DATS369 Machine Learning with Graphs
Final Project Proposal

Yuanhe Guo (yg2709)

1 Introduction

Recent advancements in generative models such as Stable Diffusion have showcased the remarkable ability of machine learning to produce high-resolution images. Beyond text-guided generation, these models also support high-quality image-to-image translation. However, this functionality is typically constrained to a single input image, significantly limiting the potential for personalization.

This raises an important challenge: given a batch of user-preferred images, how can we generate a novel image that meaningfully reflects all their preferences?

Approaches like ControlNet [1] introduce finer control by incorporating multiple input modalities, such as depth maps. Nonetheless, these methods often rely on expert human input, which hinders their scalability and accessibility.

Alternatively, Graph2Pix [2] presents a promising solution by tackling the multi-instance image-to-image translation task through a graph-based framework. By representing a sequence of input images as a graph and applying a Graph Convolutional Network (GCN) [3] to the adjacency matrix of image features, the model effectively aggregates information from multiple inputs to guide the generation process.

2 Baseline Explanation

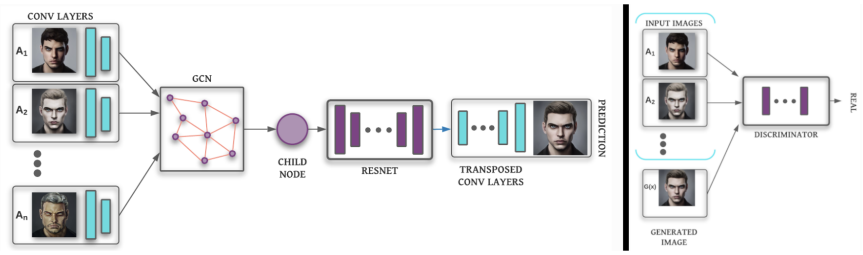

Figure 1: Overview of the Graph2Pix [2] architecture.

Graph2Pix [2] builds upon Pix2PixHD [4], a GAN-based model for high-resolution image-to-image translation. Unlike Pix2PixHD [4], which translates

from a single image, Graph2Pix replaces the source image with an embedding generated by a Graph Convolutional Network (GCN). This design allows the model to aggregate information from an entire batch of input images.

2.1 Pix2PixHD

Pix2PixHD [4] generates novel images from input data, enabling tasks such as generating photo-realistic images from segmentation masks. To handle high-resolution outputs, it uses a two-stage generator $G = \{G_1, G_2\}$. An UNet G_1 with residual blocks is trained on low-resolution images. While a residual encoder network G_2 is trained to downscale and upscale images.

For high-resolution GAN training, the real and generated images are downsampled by factors of 2 and 4 to form a three-scale pyramid. Three discriminators, D_1, D_2, and D_3, are assigned to each scale, resulting in a multi-task learning objective:

$$\min_{G} \max_{D_1, D_2, D_3} \sum_{k=1}^{3} \mathcal{L}_{GAN}(G, D_k). \tag{1}$$

To stabilize training, Pix2PixHD incorporates a feature matching loss:

$$\mathcal{L}_{FM}(G, D_k) = \mathbb{E}_{(\mathbf{s},\mathbf{x})} \sum_{i=1}^{T} \frac{1}{N_i} \left\| D_k^{(i)}(\mathbf{s}, \mathbf{x}) - D_k^{(i)}(\mathbf{s}, G(\mathbf{s})) \right\|_1, \tag{2}$$

where $D_k^{(i)}$ denotes the i-th discriminator layer, and N_i is the number of elements in that layer. Here, $\mathbf{s}$ and $\mathbf{x}$ represent the input and ground-truth images.

2.2 Graph2Pix

As shown in Fig. 1, Graph2Pix [2] modifies the G_2 network in Pix2PixHD by replacing the downscale module with a GCN [3]. To adapt GCNs for image data, the affine transformation is substituted with 3×3 convolutions. The update rule for the l-th GCN layer is:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}\mathrm{Conv}(H^{(l)})\right), \tag{3}$$

where $\tilde{D}$ is the degree matrix and $\tilde{A}$ is the adjacency matrix with self-loops, σ is the activation function.

To generate a child image conditioned on its ancestors while preserving relational structure, the child node is initialized with noise, and the model is trained to reconstruct the original image. Meanwhile, to further enhance the influence of input images on the generated image, all ancestor images are fed into the discriminator to compute the GAN loss during training.

