common pitfalls such as mode collapse. The patch-level feedback from the critics introduces a nuanced layer of scrutiny, enabling targeted improvements and fostering a more robust learning environment for the generator. By balancing the macro and micro-level assessments, the model ensures that the generated images are not only visually coherent in their overall layout but also exhibit a high degree of fidelity in their finer details, making them indistinguishable from real images in the target domain.

Loss Functions

We utilize two critical loss functions for effective training: adversarial loss and reconstruction loss. Adversarial loss is designed to create a dynamic where the generator and the discriminator are in a constant state of competition. The generator aims to produce images that are indistinguishable from real images, while the discriminator endeavors to differentiate between real and generated images. This competitive framework encourages the generator to improve its output based on the discriminator's feedback. The adversarial loss quantifies how well the generator is able to fool the discriminator. It is traditionally computed using binary cross entropy (Isola et al. 2017), a method that evaluates the discriminator's ability to classify images correctly as real or generated. However, this standard method can sometimes lead to training instabilities, particularly due to vanishing gradients and mode collapse issues. To overcome these challenges, we have adopted the Wasserstein loss (W loss) (Arjovsky, Chintala, and Bottou 2017), known for its capability to stabilize the training process. The W loss quantifies the Earth Mover's Distance (EMD) between the distribution of real and generated images, providing a more meaning-

ful and smooth gradient for the generator. This loss function is particularly effective in mitigating mode collapse, where the generator tends to produce limited output varieties. It encourages the generator to produce a diverse range of realistic images. To further enhance the stability brought by W loss, we implement a gradient penalty as a soft constraint, ensuring Lipschitz continuity. This regularization technique prevents the gradients of the critic from becoming excessively steep, which is crucial for maintaining stable and reliable training dynamics.

Alongside the adversarial loss, the reconstruction loss plays an equally important role in preserving the fidelity of the generated images to the input. The reconstruction loss ensures that key features of the input image are retained in the generated output, thus maintaining content and structural integrity. We have chosen the L1 loss, or mean absolute error, as our reconstruction loss function. L1 loss is preferred in image-to-image translation tasks because it tends to produce less blurring, resulting in sharper and more detailed images. By penalizing the absolute differences in pixel values between the generated and real images, the L1 loss effectively guides the generator to closely mimic the target images. This balance between the Wasserstein adversarial loss for ensuring diversity and stability, and the L1 reconstruction loss for maintaining image fidelity, creates a robust training framework. This combination is instrumental in our model's ability to generate high-quality, realistic images that are both varied and closely aligned with the input image characteristics.

Evaluation Metrics

For evaluating the quality of images generated, we employ the Frechet Inception Distance (FID) as the primary metric (Heusel et al. 2017). FID is particularly suited to our needs as it adeptly measures the quality and diversity of generated images in comparison to the corresponding real images. Utilizing the Inception-v3 network, pre-trained on ImageNet, FID extracts feature vectors from both sets of images. These feature vectors, representing high-level aspects of the images, are treated as samples from multivariate Gaussian distributions. The FID score is then calculated by computing the Frechet distance between these two distributions, derived from their respective means and covariances. In the context of paired image-to-image translation, a lower FID score indicates a closer resemblance between the generated images and their real counterparts, suggesting that the translation process is effectively capturing the essential qualities of the original images.

Optimization

Our model's optimization approach is tailored distinctly for the generator, global critic, and local critic. The generator is fine-tuned to generate high-quality images, aiming to minimize a combined metric of adversarial and reconstruction losses. In contrast, both the global and local critics are optimized with the primary objective of enhancing their ability to differentiate between real and generated images. This is achieved by minimizing the adversarial loss, ensuring that each critic becomes more proficient in accurately identifying and classifying images.

The optimization process is iterative, with alternating updates to the generator and critics, ensuring a balanced adversarial training environment where neither component overpowers the other. For this optimization, we employed the Adam optimizer, favored in deep learning for its efficiency in handling sparse gradients. We set the learning rate to 0.0002 and the momentum parameters to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Additionally, the optimization was conducted with a batch size of 8 and spanned over 20 epochs. Notably, the generator was initialized with weights obtained from preliminary trial experiments. This careful and iterative optimization process is crucial in guiding the model to produce high-quality, diverse, and realistic images that are closely aligned with the content and style of the paired source images. This approach ensures that the model not only learns to generate visually appealing images but also adheres to the specific requirements of our image translation task.

Inference

The inference process in our image-to-image translation model aligns closely with the methodology described in (Isola et al. 2017). In this phase, the generator network is operated in the same manner as during the training phase, which is a notable deviation from typical protocols used in generative models. One key aspect of this approach is the continued application of dropout during the inference stage. Conventionally used during training to prevent overfitting, the use of dropout at inference introduces a level of randomness or stochasticity in the output, potentially leading to a greater diversity in the generated images. This can be particularly beneficial in scenarios where variability in outputs is desired.

Furthermore, we employ batch normalization during inference, but with a crucial modification: instead of using aggregated statistics from the training batches, we utilize the statistics of the test batch itself. This approach ensures that the normalization process is more accurately aligned with the specific data characteristics present during inference. The rationale behind this is to adapt the model's behavior more precisely to the current data, potentially enhancing the quality and consistency of the generated images. Such a strategy, while diverging from standard practices, can lead to improved performance and stability in the model's outputs, ensuring that the images generated during inference closely resemble those produced during training both in terms of quality and style.

