Text Guided Image Manipulation Project Report (Information Retrieval)

Abhuday Tiwari
MT22005
Dept. of Computer Science
IIIT-Delhi, India
abhuday22005@iiitd.ac.in

Amrita Aash
MT22011
Dept. of Computer Science
IIIT-Delhi, India
amrita22011@iiitd.ac.in

Kirti Vashishtha
MT22035
Dept. of Computer Science
IIIT-Delhi, India
kirti22035@iiitd.ac.in

Ritisha Gupta
MT22056
Dept. of Computer Science
IIIT-Delhi, India
ritisha22056@iiitd.ac.in

Nikhilesh Verhwani
MT22114
Dept. of Computer Science
IIIT-Delhi, India
nikhilesh22114@iiitd.ac.in

Shubham Agarwal
MT22124
Dept. of Computer Science
IIIT-Delhi, India
shubham22124@iiitd.ac.in

I. Problem Statement

Text-guided image manipulation is an image editing technique that manipulates a given image according to the natural language text descriptions. It is a rapidly growing technique in the field of NLP (Natural Language Processing) and CV (Computer Vision). Recent advancements in Deep Learning have opened doors to various image manipulation applications [1] [2]. Despite remarkable advances in image generation methods, general domain high-fidelity image editing still needs to be improved. We propose a more approachable technique that can automatically edit a given image using natural language descriptions.

II. Motivation

With an enormous increase in the volume of data, an automated system is required to manipulate multiple images rapidly, resulting in significant time savings and increased productivity. Businesses can use text-based image manipulation to create eye-catchy posters and graphics to attract customers. Users can create memes to convey humour and sarcasm on social media platforms. Real-life applications of text-guided image manipulation help marketers create engaging social media visual content and allow customers to implement the try-on feature before buying products. Areas of gaming and virtual reality can majorly benefit from this field by creating images and environments which are imaginary to humans yet realistic. Fields of architecture and interior design will also benefit by visualizing customers' spaces even before they are built.

It enables people without expertise to change the images using simple natural language instructions. Suppose there is a photograph of a beautiful sunset, and there is a powerline which destroys the beauty of the picture. Users who are not experts may need help removing the powerline so that it looks natural. However, by using text-guided image manipulation, they can type a command such as "remove power line", which will remove the powerline so that it blends seamlessly with the sky. It can open up new possibilities in art, design and photography that were previously time-consuming.

III. Literature Review

General Adversarial Networks [3] introduces a novel approach in which generative models are trained using two Neural Networks - the Generative Model (G Model) and a Discriminator Model (D model). Both models are trained together in minimax game fashion. The generator tries to minimize the probability of the sample generated by the discriminator whether the D-model correctly recognizes the samples as fake; apart from that discriminator tries to maximize the probability. Due to this competition, both improve their ability. The critical contribution of GANs is their ability to generate realistic and synthetic images similar to training images. On the contrary, GANs can be complex to train and demand precise hyperparameter adjustment to guarantee convergence. Additional challenges include poor inversion capability and the inability to manipulate novel images. Deep Unsupervised Learning using Nonequilibrium Thermodynamics [4] proposes a new approach to unsupervised deep learning that uses concepts from nonequilibrium thermodynamics. The fundamental idea is based on statistical physics, i.e. to slowly demolish a given data distribution structure by an iterative forward diffusion process by introducing noise to the system in a controlled manner, causing the data distribution structure to dissipate. After that, a reverse diffusion process gives data their original structure, producing a highly adaptable and manageable generative data model that can be used for various downstream tasks. The algorithm estimates the reversal of a Markov diffusion chain that transforms data into a noise distribution. The algorithm can learn any data distribution and is tractable, sample-able, and able to manipulate distributions. Diffusion Models are more reliable compared to any other model. Style-Based Generator Architecture for Generative

