Domain transfer of sketched facial image into realistic facial image to prevent crime

Wonseok Oh
Electrical and Computer Engineering
University of Michigan, Ann Arbor
okong@umich.edu

Abstract

This project focuses on the development of an innovative framework to enhance modern forensic procedures by accurately recreating vivid images of a criminal from a provided sketch. In stage one, our approach relies on a technique called pSp Richardson et al. (2021), in which feature maps are derived from a pyramid network and introduced into a StyleGAN generator. The generator has undergone prior tutelage using datasets like CelebA-HQ Lee et al. (2020), and it's designed to recreate images from the provided sketches. Subsequently, the second phase involves stylizing the produced output according to particular instructions (for instance, 'blue eyes and brown hair') utilizing a mechanism known as InstructPix2Pix Brooks et al. (2023). To conclude, we not only enhance the existing pSp framework by incorporating an additional loss function but also conduct a comprehensive evaluation of the method with real-life sketch drawings.

1 Introduction

In the intricate field of crime detection and prevention, the effective identification of suspects remains a pressing issue. Traditional reliance on closed-circuit television (CCTV) footage, in spite of its ubiquity, often faces challenges due to the low resolution and limited datasets that hinder effective facial recognition. Concurrently, an alternative approach employed by law enforcement is the utilization of skilled artists to render high-quality sketches based on eyewitness accounts. Despite its value, this method relies heavily on subjective interpretation and does not result in an objective, detailed image which can be crucial for criminal identification. In light of these limitations, there is growing recognition of the need to enhance both the quantifiable accuracy and utilitarian value of the method to generate a suspect's image.

We propose a prospective approach to this problem by embracing the potential of generative AI. Recent advancements such as the Encoding networks Richardson et al. (2021) and InstructPix2Pix models Brooks et al. (2023) have demonstrated the ability to translate high-level descriptions into intricate colored portraits, a capability that we aim to tap into.

The objective here is to create a cohesive system that amalgamates the distinguished intricacy of artists' sketches with the precision of AI to form reconstructed, high-definition, colored images of suspects. By doing so, we aim to markedly improve the efficacy of crime detection and prevention without compromising the invaluable contribution of sketch artists. More broadly, we hope this work will instigate a constructive dialogue on the interface between art, cutting-edge technology, and law enforcement.

Our results suggest the proposed model not only successfully reconstructs detailed, colored portraits from sketches but also provides the potential to revolutionize the procedure of creating suspect montages, facilitating the collaboration between human creativity and advanced technology. By enhancing the overall process of sketch-to-image translation in suspect identification, our method promises to be a robust, innovative contribution to the existing landscape.

2 Related Works

2.1 Latent Space in GAN Inversion

With the development of GAN, many studies have emerged to control latent space. Many recent papers have used StyleGANs Karras et al. (2019, 2020, 2021) because of their efficient image generation performance and semantic abundance of latent space. GAN Inversion is a method of extracting the latent vector from the image for the generator. This GAN Inversion method has suggested $\mathcal{W}$ and its extension, $\mathcal{W}+$ Abdal et al. (2019) space, depending on the shape of the target latent space for converting.

Our model employs the GAN inversion technique to create an encoder that performs an inversion in $\mathcal{W}+$ space and then uses the created latent vector to turn it back into an image.

2.2 Inversion Method

There are two methods of inversion. First, the optimization-based method directly tunes the parameter of the generator to express the target image. It exhibits high-quality results without additional learnable parameters, but it takes a long time. I2S Abdal et al. (2019) embeds the image into the extension latent space $\mathcal{W}+$ of StyleGAN. The author of StyleGAN2 Karras et al. (2020) proposed a method of optimizing the noise map corresponding to each layer of the synthesis network and latent code. PTI Roich et al. (2021) performs pivotal tuning in the latent space for real images. It finds the latent code representing the image, then fixes it and fine-tunes the generator to derive the results.

Second, learning-based methods train encoders that create latent codes that can represent real images. Although there is a time advantage because a single inference produces results, it usually results in poor visual quality compared to optimization methods. pSp Richardson et al. (2021) and e4e Tov et al. (2021) directly extract the late vector of $\mathcal{W}+$ space using the encoder network. It extended the GAN inversion problem to the image-to-image translation problem by changing the input image because only the encoder network needs to be learned. Another study Restyle Alaluf et al. (2021) also improves visual quality by allowing encoders to modify latent codes repeatedly.

In our work, we apply networks such as pSp and Restyle to use a method that allows us to learn the encoder once and continue to get results quickly. Recently, studies Alaluf et al. (2022); Dinh et al. (2022) using a hypernetwork that allows the addition of offset to the generator parameter have revealed promising results.

2.3 Attention Mechanizm

The attention mechanism was developed to handle long sequences of data in Neural Machine Translation (NMT). "Attention is All You Need" Vaswani et al. (2017). introduced the Transformer model, using a self-attention mechanism, allowing models to focus on different parts of the input, enhancing accuracy and efficiency.

Convolutional Block Attention Module (CBAM)) Woo et al. (2018). then improved the attention mechanism by introducing attention modules on both channel and spatial aspects in CNNs, allowing the models to focus on more representative features in the data.

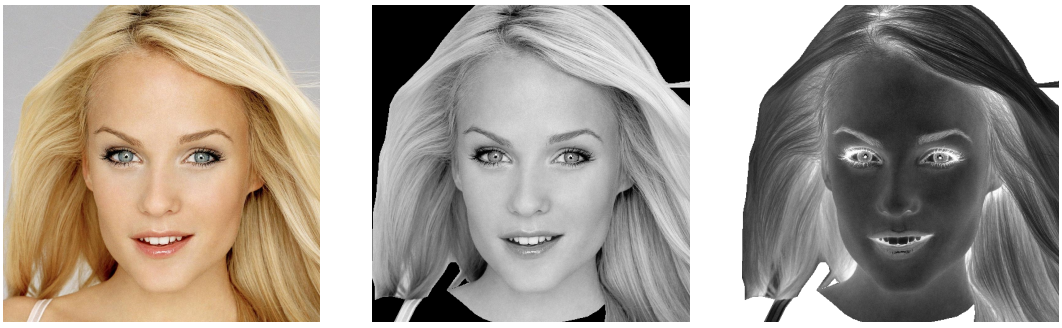
We used CBAM to make the new loss function called attention loss to train the encoder. This extracts the attention map from the images and compares the distance.

3 Data

3.1 CelebA-HQ dataset

The base dataset is the CelebA-HQ dataset for our study. This data is the image part of the CelebAMask-HQ dataset Lee et al. (2020). [1] This high-quality, large-scale celebrity identification database was carefully curated by researchers at the Chinese University of Hong Kong and has been pivotal in several ground-breaking advancements in the field of facial recognition. This

[1] https://mmlab.ie.cuhk.edu.hk/projects/CelebA/CelebAMask_HQ.html



(a) Original Image (b) Segmented Image (c) Inverted Image

(d) Blurred Image (e) Final Sketch (f) Canny Edge Detector

Figure 1: Process of building sketches

data can be easily downloaded through the URL on the official site of the Chinese University of Hong Kong. The comprehensiveness of CelebA-HQ is reflected in its sheer volume and variety, consisting of 30,000 high-resolution celebrity images. Each image is meticulously detailed with numerous attributes, making it a potent tool for data-driven studies. The dataset equally represents approximately 10,202 unique identities, covering diverse demographic parameters such as age and ethnicity. Beyond the quantitative aspects, the CelebA-HQ dataset stands out for its qualitative richness. Each image is annotated with 40 different attribute labels, capturing a broad array of features such as hair color, gender, and presence of accessories. This meticulous detail allows for in-depth analysis - from identifying distinctive features obscured by eyewear to detecting nuanced facial expressions, thus painting a comprehensive picture of the individual in question.

3.2 Preprocessing sketches

In our study, we employ key computer vision methodologies during the preprocessing phase when examining sketches. These include both segmentation and filtering techniques with the core objective of creating refined sketches of images from the CelebA-HQ dataset. The process emulates the approach of delineating the contours of the face.