Method	FID ↓	LPIPS ↓
Pix2PixHD [4]	23.89	0.32
Graph2Pix [2] Paper	**19.29**	**0.25**
Graph2Pix [2] Ours	62.88	0.32

Table 1: Comparison between our reproduced model and the results reported in the Graph2Pix [2] paper. Due to the significantly smaller scale of data used in our reproduction, this comparison serves as a qualitative reference rather than a direct benchmark.

3 Baseline Reproduction

3.1 Dataset

Graph2Pix [2] curated a dataset from Artbreeder[1], a platform that allows users to mix multiple images. As a result, the dataset contains a large number of image graphs, represented as binary trees. The full dataset consists of $86,373$ training images and $15,239$ testing images.

In our experiment, we use a released subset of 7,000 images. Each graph is a binary tree of depth 2, resulting in 7 images per graph. This yields a total of 1,000 graphs, with 850 used for training and 150 for testing.

3.2 Experiments and Results

We trained the Graph2Pix model for 200 epochs with a batch size of 8. All images were resized to 256×256, following the default configuration.

For evaluation, we used the trained checkpoint to generate images on the test set and computed FID and LPIPS scores over 150 test images. Tab. 1 compares the performance of our reproduced model with the results reported in the original paper. We also include the Pix2PixHD [4] results for reference. Due to differences in the scale of the training and testing datasets, these comparisons are not directly equivalent. Nevertheless, the relatively low LPIPS score demonstrates the effectiveness of the method.

Fig. 2 presents qualitative comparisons between the generated images and the ground truth, illustrating the model's ability to preserve structure and style across different instances.

4 Future Plan

Dataset. The currently available subset of the dataset is relatively small and may not produce compelling results. To address this limitation, we plan to contact the original authors to request access to the full dataset.

[1] `https://www.artbreeder.com/`

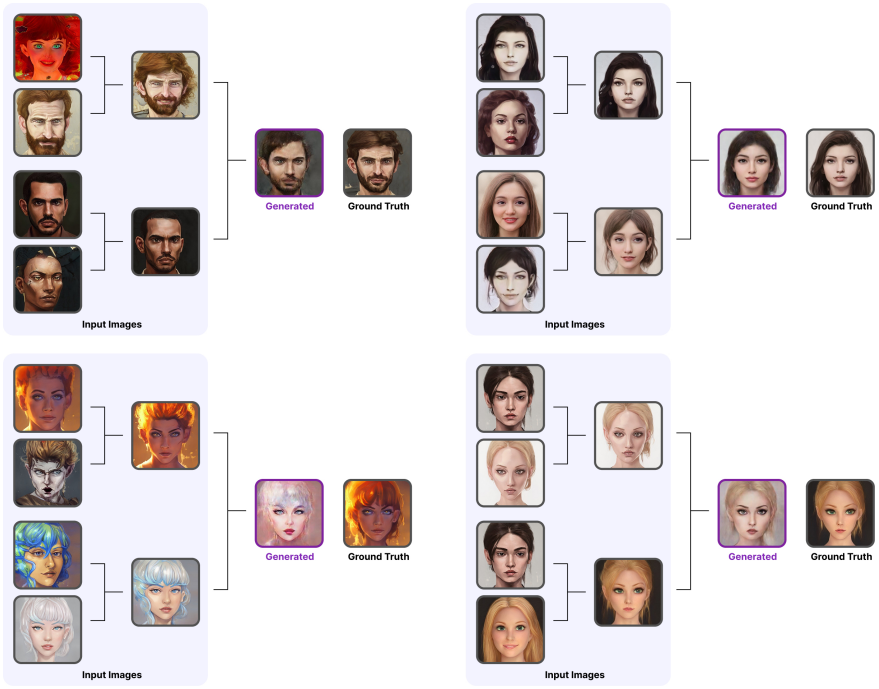

Figure 2: Qualitative results generated by our trained model.

Model Improvement. In Graph2Pix [2], the target image is modeled as noise during training, a setup that aligns well with the principles of diffusion models and may outperform GAN-based approaches. Therefore, we propose integrating the GCN module into the existing two-stage pipeline of the Latent Diffusion Model [5]. Instead of using a 3×3 convolution kernel, we can use the pretrained Variational Auto-Encoder (VAE) [6] to compute the image representation. We will then evaluate the performance of this hybrid approach against a reproduced Graph2Pix baseline.

References

[1] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023.

[2] D. Gokay, E. Simsar, E. Atici, A. Ahmetoglu, A. E. Yuksel, and P. Yanardag, "Graph2pix: A graph-based image to image translation framework," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 2001–2010.

[3] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl

[4] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021.

[6] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014. [Online]. Available: http://arxiv.org/abs/1312.6114