Adversarial Networks [5] proposes a new architecture of the GAN's [3] generative model, which provides a new method for controlling the image generation process. It comprises several convolutional layer blocks that transform a randomly generated latent vector into an output image. The key feature of the StyleGAN architecture is the ability to provide fine-grained control over the visual features of the generated snapshots, which implies it has precise control over various aspects of the generated images, such as their colours, textures, shapes, and other visual characteristics. Overall, the StyleGAN architecture is a significant advancement in GAN technology, allowing the creation of high-quality images with fine-grained control over visual features. Another notable work using GANs [3] includes" ManiGAN: Text-guided image manipulation." [6], where the author proposes a novel method for image manipulation that leverages the strength of NLP (natural language processing) and deep learning techniques. The new model combines a Generative Adversarial Network [3] with a text-to-image synthesis model. A text encoder module in ManiGAN converts textual descriptions of desired image manipulations into latent code fed into the generator network. This contrasts with conventional GANs, which produce images from a fixed set of predefined latent codes. The ManiGAN model presents a promising approach to text-guided image manipulation that has the potential to be applied in various domains, such as computer vision, visual storytelling, and advertising. One of the most popular loss functions used in fine-tuning the models of text-guided image manipulation is CLIP [7] loss which is inspired by the CLIP model. The model predicts which text caption is mapped with which image. The CLIP model is trained on 400 million image text pairs, which the paper's authors have curated. The main idea of the work is to learn perception from supervision present in natural language. The model is trained to maximize the similarity between the representations of positive pairs and minimize the similarity between the representations of opposing pairs. In other words, images and texts with similar semantic meanings will be mapped together. The model is fine-tuned on a task-specific dataset using a small amount of labelled data. To incorporate the GAN [3] model together with the CLIP [7] loss, StyleClip [8] proposed an architecture which will allow the manipulation of multiple attributes of images given by a complex text prompt. The architecture combines StyleGAN as a pre-trained GAN [3] model and CLIP [7] loss as the loss function. CLIP [7] loss helps to minimize the cosine similarity between the text prompt and the generated image. The novel idea of introducing multiple mappers or layers for different detail levels allows StyleClip to modify an image given multiple attributes with a good performance. As StyleClip depends on a pre-trained StyleGAN generator, unseen images outside the domain of StyleGAN, do not manipulate the images as expected. Also, only those text prompts already mapped to the pre-trained CLIP model will produce images with faithful manipulation. The out-of-domain limitation of StyleCLIP [8] is overcome by CLIP-Guided Domain Adaptation of Image Generators, StyleGAN-NADA [9]. The key idea in this paper is a training scheme that will shift the existing pre-trained domain to a new one so that unseen mages can also be manipulated. The proposed architecture uses two main components, StyleGAN2 [10] and CLIP [7] loss. Instead of one, two generators are used while training. In the training phase, it is fine-tuned to produce images different from one of the generators in the direction of the textual description. This direction is determined in the CLIP space between the source and target text prompts. As StyleGAN-NADA uses CLIP [7], it is limited by the concepts and the pre-trained data the model uses. Also, text prompts are limited due to ambiguity in the user's natural language. In other words, it might be challenging to understand the meaning of the prompt the user tries to convey.

IV. Novelty

One of the most significant difficulties in image manipulation is maintaining facial identity as seen in Fig.1. Since our dataset comprises human faces, we aimed to integrate the Face Identity Loss [12] function to preserve facial identity. The Face Identity Loss [12] function prevents any undesired modifications, and the extent of feature preservation of the input human face data depends on the weight added. Consequently, our final loss function is the sum of the CLIP [7] loss and the Face Identity Loss [12] function. Fig 11 suggests that diffusion in unable to maintain certain features such as eyes and smile during the manipulation process.

Fig. 1. The figure suggests that diffusion is unable to maintain certain features such as eyes and smile during the manipulation process.

V. Methodology

Our proposed methodology for text-guided image manipulation includes two major components: Diffusion [4] for image generation and CLIP [7] loss for fine-tuning the generation

model based on the text prompt. Fig.2 represents the flow of our above proposed approach and Fig.3 represents the image manipulation and retrieval.

