Learning an Encoder for Image Editing within a StyleGAN Latent Space

Pascal Herrmann
pascal.herrmann@tum.de

Tobias Zengerle
tobias.zengerle@tum.de

1. Introduction & Related Work

StyleGAN [2] is a state of the art network for synthesizing realistic looking face images. The random nature of the latent code sampling however doesn't allow for any control over the type of output that should be generated. The method proposed by [6] introduces a novel approach to learn a GAN-Inversion encoder to embedd given images into the StyleGAN latent space. Their method is supervised in the image-space and regularized by a discriminator. In this work we present a training approach based on this new type of latent space embedding. It allows us to generate StyleGAN latent space embeddings based on easy to control input constraints for the desired face shape and appearance. We will explain our method and evaluate its performance and quality in the context of face image manipulation.

2. Method

We train an encoder network that takes as input a *source image portrait* and a *target landmark image*, depicting the outlines of the desired face expression and pose. The network output should be the latent embedding code corresponding to the manipulated image, i.e., the target landmarks applied to the source image. Similar to [6], the latent code is passed through a pre-trained StyleGAN-network resulting in an output image, which is then evaluated by a discriminator network. The proposed architecture is depicted in figure 1.

Latent Difference Encoder. In order to minimize the entanglement of the embedding and manipulation task, we decided to split the encoder into two parts. Using a pre-trained inversion-network from [6], we first embed the source image into the StyleGAN latent space, obtaining the source image's latent representation z. Then, inspired from [4], we pass this embedded latent code z, alongside with the target landmark image into the actual encoder network. This trainable encoder will return a *difference vector*, representing the required "shift" in latent space to perform the image manipulation. By adding this difference vector to the source image's latent code z, we obtain the *manipulated* latent code $\hat{z}$.

Conditional Discriminator. In order to ensure the generated images look realistic, and at the same time incorporate the target landmarks, we train the network with a conditional discriminator [3]. The discriminator receives a target landmark image and the corresponding created manipulated image as input. Its output score represents the realism of the fed-in real- or fake-pair. The resulting GAN-loss L_{adv} enables us to maintain overall image quality and simultaneously enforce the adoption of the desired landmarks.

Cycle Consistency Training. To make the model learn how to preserve the appearance of the person from the source image we use a Cycle Consistency based approach.After the source image has been manipulated to match the pose and expression of the target landmark, we then pass this manipulated image with its original source landmarks to the same encoder, to map it back to the initial pose & expression. Ideally the model should be able to carry over the appearance of the person from the source image throughout these two manipulation passes and we arrive at a recreation of the original image. To assure this we employ a L2 cyle-reconstruction loss L_{rec_cycle} between the latent space embeddings of the source image and the cycle-reconstructed image.

Furthermore, we also apply the adversarial loss on the cycle-reconstructed image, such that the total loss of the Encoder is: $L_{Enc} = L_{adv\_manipulated} + L_{rec\_cycle} + L_{adv\_cycle}$

3. Experiments & Results

We apply our experiments on the FFHQ dataset. Due to limited computational resources, we work with a downscaled version of resolution 128×128 pixels. Additionally, we use an off-the-shelf face-alignment network [1] to extract facial keypoints from the portrait images. These keypoints were used to create the target landmark images for our dataset, similar to [5]. Since the publicly available versions of StyleGAN [2] as well as the inversion network from [6] are only available for higher resolutions we trained both these networks ourselves at the beginning of the project.

Metrics. We evaluate our models using five quantitative metrics: Firstly, we use the Frechet-Inception-Distance (FID) to assess the realism of the generated images. Besides evaluating FID for the manipulated images, we

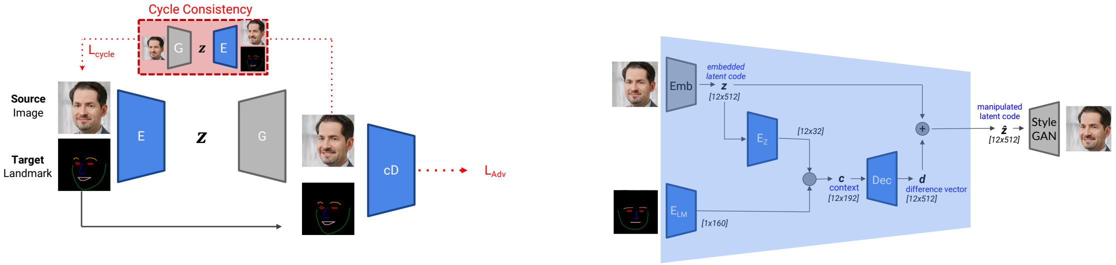

Figure 1. **Overall Architecture** (left) **In-Depth View of the Latent Difference Encoder** (right)