Our first step involves extracting the foreground, i.e., the area of interest from the background. This is accomplished through segmentation, leveraging the annotated data from the CelebA-HQ dataset. The image is then inverted for contrast, and subsequently divided by a Gaussian blurred image of the segmented original, further aiding in the enhancement of edge detection.

Upon normalization of the image within the range of $[0, 255]$, we obtain the desired sketch representation. Figure 1 illustrates this step-by-step process in a comprehensive manner.

Moreover, our method has demonstrated superior performance in relation to the traditional Canny edge detector, as evidenced in Figure 1f. The prime reason behind this enhanced performance is our method's ability to minimize noise in the image, a factor considerably influencing the quality of the extracted sketch.

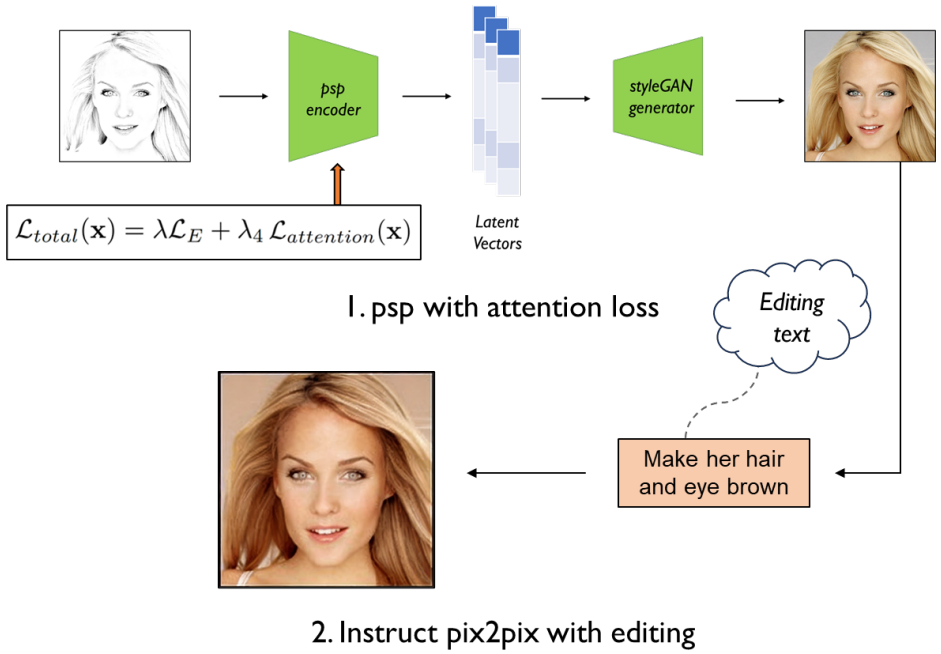


Figure 2: Overall model structure

An important component in the Gaussian blurred image is the tuning parameter σ. Empirical analysis revealed that the most effective results were achieved when the σ parameter was set approximately to a quarter of the image size, which is 255, thereby generating sketches of optimum quality.

4 Approach

Overall Approach Our model comprises two modules 2. The first, the Image2Image translation module, generates a preliminary draft image after training through a pSp encoder Richardson et al. (2021) using sketch data. This draft image purely reflects the result of learning, without any additional information verified by the witness. Therefore, the second module inputs editing text, transforming the image in accordance with the text information to yield the final result. This two-fold process allows for a highly adaptive and purposeful image generation, demonstrating the synergistic interplay between artificial intelligence and graphic rendering.

4.1 Image2Image translation model

pSp Framework. To reconstruct a given sketch with high image quality, we utilize a prominent example of image translation, namely pixel2style2pixel (pSp) Richardson et al. (2021). pSp is built on a pre-trained StyleGAN encoder which can directly encode the real image into the latent domain, thus handling a variety of tasks without following the "invert first, edit later" standard. Motivated by this paper, we use two versions of the pSp model (one as the baseline and one with an additional attention loss) to reconstruct sketches generated by our method and compare the results with ground truth images.

Convolutional Block Attention Module (CBAM) Our attention loss uses the Convolutional Block Attention Module (CBAM) Woo et al. (2018) which takes two parts: a channel attention module and a spatial attention module 3. The channel attention module computes a one-dimensional channel attention map, denoted as $\mathbf{M_c} \in \mathbb{R}^{C\times1\times1}$, which assigns significance to distinct channels within the feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$.