Results

In analyzing the generator loss curves from our Pix2Pix GAN experiment, which used both Binary Cross Entropy (BCE) and Wasserstein (W) loss, we found insightful differences in the training stability of the two methods. Figure 4 illustrates these differences. While direct comparisons of the absolute loss values between BCE and W loss are not meaningful due to their distinct mathematical formulations and scales, observing the pattern of variations over time reveals

crucial insights. The W loss exhibits a stable, consistent decline in generator loss, indicative of a more controlled and predictable training process. Such stability is particularly advantageous in complex tasks like image translation. In contrast, the BCE loss curve displays an initial increase and then a decrease in loss, accompanied by slight fluctuations. This pattern suggests a less predictable training journey, common in GAN frameworks using BCE loss.

Shifting focus to the quality of the output, as displayed in Figure 5, our model's performance is noteworthy. The enhanced Pix2Pix GAN achieves a Frechet Inception Distance (FID) of 80.71 on the validation set, a clear indication of its effectiveness in generating images that closely resemble their real counterparts. The model successfully translates the color and texture diversity from the original photographs to the generated images, even in instances where the input edge maps are incomplete. This ability to accurately replicate complex features from the input demonstrates the model's adeptness at high-fidelity image translation, making it a valuable tool for practical applications where capturing detailed nuances of the source data is essential.

In addition to the previous analyses, we incorporated Figure 6 to evaluate our enhanced Pix2Pix GAN model's performance in scenarios where the original Pix2Pix GAN (Isola et al. 2017) tends to fail. Specifically, we focused on cases with densely populated edge maps, which traditionally challenge the original Pix2Pix GAN, often resulting in generated images with overly complex textures and a chaotic mix of colors. Our model exhibits a marked improvement in handling such intricate edge maps. The images translated by our model, although deviating in color from the original photographs, maintain a realistic appearance with well-balanced textures and colors. This enhanced capability can be attributed to our model's refined feature extraction process, enabled by the modified U-Net architecture and the stability offered by the Wasserstein loss. This loss function, in contrast to BCE, promotes a more nuanced gradient flow, allowing the generator to better navigate the complex feature landscape presented by densely edged inputs.

However, our model is not without its limitations. In cases with sparser edge maps, it occasionally struggles to convey the three-dimensional nature of the shoes as depicted in Figure 7. This shortcoming could be due to the model's tendency to focus more on texture and color replication, potentially at the expense of accurately capturing depth and volume. This issue might stem from the model's architecture, which may require further tuning to balance the rendering of textures and the portrayal of depth. Future iterations of the model could benefit from incorporating additional depth cues or exploring different architectural modifications to address this specific challenge.

Overall, while our enhanced Pix2Pix GAN model demonstrates significant improvements in handling complex edge maps, the limitations underscore the importance of continued refinement, particularly in balancing the model's proficiency in texture and color replication with its ability to convey depth and volume.

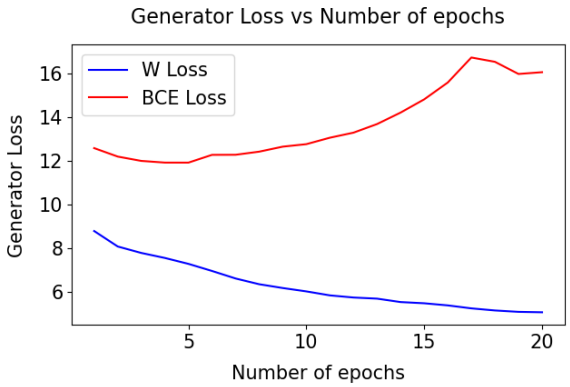

Figure 4: Comparison of Generator Loss Curves. The plot shows the progression of generator loss across training epochs for Binary Cross Entropy (BCE) loss and Wasserstein (W) loss.

Conclusion

Our project enhances the Pix2Pix GAN framework for image-to-image translation, focusing on translating edge maps to photographic images. The integration of a modified U-Net generator and a dual-critic system, alongside the use of Wasserstein and L1 losses, has been pivotal in improving the model's training stability and output fidelity. This approach, developed without leveraging existing source codes, demonstrates the model's capability in generating realistic translations, especially in processing complex edge maps. While the model shows a marked improvement in certain aspects over the original Pix2Pix GAN, it also brings to light areas for further refinement, such as enhancing depth and volume representation in images with sparser edge maps.

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein GAN: Machine Learning. *stat. ML*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and 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.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Figure 5: Sample results from the enhanced Pix2Pix GAN. Each column displays the edge map of a shoe (top) and its corresponding translated image (middle), alongside the real photograph (bottom).

Figure 6: Comparison of Translated Images for Dense Edge Maps. Showcases the improved handling of complex edge maps by our enhanced Pix2Pix GAN model versus the original Pix2Pix GAN.

Figure 7: Model Performance on Sparse Edge Maps. Demonstrates challenges in rendering 3D aspects of shoes in scenarios with less detailed edge maps.