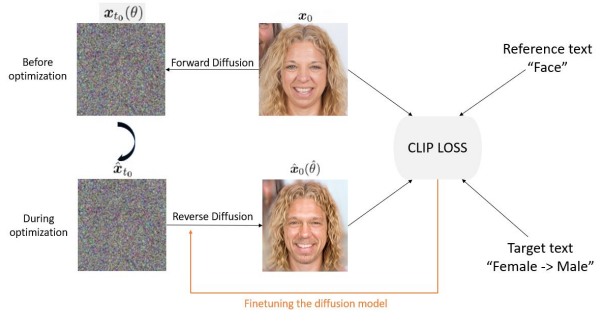

Fig. 2. Model Fine Tuning

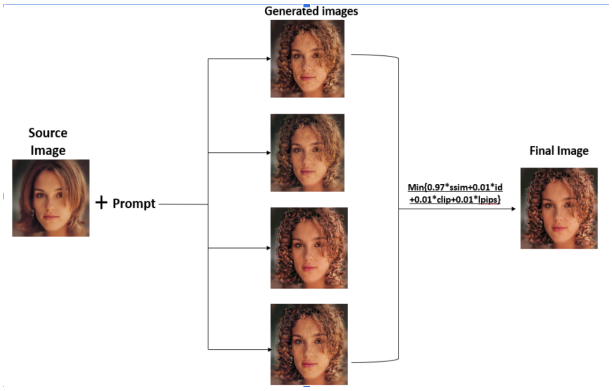

Fig. 3. Image Manipulation

The input image is first converted to latent vectors using the forward diffusion process by gradually adding Gaussian noise. The noise is iteratively removed from the sample in the reverse denoising process, accompanied by CLIP [7] loss. Only the reverse denoising step is performed iteratively to minimize the CLIP [7] loss and fine-tune the model. As DDIM (Denoising Diffusion Implicit Models) [11] is deterministic in both the forward and reverse diffusion process, it was adopted as it guarantees the reconstruction of the original image.

CLIP [7] and identity loss were used to fine-tune the model. The CLIP [7] loss used here is the local directional loss which aids in aligning the latent vectors of a pair of images to the latent vectors of a pair of texts, where the pair of images will be the original and generated image by a diffusion process. In contrast, the pair of texts is the reference text given to the original image and the target text prompt given to manipulate it. The reference texts are concise words used to refer to each input image. Identity loss aids in preserving the identity and detailing of the image after manipulation. It comprises two parts. The first is L1 loss which computes the loss between the original and manipulated image, and the other is the face identity loss. For different domains, the weights added to the above two parts vary; for example, for human face images retaining facial identity will be significant.

VI. Dataset

We trained our model on CelebA-HQ dataset https://github.com/IIGROUP/MM-CelebA-HQ-Dataset. For testing we used FFHQ dataset https://github.com/NVlabs/ffhq-dataset.

VII. Code

Link to github repository: https://github.com/agrawals1/IR_Project

VIII. Evaluation

A. Qualitative Analysis

The following figures compares the result of our baseline model with the final model against four text prompts. It also highlights our final models' qualitative performance. Fig.4 is comparison for the text prompt "tanned", Fig.5 is comparison for the text prompt "with makeup", Fig.7 is comparison for the text prompt "pixar", Fig.7 is comparison for the text prompt "female to male". Fig.8 shows more examples on different prompts.

Fig. 4. The first row corresponds to original images, the second row corresponds to baseline results for the prompt "tanned" and the third row corresponds to the final models' results for the prompt "tanned"

B. Quantitative Analysis

The table below shows the quantitative result for the LPIPS [13] score and the directional CLIP loss for the baseline model vs the final model. LPIPS score is the average score over 50 images. Github link to LPIPS code: https://github.com/richzhang/PerceptualSimilarity. We calculated directional clip loss for each prompt (100 images for each of the 4 prompts).

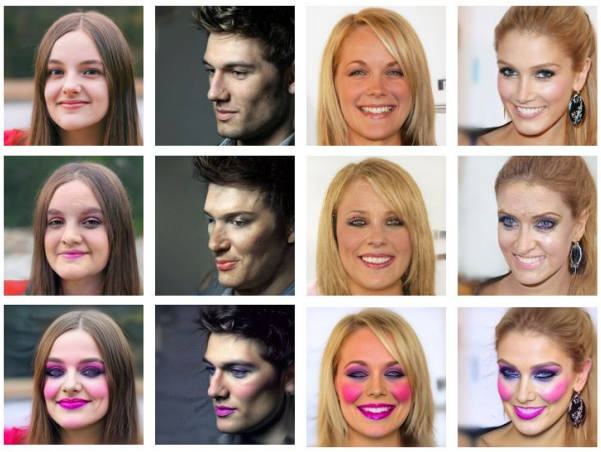

Fig. 5. The first row corresponds to original images, the second row corresponds to baseline results for the prompt ”with makeup” and the third row corresponds to the final models’ results for the prompt ”with makeup”