On the other hand, the spatial attention module estimates a two-dimensional spatial attention map, denoted as $\mathbf{M_s} \in \mathbb{R}^{1\times H\times W}$, which underscores essential regions in the feature map.

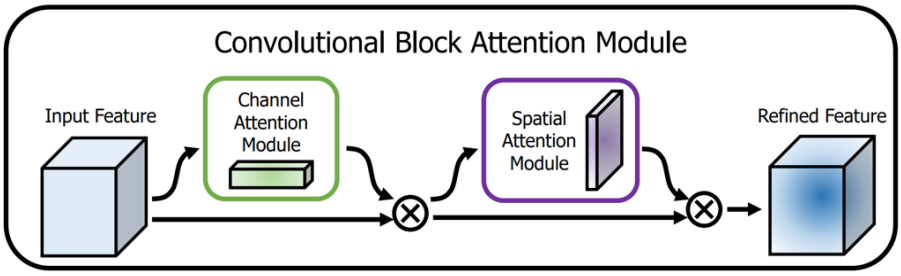


Figure 3: Structure of CBAM module

This process generates channel context descriptors: $\mathbf{F}_{avg}$ and $\mathbf{F}_{max}$. It focuses primarily on the "what" of the feature map. The spatial attention module instead emphasizes the "where" of the feature map, identifying spatial significance through average and max pooling operations. This results in the production of two spatial context descriptors which are combined and processed by a convolutional layer to produce the spatial attention map $\mathbf{M_s}(\mathbf{F}')$. Finally, the attention map is used to calculate the attention loss as follows:

$$\mathbf{F}'(\mathbf{x}) = \mathbf{M_c}(\mathbf{F}) \otimes \mathbf{F}$$
$$\mathbf{F}''(\mathbf{x}) = \mathbf{M_s}(\mathbf{F}') \otimes \mathbf{F}$$

$$\mathcal{L}_{attention}(\mathbf{x}) = ||\mathbf{F}''(\mathbf{x}) - \mathbf{F}''(G(E(\mathbf{x})))||_2$$

Loss Functions. Our framework utilizes a weighted combination of several objectives to train the encoder. The pixel-wise $\mathcal{L}_2$ loss is utilized as follows:

$$\mathcal{L}_2(\mathbf{x}) = ||\mathbf{x} - D(E(\mathbf{x}))||_2$$

To learn perceptual similarities, we incorporate the $\mathcal{L}_{LPIPS}$ loss, which has been shown to better preserve image quality compared to the standard perceptual loss:

$$\mathcal{L}_{LPIPS}(\mathbf{x}) = ||F(\mathbf{x}) - F(D(E(\mathbf{x})))||_2$$

Here, F(·) denotes the perceptual feature extractor. In addition, to encourage the encoder to output latent style vectors closer to the average latent vector, we include the following regularization loss:

$$\mathcal{L}_{reg}(\mathbf{x}) = ||E(\mathbf{x}) - \bar{\mathbf{w}}||_2$$

To tackle the challenge of preserving the input identity when encoding facial images, we incorporate a dedicated recognition loss measuring the cosine similarity between the output image and its source:

$$\mathcal{L}_{ID}(\mathbf{x}) = 1 - \langle(\mathbf{x}), R(D(E(\mathbf{x})))\rangle$$

The loss function defined up to this point is referred to as the encoder loss and is defined as:

$$\mathcal{L}_E(\mathbf{x}) = \lambda_1\mathcal{L}(\mathbf{x}) + \lambda_2\mathcal{L}_{LPIPS}(\mathbf{x}) + \lambda_3\mathcal{L}_{ID}(\mathbf{x}) + \lambda_4\mathcal{L}_{reg}(\mathbf{x})$$

So eventually the attention map is used to make the attention loss. Therefore the overall loss function is as follows,

$$\mathcal{L}_{total}(\mathbf{x}) = \lambda\mathcal{L}_E + \lambda_4\mathcal{L}_{attention}$$

Here, these hyperparameters λ are set as different values when performing different tasks. For example, we usually set the weight of $\mathcal{L}_2$ to be 1 for the StyleGAN inversion, while it is set to be 0.01 for the face part when performing face frontalization tasks. Therefore the loss function used in the pretrained model is unlikely to be optimal for our sketch reconstruction task. Given this, we train our model from scratch on Great Lakes including a new loss function ($\mathcal{L}_{attention}$) and compare them with the baseline.

4.2 Stylizing Montages

InstructPix2Pix. Our objective after reconstructing images from sketches is to derive stylized outputs. We utilize a model known as InstructPix2Pix Brooks et al. (2023) which, when provided with specific instructions such as "change the hair color to brown", has the ability to modify images accordingly. This model combines the capabilities of a vast language model, GPT-3, and a text-to-image diffusion model, Stable Diffusion, to generate an extensive dataset of various image editing examples.

5 Experiments

5.1 Experiments Process

We provided an accessible GitHub repository at the given URL [2] to evaluate and deliver a comprehensive view of our experiments. This repository extends the fundamental framework of pSp Richardson et al. (2021) and adds a few additional files. These include "*models/attentionmodule/MODELS*" and "*models/attentionmodule/MODELS/cbam.py*" for CBAM Woo et al. (2018), "*utils/img2sketch.py*" to manage the sketch preprocessor, and "metrics.py", which is used for evaluation purposes. Moreover, this contained instructpix2pix Brooks et al. (2023) module which has the purpose of visualizing individual results and facilitating stylization.

In terms of our experiment parameters, we've conducted training from scratch for the baseline as well as all other model variants, ours included. During this process, we adopted an 8:1:1 ratio for arranging the train/val/test datasets, implemented a batch size of 4, and executed 100,000 training steps on the Great Lakes computing cluster. The difference in the model setup revolves around the new loss's ($\mathcal{L}_{attention}$) weight, denoted as λ_4.

5.2 Quantitative Results

Model	Runtime	MSE ↓	LPIPS ↓	Similarity −
pSp	0.0398 ± 0.1031	0.0780	0.291	0.340
Ours ($\lambda_4 = 1$)	0.0478 ± 0.0931	0.0777	**0.288**	**0.340**
Ours ($\lambda_4 = 1e3$)	0.0390 ± 0.0710	**0.0773**	0.289	0.341
Ours ($\lambda_4 = 1e6$)	0.0353 ± 0.1004	0.0789	0.295	0.341
Ours ($\lambda_4 = 1e9$)	0.0345 ± 0.1074	0.0794	0.294	0.341

Table 1: Quantitative Comparison

As per the data delineated in Table 1, it is discernible that our attention loss offers a subtle enhancement in performance related to Mean Squared Error (MSE) and Learned Perceptual Image Patch Similarity (LPIPS). A slight performance improvement was observed when the hyperparameter lambda was assigned a value of 1 or $1e3$. The value of LPIPS exhibited marginal disparity, however, the MSE value witnessed a more conspicuous deviation. Such variance evinced that maintaining the hyperparameter of the attention loss at 1e3 was indeed substantial and impactful. The similarity loss indicates the cosine similarity between the output image and its source based on the pre-trained ArcFace network.

5.3 Qualitative Results

CelebA-HQ We provide a qualitative comparison of various models alongside the stylization of our reconstructed images, as depicted in Figure 4. Our approach, with $\lambda = 1e3$, demonstrates superior capacity in the retrieval of intricate details such as eye shadow on the image featured in the first row. Additionally, the application of stylization using given prompts yields intuitive outcomes. This method enables us to modify the image towards a preferred output, potentially aligning it more closely with the witness's account.

[2] https://https://github.com/prbs5kong/sketch2face

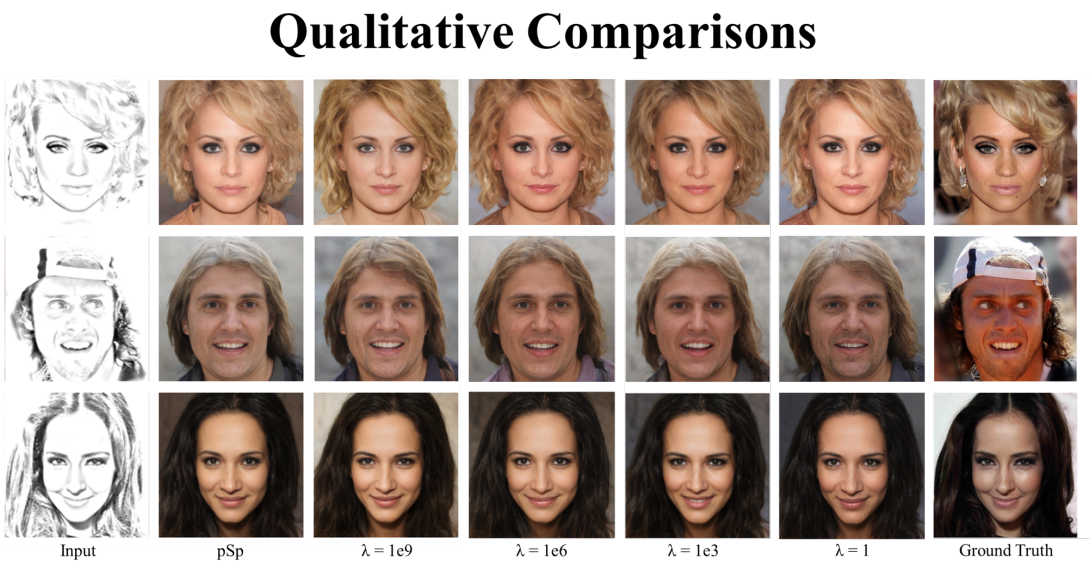


Figure 4: Evaluation on CelebA-HQ dataset

Real Sketches Our technique has been tested on authentic sketches acquired from the internet and created by montage artists. Since these images are literal internet-derived montages, we lack their ground truths. However, the multimodal inheritance attribute of the pSp encoder permits us to generate a range of plausible outputs for the sketch. The results can be referenced in Figure 5.

Figure 5: Mutilmodal Images from realsketch

5.4 Discussion

Accessories While the pSp network provides good reconstruction results from sketch, it still has some limitations to our project. The network occasionally fails in distinguishing faces from accessories which can affect the process of reconstructing montages as in Figure 6.

Information Leak We trained our network on computed sketches, which are sketches that contain information from original images. However, the drawings by sketch artists are not directly drawn from prior knowledge of the ground truth and thus underperform as shown in the previous real application. We thus leave the generalization of image-to-image translation to real datasets as future work.

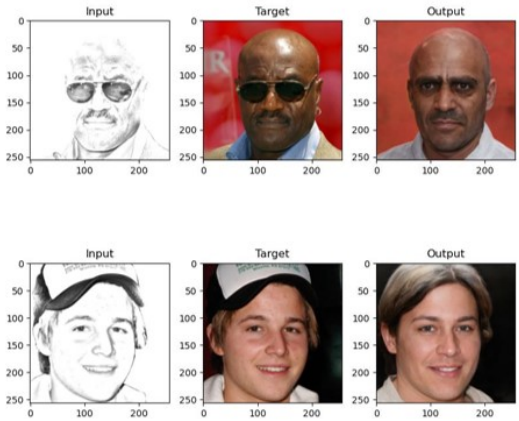


Figure 6: Failure Case

6 Conclusion

Our project brings forward a novel framework for creating an accurate visual representation of criminals from sketches, using the pSp Richardson et al. (2021) image translation technique and the InstructPix2Pix Brooks et al. (2023) mechanism for specific image styling. As we have seen, the attention loss function we incorporated showed promising improvements in performance, as evidenced by our Mean Squared Error (MSE) and Learned Perceptual Image Patch Similarity (LPIPS) scores.

However, some limitations were noticed, particularly when dealing with accessories and faces during the image reconstruction process and in generalizing the translation approach to real datasets. Nonetheless, when broadly applied, our technique demonstrated encouraging potential, even when tested with real sketches from the internet, indicating promising avenues for future research.

References

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.

Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021.

Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18511–18521, 2022.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

Tan M. Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. Hyperinverter: Improving stylegan inversion via hypernetwork. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, 2021.

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021.

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *arXiv preprint arXiv:2102.02766*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module, 2018.