also compute the FID for the reconstruction capability of our network, by passing in the images with their matching landmarks to the encoder, such that ideally no manipulation should be performed. In order to evaluate how well the desired pose and expression are adapted in the manipulated output images, we use the Hausdorff Distance between the target landmarks and the extracted landmarks from the manipulated image. Lastly, we use the cosine similarity (CSIM) between the Facenet-embeddings of the pairs of the source image and the generated manipulated image, to measure the preservation of the person's appearance. The results are depicted in the following table.

Evaluation. To the best of our knowledge, there are no existing methods for our exact task, "conditioning a pre-trained StyleGAN with landmark images". Therefore, we focus on an ablation study to investigate the effects of our architectural choices.

We compare our proposed *Latent Difference Encoder* with two other encoder architectures. Another difference based encoder that directly takes the face *keypoint vectors* as input instead of the target landmark images and a *Naive Encoder* structure, where the portrait and landmark input images are simply concatenated before being fed into a simple CNN Encoder. Further we analyze the impact of applying the cycle reconstruction loss in *image space* as well as in *latent space*.

We come to the result, that the full proposed architecture works best, and outperforms methods that train a naive encoder-network that jointly performs embedding and manipulation at the same time. We also observe that performing the cycle-reconstruction loss in the latent space leads to better results than performing it in the image space. We conduct a qualitative analysis of the different architectures. The results of our proposed architecture are displayed in figure 2. More qualitative results can be found in our code repository.

Model	**FID** ↓ (Recon.)	**FID** ↓ (Manip.)	**Haus-dorff** ↓	**CSIM** ↑
GAN-Embedding	37.4	-	-	-
Naive Enc.	60.2	81.3	0.221	0.263
Naive Enc. + Cycle Consistency (Image Space)	87.7	84.3	**0.172**	0.235
Latent Diff. Enc. + Cycle Consistency (Image Space)	36.8	42.9	0.192	0.438
Keypoint Enc. + Cycle Consistency (Latent Space)	38.1	38.3	0.182	0.459
Latent Diff. Enc. **+ Cycle Consistency** **(Latent Space)**	**33.2**	**36.7**	0.186	**0.480**

Table 1. Results of the Ablation Study. (for comparison: The pre-trained StyleGAN we use has an FID Score of 17.1)

4. Conclusion

Our experiments have shown that a cGAN architecture can be used to train an encoder for effective latent space manipulation for a pre-trained StyleGAN model, based on shape and apperance input images. We also show that using a pretrained embedding network to disentangle the learning of appearance identity and shape leads to significantly better results than a joint naive encoding architecture. Naturally the results of our approach depend heavily on the quality of the pre-trained StyleGAN- and Embedding-network. While we had to train these networks from scratch in limited time, we expect that more fine-tuned models, such as the officially provided pre-trained models for higher resolutions, will lead to better results.

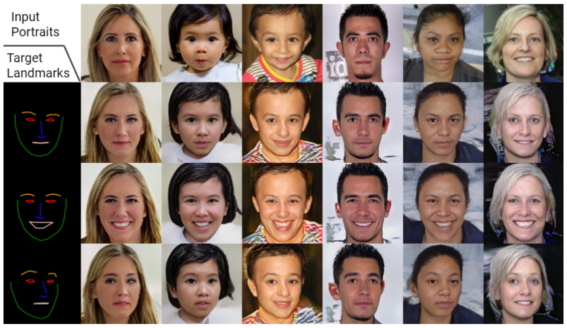

Figure 2. Results of our full proposed architecture

References

[1] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks), 2017.

[2] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2018.

[3] Takeru Miyato and Masanori Koyama. cgans with projection discriminator, 2018.

[4] Ayush Tewari, Mohamed Elgharib, and Gaurav Bharaj. Stylerig: Rigging stylegan for 3d control over portrait images, cvpr 2020, june 2020.

[5] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor S. Lempitsky. Few-shot adversarial learning of realistic neural talking head models, 2019.

[6] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing, 2020.

Additional Resources

Code Repository

https://github.com/pascalherrmann/ADL4CV-Project

Final Presentation Video

https://www.youtube.com/watch?v=6K5CvxyZG74&feature=youtu.be