Fig. 6. The first row corresponds to original images, the second row corresponds to baseline results for the prompt ”pixar” and the third row corresponds to the final models’ results for the prompt ”pixar”

Fig. 7. The first row corresponds to original images, the second row corresponds to baseline results for the prompt ”female to male” and the third row corresponds to the final models’ results for the prompt ”female to male”

	Baseline Model	Final Model
Lpips score for 50 images (Original vs generated)	0.1992	0.1220
Directional Clip loss for ”tanned” prompt	0.6260	0.5674
Directional Clip loss for ”pixar” prompt	0.6104	0.5273
Directional Clip loss for ”with makeup” prompt	0.5903	0.5713
Directional Clip loss for ”female to male” prompt	0.5415	0.5239
Average Directional Clip Loss over 500 images	0.5920	0.5474

LPIPS, Learned Perceptual Image Patch Similarity is a metric used for measuring the similarity between two images. We use LPIPS score to evaluate the inversion capability of our model. Inversion in the context of image generation is the ability of the model to generate the same image from the latents of the input image. A higher value of LPIPS indicates that the two images are further or more different. A lower value indicates that the two images are structurally more similar.

Directional CLip Loss is the same which is explained in Methodology. The lower the loss the better.

As is seen from the table above the final model outperforms the baseline model in both the metrics, LPIPS score and directional clip loss.

C. Comparison with SOTA

We faced the following difficulties while comparing with the state of the art:-

1. Unavailibility of evaluation metric: The quality of editing can only be gauged through human evaluation and all the papers concerning image editing rely only on human inspection for evaluation. No quantitative metric is available for the same.

2. Unavailibility of codebase: due to unavailability of codebase of the sota (muse [14]), We could not generate images for the sota model, hence even human evaluation was not possible.

D. Results on diverse prompts

The following Fig.8 shows the manipulated images by our final model on various prompts. The results convey that our model is robust is handling diverse set of prompts.

Fig. 8. The first row corresponds to original images and the second row corresponds the final models' results for the given prompts mentioned below

REFERENCES

[1] Hong, S., Yan, X., Huang, T. S., Lee, H. (2018). Learning hierarchical semantic image manipulation through structured representations. Advances in Neural Information Processing Systems, 31.

[2] Bayar, B., Stamm, M. C. (2018). Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. IEEE Transactions on Information Forensics and Security, 13(11), 2691-2706.

[3] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial networks. Communications of the ACM, 63(11), 139-144.

[4] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015, June). Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning (pp. 2256-2265). PMLR.

[5] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4401-4410).

[6] Li, B., Qi, X., Lukasiewicz, T., & Torr, P. H. (2020). Manigan: Text-guided image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7880-7889).

[7] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

[8] Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., & Lischinski, D. (2021). Styleclip: Text-driven manipulation of stylegan imagery. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2085-2094).

[9] Gal, R., Patashnik, O., Maron, H., Chechik, G., & Cohen-Or, D. (2021). Stylegan-nada: Clip-guided domain adaptation of image generators. arXiv preprint arXiv:2108.00946.

[10] 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 (pp. 8110-8119).

[11] Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502.

[12] JiankangDeng,JiaGuo,NiannanXue,andStefanosZafeiriou. Arcface: Additive angular margin loss for deep face recogni- tion. In Proceedings of the IEEE/CVF Conference on Com- puter Vision and Pattern Recognition, pages 4690–4699, 2019.

[13] Zhang, Richard, et al. "The unreasonable effectiveness of deep features as a perceptual metric." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[14] Chang, Huiwen, et al. "Muse: Text-To-Image Generation via Masked Generative Transformers." arXiv preprint arXiv:2301.00704 (2023).