StyleGAN2 Distillation for Feed-forward Image Manipulation

Yuri Viazovetskyi[⋆1], Vladimir Ivashkin[⋆1,2], and Evgeny Kashin[⋆1]

[1] Yandex
[2] Moscow Institute of Physics and Technology
{iviazovetskyi,vlivashkin,evgenykashin}@yandex-team.ru

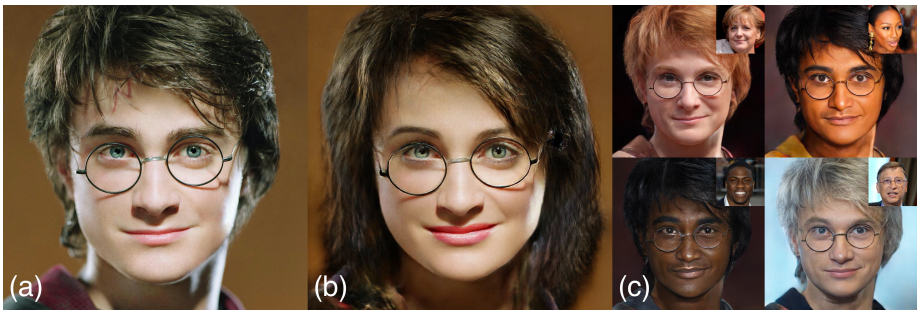


Fig. 1: Image manipulation examples generated by our method from (a) source image sampled from Celeba-HQ: (b) gender swap at 1024x1024 and (c) style mixing at 512x512. Samples are generated feed-forward, StyleGANv2 which we distilled was trained on FFHQ

Abstract. StyleGAN2 is a state-of-the-art network in generating realistic images. Besides, it was explicitly trained to have disentangled directions in latent space, which allows efficient image manipulation by varying latent factors. Editing existing images requires embedding a given image into the latent space of StyleGAN2. Latent code optimization via backpropagation is commonly used for qualitative embedding of real world images, although it is prohibitively slow for many applications. We propose a way to distill a particular image manipulation of StyleGAN2 into image-to-image network trained in paired way. The resulting pipeline is an alternative to existing GANs, trained on unpaired data. We provide results of human faces' transformation: gender swap, aging/rejuvenation, style transfer and image morphing. We show that the quality of generation using our method is comparable to StyleGAN2 backpropagation and current state-of-the-art methods in these particular tasks.

Keywords: Computer Vision, StyleGAN2, distillation, synthetic data

⋆ equal contribution

1 Introduction

Generative adversarial networks (GANs) [18] have created wide opportunities in image manipulation. General public is familiar with them from the many applications which offer to change one's face in some way: make it older/younger, add glasses, beard, etc.

There are two types of network architecture which can perform such translations feed-forward: neural networks trained on either paired or unpaired datasets. In practice, only unpaired datasets are used. The methods used there are based on cycle consistency [61]. The follow-up studies [24, 11, 12] have maximum resolution of 256x256.

At the same time, existing paired methods (e.g. pix2pixHD [55] or SPADE [42]) support resolution up to 2048x1024. But it is very difficult or even impossible to collect a paired dataset for such tasks as age manipulation. For each person, such dataset would have to contain photos made at different age, with the same head position and facial expression. Close examples of such datasets exist, e.g. CACD [8], AgeDB [40], although with different expressions and face orientation. To the best of our knowledge, they have never been used to train neural networks in a paired mode.

These obstacles can be overcome by making a synthetic paired dataset, if we solve two known issues concerning dataset generation: appearance gap [22] and content gap [28]. Here, unconditional generation methods, like StyleGAN [30], can be of use. StyleGAN generates images of quality close to real world and with distribution close to real one according to low FID results. Thus output of this generative model can be a good substitute for real world images. The properties of its latent space allow to create sets of images differing in particular parameters. Addition of path length regularization (introduced as measure of quality in [30]) in the second version of StyleGAN [31] makes latent space even more suitable for manipulations.

Basic operations in the latent space correspond to particular image manipulation operations. Adding a vector, linear interpolation, and crossover in latent space lead to expression transfer, morphing, and style transfer, respectively. The distinctive feature of both versions of StyleGAN architecture is that the latent code is applied several times at different layers of the network. Changing the vector for some layers will lead to changes at different scales of generated image. Authors group spatial resolutions in process of generation into coarse, middle, and fine ones. It is possible to combine two people by using one person's code at one scale and the other person's at another.

Operations mentioned above are easily performed for images with known embeddings. For many entertainment purposes this is vital to manipulate some existing real world image on the fly, e.g. to edit a photo which has just been taken. Unfortunately, in all the cases of successful search in latent space described in literature the backpropagation method was used [1, 2, 16, 31, 47]. Feed-forward is only reported to be working as an initial state for latent code optimization [5]. Slow inference makes application of image manipulation with StyleGAN2 in production very limited: it costs a lot in data center and is almost impossible

to run on a device. However, there are examples of backpropagation run in production, e.g. [48].

In this paper we consider opportunities to distill [21, 4] a particular image manipulation of StyleGAN2 generator, trained on the FFHQ dataset. The distillation allows to extract the information about faces' appearance and the ways they can change (e.g. aging, gender swap) from StyleGAN into image-to-image network. We propose a way to generate a paired dataset and then train a "student" network on the gathered data. This method is very flexible and is not limited to the particular image-to-image model.

Despite the resulting image-to-image network is trained only on generated samples, we show that it performs on real world images on par with StyleGAN backpropagation and current state-of-the-art algorithms trained on unpaired data.

Our contributions are summarized as follows:

- We create synthetic datasets of paired images to solve several tasks of image manipulation on human faces: gender swap, aging/rejuvenation, style transfer and face morphing;
- We show that it is possible to train image-to-image network on synthetic data and then apply it to real world images;
- We study the qualitative and quantitative performance of image-to-image networks trained on the synthetic datasets;
- We show that our approach outperforms existing approaches in gender swap task.

We publish all collected paired datasets for reproducibility and future research: https://github.com/EvgenyKashin/stylegan2-distillation.

2 Related work

Unconditional image generation Following the success of ProgressiveGAN [29] and BigGAN [6], StyleGAN [30] became state-of-the-art image generation model. This was achieved due to rethinking generator architecture and borrowing approaches from style transfer networks: mapping network and AdaIN [23], constant input, noise addition, and mixing regularization. The next version of StyleGAN – StyleGAN2 [31], gets rid of artifacts of the first version by revising AdaIN and improves disentanglement by using perceptual path length as regularizer.

Mapping network is a key component of StyleGAN, which allows to transform latent space $\mathcal{Z}$ into less entangled intermediate latent space $\mathcal{W}$. Instead of actual latent $z \in \mathcal{Z}$ sampled from normal distribution, $w \in \mathcal{W}$ resulting from mapping network $f : \mathcal{Z} \to \mathcal{W}$ is fed to AdaIN. Also it is possible to sample vectors from extended space $\mathcal{W}+$, which consists of multiple independent samples of $\mathcal{W}$, one for each layer of generator. Varying w at different layers will change details of generated picture at different scales.

Latent codes manipulation It was recently shown [17, 27] that linear operations in latent space of generator allow successful image manipulations in a variety of domains and with various GAN architectures. In GANalyze [17], the attention is directed to search interpretable directions in latent space of BigGAN [6] using MemNet [32] as "assessor" network. Jahanian et al. [27] show that walk in latent space lead to interpretable changes in different model architectures: BigGAN, StyleGAN, and DCGAN [43].

To manipulate real images in latent space of StyleGAN, one needs to find their embeddings in it. The method of searching the embedding in intermediate latent space via backprop optimization is described in [1, 2, 16, 47]. The authors use non-trivial loss functions to find both close and perceptually good image and show that embedding fits better in extended space $\mathcal{W}+$. Gabbay et al. [16] show that StyleGAN generator can be used as general purpose image prior. Shen et al. [47] show the opportunity to manipulate appearance of generated person, including age, gender, eyeglasses, and pose, for both PGGAN [29] and StyleGAN. The authors of StyleGAN2 [31] propose to search embeddings in $\mathcal{W}$ instead of $\mathcal{W}+$ to check if the picture was generated by StyleGAN2.

Paired Image-to-image translation Pix2pix [26] is one of the first conditional generative models applied for image-to-image translation. It learns mapping from input to output images. Chen and Koltun [9] propose the first model which can synthesize 2048x1024 images. It is followed by pix2pixHD [55] and SPADE [42]. In SPADE generator, each normalization layer uses the segmentation mask to modulate the layer activations. So its usage is limited to the translation from segmentation maps. There are numerous follow-up works based on pix2pixHD architecture, including those working with video [7, 53, 54].

Unpaired Image-to-image translation The idea of applying cycle consistency to train on unpaired data is first introduced in CycleGAN [61]. The methods of unpaired image-to-image translation can be either single mode GANs [61, 59, 36, 11] or multimodal GANs [62, 24, 33, 34, 37, 12]. FUNIT [37] supports multi-domain image translation using a few reference images from a target domain. StarGAN v2 [12] provide both latent-guided and reference-guided synthesis. All of the above-mentioned methods operate at resolution of at most 256x256 when applied to human faces.

Gender swap is one of well-known tasks of unsupervised image-to-image translation [11, 12, 38].

Face aging/rejuvenation is a special task which gets a lot of attention [60, 50, 19]. Formulation of the problem can vary. The simplest version of this task is making faces look older or younger [11]. More difficult task is to produce faces matching particular age intervals [35, 56, 58, 38]. S^2GAN [19] proposes continuous changing of age using weight interpolation between transforms which correspond to two closest age groups.

Training on synthetic data Synthetic datasets are widely used to extend datasets for some analysis tasks (e.g. classification). In many cases, simple graphical engine can be used to generate synthetic data. To perform well on real world images, this data need to overcome both appearance gap [22, 15, 51, 52, 49] and content gap [28, 46].

Ravuri et al. [44] study the quality of a classificator trained on synthetic data generated by BigGAN and show [45] that BigGAN does not capture the ImageNet [14] data distributions and is only partly successful for data augmentation. Shrivastava et al. [49] reduce the quality drop of this approach by revising train setup. Chen et al. [10] make paired dataset with image editing applications to train image2image network.

Synthetic data is what underlies knowledge distillation, a technique that allows to train "student" network using data generated by "teacher" network [21, 4]. Usage of this additional source of data can be used to improve measures [57] or to reduce size of target model [39]. Aguinaldo et al. [3] show that knowledge distillation is successfully applicable for generative models.

3 Method overview

3.1 Data collection

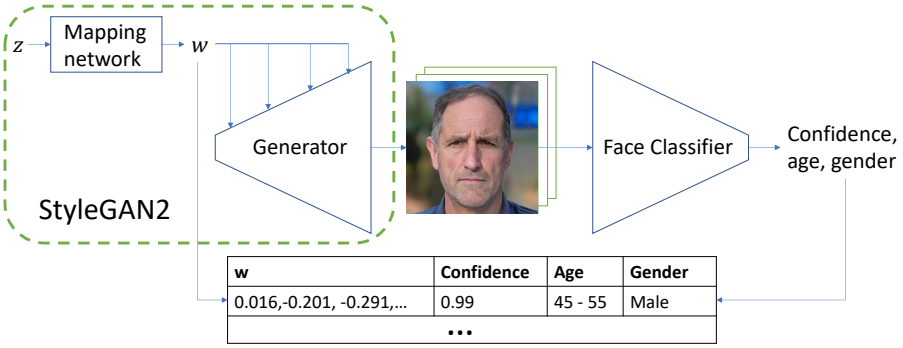


Fig. 2: Method of finding correspondence between latent codes and facial attributes

All of the images used in our datasets are generated using the official implementation of StyleGAN2[3]. In addition to that we only use the config-f version checkpoint pretrained by the authors of StyleGAN2 on FFHQ dataset. All the manipulations are performed with the disentangled image codes w.

[3] https://github.com/NVlabs/stylegan2

We use the most straightforward way of generating datasets for style mixing and face morphing. Style mixing is described in [30] as a regularization technique and requires using two intermediate latent codes w_1 and w_2 at different scales. Face morphing corresponds to linear interpolation of intermediate latent codes w. We generate 50 000 samples for each task. Each sample consists of two source images and a target image. Each source image is obtained by randomly sampling z from normal distribution, mapping it to intermediate latent code w, and generating image $g(w)$ with StyleGAN2. We produce target image by performing corresponding operation on the latent codes and feeding the result to StyleGAN2.

Face attributes, such as gender or age, are not explicitly encoded in StyleGAN2 latent space or intermediate space. To overcome this limitation we use a separate pretrained face classification network. Its outputs include confidence of face detection, age bin and gender. The network is proprietary, therefore we release the final version of our gender and age datasets in order to maintain full reproducibility of this work[4].

We create gender and age datasets in four major steps. First, we generate an intermediate dataset, mapping latent vectors to target attributes as illustrated in Fig. 2. Second, we find the direction in latent space associated with the attribute. Third, we generate raw dataset, using above-mentioned vector as briefly described in Fig. 3. Finally, we filter the images to get the final dataset. The method is described below in more detail.

1. Generate random latent vectors $z_1 \dots z_n$, map them to intermediate latent codes $w_1 \dots w_n$, and generate corresponding image samples $g(w_i)$ with StyleGAN2.
2. Get attribute predictions from pretrained neural network f, $c(w_i) = f(g(w_i))$.
3. Filter out images where faces were detected with low confidence[5]. Then select only images with high classification certainty.
4. Find the center of every class $C_k = \frac{1}{n_{c=k}} \sum_{c(w_i)=k} w_i$ and the transition vectors from one class to another $\Delta_{c_i,c_j} = C_j - C_i$
5. Generate random samples z_i and pass them through mapping network. For gender swap task, create a set of five images $g(w-\Delta), g(w-\Delta/2), g(w), g(w+\Delta/2), g(w+\Delta)$ For aging/rejuvenation first predict faces' attributes $c(w_i)$, then use corresponding vectors $\Delta_{c(w_i)}$ to generate faces that should be two bins older/younger.
6. Get predictions for every image in the raw dataset. Filter out by confidence.
7. From every set of images, select a pair based on classification results. Each image must belong to the corresponding class with high certainty.

As soon as we have aligned data, a paired image-to-image translation network can be trained.

[4] https://github.com/EvgenyKashin/stylegan2-distillation

[5] This helps to reduce generation artifacts in the dataset, while maintaining high variability as opposed to lowering truncation-psi parameter.

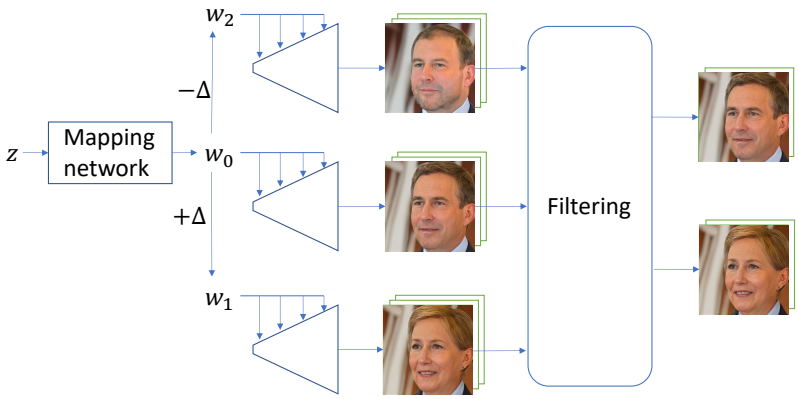


Fig. 3: Dataset generation. We first sample random vectors z from normal distribution. Then for each z we generate a set of images along the vector Δ corresponding to a facial attribute. Then for each set of images we select the best pair based on classification results

3.2 Training process

In this work, we focus on illustrating the general approach rather than solving every task as best as possible. As a result, we choose to train pix2pixHD[6] [55] as a unified framework for image-to-image translation instead of selecting a custom model for every type of task.

It is known that pix2pixHD has blob artifacts[7] and also tends to repeat patterns [42]. The problem with repeated patterns is solved in [30, 42]. Light blobs is a problem which is solved in StyleGAN2. We suppose that similar treatment also in use for pix2pixHD.

Fortunately, even vanilla pix2pixHD trained on our datasets produces sufficiently good results with little or no artifacts. Thus, we leave improving or replacing pix2pixHD for future work. We make most part of our experiments and comparison in 512x512 resolution, but also try 1024x1024 for gender swap.

Style mixing and face averaging tasks require two input images to be fed to the network at the same time. It is done by setting number of input channels to 6 and concatenating the inputs along channel axis.

4 Experiments

Although StyleGAN2 can be trained on data of different nature, we concentrate our efforts only on face data. We show application of our method to several

[6] https://github.com/NVIDIA/pix2pixHD
[7] https://github.com/NVIDIA/pix2pixHD/issues/46

tasks: gender swap, aging/rejuvenation and style mixing and face morphing. In all our experiments we collect data from StyleGAN2, trained on FFHQ dataset [30].

4.1 Evaluation protocol

Only the task of gender transform (two directions) is used for evaluation. We use Frechét inception distance (FID) [20] for quantitative comparison of methods, as well as human evaluation.

For each feed-forward baseline we calculate FID between 50 000 real images from FFHQ datasets and 20 000 generated images, using 20 000 images from FFHQ as source images. For each source image we apply transformation to the other gender, assuming source gender is determined by our classification model. Before calculating FID measure all images are resized to 256x256 size for fair comparison.

Also human evaluation is used for more accurate comparison with optimization based methods. Our study consists of two surveys:

1. **Quality.** Task for female to male translation (male to female one is similar): "For the same image on the left, there are two different options on the right. Choose the best face, which is: turned into a male (most important), similar to the original person, the position of the face and emotions are preserved, the original items in the photo are preserved."
2. **Realism.** In this task, sources are different and not shown. "Choose the image, which is: more realistic (the most important), better in quality, with fewer artifacts."

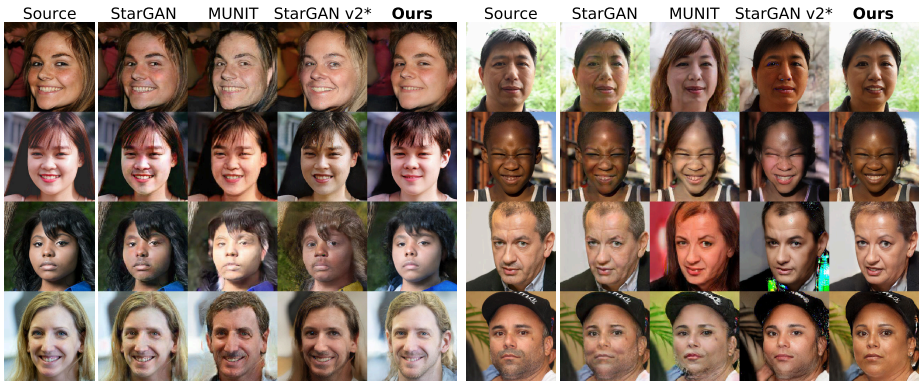
All images were resized to 512x512 size in this comparison. The first task should show which method is the best at performing transformation, the second – which looks the most real regardless of the source image. We use side-by-side experiments for both tasks where one side is our method and the other side is one of optimization based baselines. Answer choices are shuffled. For each comparison of our method with a baseline, we generate 1000 questions and each question is answered by 10 different people. For answers aggregation we use Dawid-Skene method [13] and filter out the examples with confidence level less than 95% (it is approximately 4% of all questions).

4.2 Distillation of image-to-image translation

Gender swap We generate a paired dataset for male and female faces according to the method described above and than train a separate pix2pixHD model for each gender translation.

We compete with both unpaired image-to-image methods and different StyleGAN embedders with latent code optimization. We choose StarGAN[8] [11], MU-

[8] https://github.com/yunjey/stargan



(a) Female to male (b) Male to female

Fig. 4: Gender transformation: comparison with image-to-image translation approaches. MUNIT and StarGAN v2* are multimodal so we show one random realization there

NIT[9] [25] and StarGAN v2*[10] [12] for a competition with unpaired methods. We train all these methods on FFHQ classified into males and females.

Fig. 4 shows qualitative comparison between our approach and unpaired image-to-image ones. It demonstrates that distilled transformation have significantly better visual quality and more stable results. Quantitative comparison in Table 1a confirms our observations. We also checked that our model is perform well on other datasets without retraining. Table 1b shows comparison of gender swap of CelebA-HQ images with models trained on CelebA. Our model wins despite it has no CelebA samples during training. The results indicate that the method can potentially be applied to real world images without retraining.

StyleGAN2 provides an official projection method. This method operates in $\mathcal{W}$, which only allows to find faces generated by this model, but not real world images. So, we also build a similar method for $\mathcal{W}+$ for comparison. It optimizes separate w for each layer of the generator, which helps to better reconstruct a given image. After finding w we can add transformation vector described above and generate a transformed image.

Also we add projection methods made by Dmitry Nikitko (Puzer) [41] and Peter Baylies (pbaylies) [5] for finding latent code to comparison, even though they are based on the first version of StyleGAN. These encoders are the most known implementations, they use custom perceptual losses for better perception. StyleGAN encoder by Peter Baylies is mode advanced one. In addition to more precisely selected loss functions, it uses background masking and forward pass approximation of optimization starting point.

[9] https://github.com/NVlabs/MUNIT

[10] https://github.com/taki0112/StarGAN_v2-Tensorflow (unofficial implementation, so its results may differ from the official one)

(a) Evaluate on FFHQ

Method	FID
StarGAN [11]	29.7
MUNIT [24]	40.2
StarGANv2* [12]	25.6
Ours	**14.7**
Real images	3.3

(b) Evaluate on Celeba-HQ

Method	FID
StarGANv2 [12][11]	27.3
Ours	**21.3**

Table 1: Quantitative comparison with unpaired methods. Unpaired methods trained on the same datasets we evaluate them, although ours trained on FFHQ in both cases. Table 1b shows that our method is robust regarding dataset.

Since unpaired methods show significantly worse quality, we put more effort into comparisons between different methods of searching embedding through optimization. We avoid using methods that utilize FID because all of them are based on the same StyleGAN model. Also, FID cannot measure "quality of transformation" because it does not check keeping of personality. So we decide to make user study our main measure for all StyleGAN-based methods. Fig. 5 shows qualitative comparison of all the methods. It is visible that our method performs better in terms of transformation quality. And only StyleGAN Encoder [5] outperforms our method in realism. However this method generates background unconditionally.

Table 2: User study of StyleGAN-based approaches. Winrate "method vs ours". We measure user study for all StyleGAN-based approaches because we consider human evaluation more reliable measure for perception

Method	Quality	Realism
StyleGAN Encoder (Nikitko)	18%	14%
StyleGAN Encoder (Baylies)	30%	68%
StyleGAN2 projection ($\mathcal{W}$)	22%	22%
StyleGAN2 projection ($\mathcal{W}+$)	11%	14%
Real images	-	85%

We find that pix2pixHD keeps more details on transformed images than all the encoders. We suppose that this is achieved due to the ability of pix2pixHD to pass part of the unchanged content through the network. Pix2pixHD solves an easier task compared to encoders which are forced to encode all the information about the image in one vector.

[11] Official model and weights.

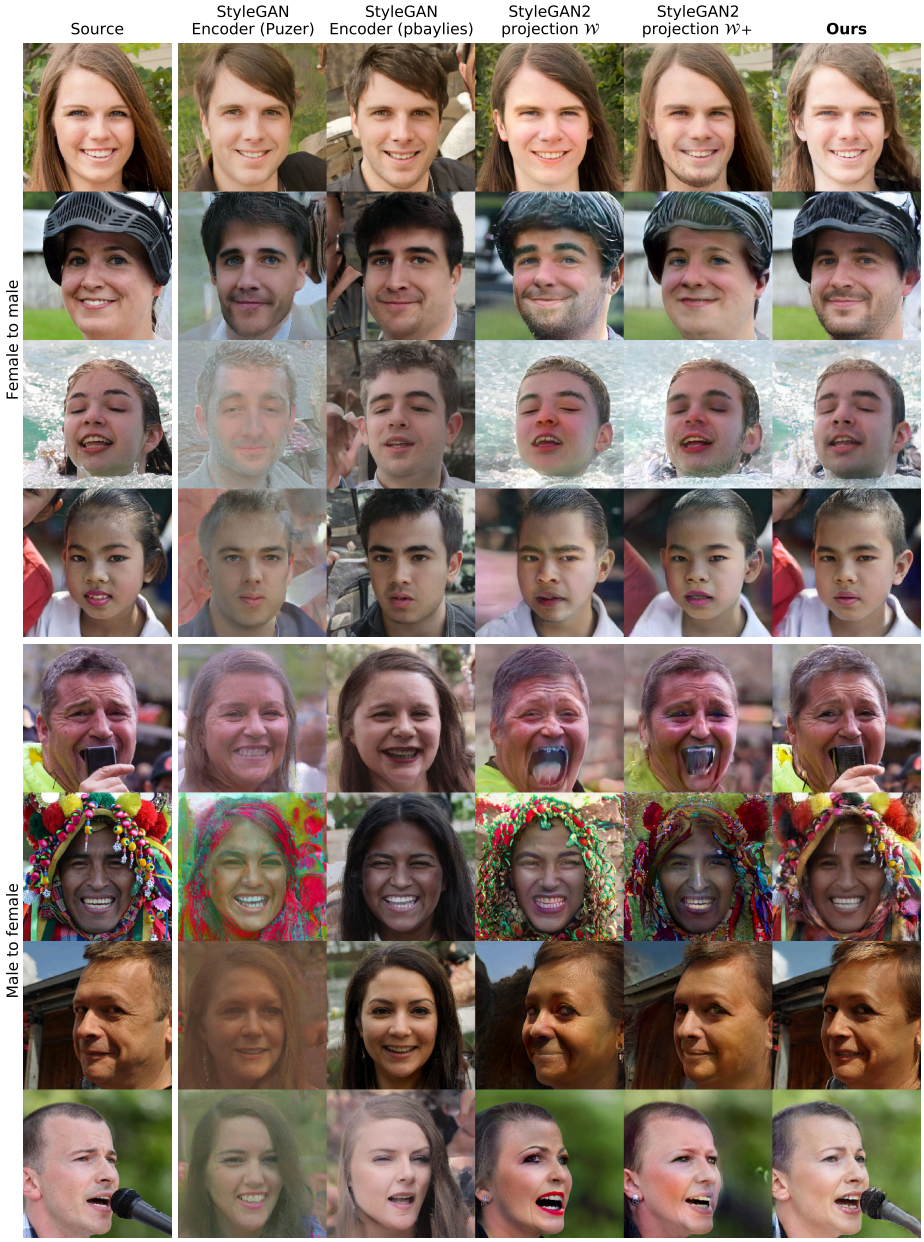


Fig. 5: Gender transformation: comparison with StyleGAN2 latent code optimization methods. Input samples are real images from FFHQ. Notice that unusual objects are lost with optimization but kept with image-to-image translation

Fig. 4 and 5 also show drawbacks of our approach. Vector of "gender" is not perfectly disentangled due to some bias in attribute distribution of FFHQ and, consequently, latent space correlation of StyleGAN[47]. For example, it can be seen that translation into female faces can also add smile.

We also encounter problems of pix2pixHD architecture: repeated patterns, light blobs and difficulties with finetuning 1024x1024 resolution. We show an uncurated list of generated images in supplementary materials.

Aging/rejuvenation To show that our approach can be applied for another image-to-image transform task, we also carry out similar experiment with face age manipulation. First, we estimate age for all generated images, then group them into several bins. After that, for each bin we find vectors of "+2 bins" and "-2 bins". Using these vectors, we generate united paired dataset. Each pair contains younger and older versions of the same face. Finally, we train two pix2pixHD networks, one for each of two directions. Examples of the application

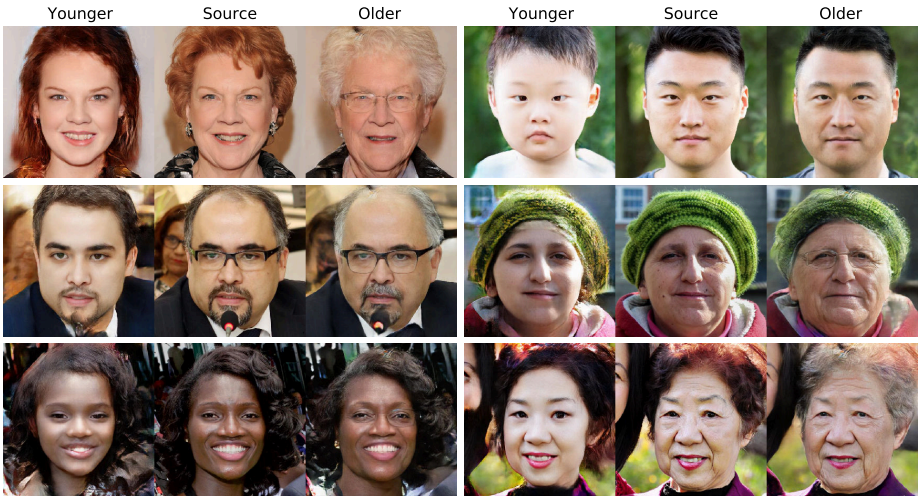


Fig. 6: Aging/rejuvenation. Source images are sampled from FFHQ

of this approach are presented in Fig. 6.

4.3 Distillation of style mixing

Style mixing and face morphing There are 18 AdaIN inputs in StyleGAN2 architecture. These AdaINs work with different spatial resolutions, and changing different input will change details of different scale. The authors divide them into three groups: coarse styles (for 4^2 – 8^2 spatial resolutions), middle styles (16^2 –

Fig. 7: Style mixing with pix2pixHD. (a), (b), (c) show results of distillated crossover of two latent codes in $\mathcal{W}+$, (d) shows result of average latent code transformaiton. Source images are sampled from FFHQ

32^2) and fine styles (64^2 – 1024^2). The opportunity to change coarse, middle or fine details is a unique feature of StyleGAN architectures.

We collect datasets of triplets (two source images and their mixture) and train our models for each transformation. We concatenate two images into 6 channels to feed our pix2pixHD model. Fig. 7(a,b,c) show the results of style mixing.

Another simple linear operation is to average two latent codes. It corresponds to morphing operation on images. We collect another dataset with triplet latent codes: two random codes and an average one. The examples of face morphing are shown in Fig. 7(d).

5 Conclusions

In this paper, we unite unconditional image generation and paired image-to-image GANs to distill a particular image manipulation in latent code of StyleGAN2 into single image-to-image translation. The resulting technique shows both fast inference and impressive quality. It outperforms existing unpaired image-to-image models in FID score and StyleGAN Encoder approaches both in user study and time of inference on gender swap task. We show that the approach is also applicable for other image manipulations, such as aging/rejuvenation and style transfer.

Our framework has several limitations. StyleGAN2 latent space is not perfectly disentangled, so the transformations made by our network are not perfectly pure. Despite the latent space is not disentangled enough to make pure transformations, impurities are not so severe.

We use only pix2pixHD network although different architectures fit better for different tasks. Besides, we distil every transformation to a separate model, although some universal model could be trained. This opportunity should be investigated in future studies.

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan++: How to edit the embedded images? arXiv preprint arXiv:1911.11544 (2019)
2. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4432–4441 (2019)
3. Aguinaldo, A., Chiang, P.Y., Gain, A., Patil, A., Pearson, K., Feizi, S.: Compressing gans using knowledge distillation. arXiv preprint arXiv:1902.00159 (2019)
4. Ba, J., Caruana, R.: Do deep nets really need to be deep? In: Advances in neural information processing systems. pp. 2654–2662 (2014)
5. Baylies, P.: Stylegan encoder - converts real images to latent space. https://github.com/pbaylies/stylegan-encoder (2019)
6. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)

7. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5933–5942 (2019)
8. Chen, B.C., Chen, C.S., Hsu, W.H.: Cross-age reference coding for age-invariant face recognition and retrieval. In: European conference on computer vision. pp. 768–783. Springer (2014)
9. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: Proceedings of the IEEE international conference on computer vision. pp. 1511–1520 (2017)
10. Chen, