Text Guided Image Manipulation

Ayush Agrawal
MT22095
Dept. of Computer Science
IIIT-Delhi, India
ayush22095@iiitd.ac.in

Janak Kapuriya
MT22032
Dept. of Computer Science
IIIT-Delhi, India
kapuriya22032@iiitd.ac.in

Shubham Agrawal
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 robust technique that can automatically edit a given image using natural language prompts.

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. Marketers can 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.

The proposed solution 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 and cost-inefficient.https://www.overleaf.com/project/643e605a5469abac9e0caadb

III. Related Works

Deep Unsupervised Learning based Diffusion models [4] use concepts from non-equilibrium 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. 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. 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. 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.

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

We initially implemented two baselines for image manipulation - StyleClip [7] and TediGAN. [15] StyleClip modifies the latent vector to generate back the original image optimised for the given textual manipulation using clip loss. TediGAN applies three sequential processes for the achievement of our objective - inversion module, text-image similarity learning and instance level optimisation. 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.

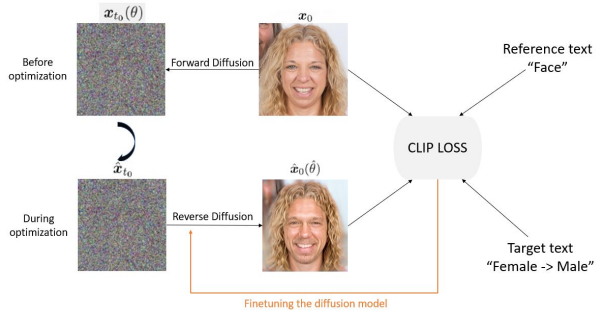

Fig. 2. Model Fine Tuning

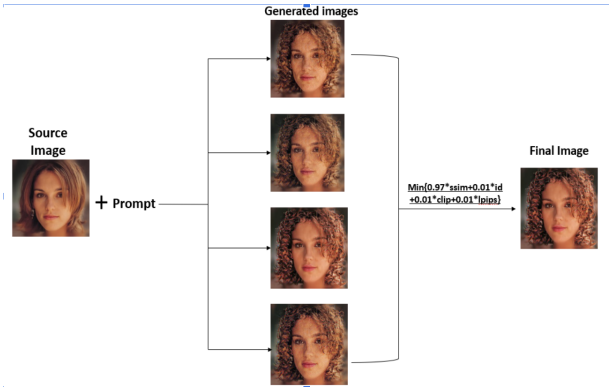

Fig. 3. Image Manipulation

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 models on CelebA-HQ dataset, which contains 30K high resolution face images and their associated caption explaining that image. Out of these, we used 50 samples for training our baselines and diffusion models. The dataset can be found at:https://github.com/IIGROUP/MM-CelebA-HQ-Dataset.

VII. Evaluation

We performed human-evaluation on the manipulated images generated by models. This qualitative method is inevitable as the models being trained on the same notion for manipulation effectiveness as training loss can’t be de facto used for evaluation itself.

We performed scoring of each manipulation done by each of the two baselines and the main model for the following metrics - clarity/sharpness, justifying prompt, and maintaining faithfulness in preserving the original image.

These were individually scored on a scale of 1-3 with 1 for bad, 2 for average, and 3 for good performance respectively. we took the mode of the values of each image's result. we did human annotation on 50 images with 3 different annotators.

Our Human Evaluation is given in the below Table.

	TediGAN	StyleCLIP	Final Model
Modification (P)	1	1	3
Faithfulness (P)	2	1	2
Sharpness (P)	2	3	3
Modification(F2M)	2	3	3
Faithfulness (F2M)	2	1	3
Sharpness (F2M)	2	2	2

In the above table, P denotes the Pixar prompt and F2M denotes the Female to Male prompt. As we can see Final model performs well compared to other models in all three components compared to two other models.

The complete evaluation results could be valued at the following location : Human Evaluation Data. The following figures compares the result of our baseline model with the final model against two text prompts. Fig.5 is comparison for the text prompt "pixar", Fig.5 is comparison for the text prompt "female to male".

Fig. 4. The first row corresponds to original images, the second row corresponds to baseline one, the third row corresponds to the final model and forth row corresponds to baseline two results for prompt "pixar".

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

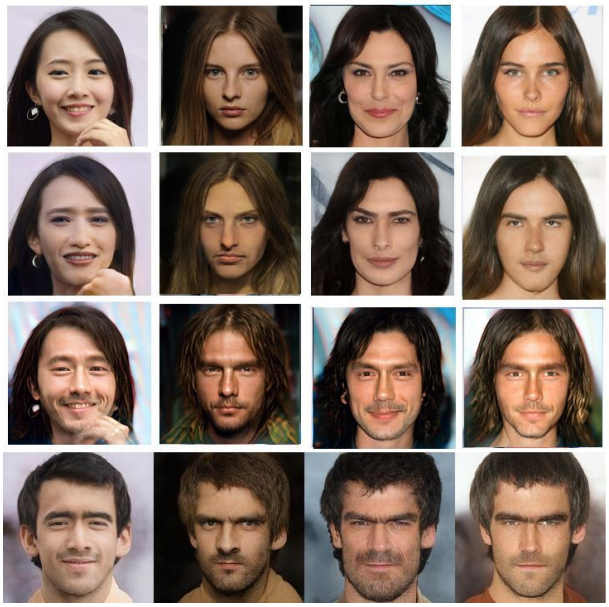

Fig. 5. The first row corresponds to original images, the second row corresponds to baseline one, the third row corresponds to the final model and the forth row corresponds to baseline two results for prompt "female to male"

VIII.

A. *Error Analysis*

As evident from the evaluation results on the given prompts to manipulate images, the baselines perform poorly on the faithfulness test. the manipulation of the images is not able to properly preserve the underlying distribution of the image and might completely change the image. However, our diffusion model is sufficiently robust and works well when measured for faithfulness, which is regularised using the faceID loss. Some models are failing to retain the sharpness.

Fig. 6. The first on the left side corresponds to the original images, and the right side image corresponds to the prompt Female to male. As we can see model fails to obtain a prompt on the original image as well as loose sharpness.

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).

[15] Xia, W., Yang, Y., Xue, J. H., Wu, B. (2021). Tedigan: Text-guided diverse face image generation and manipulation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2256-2265).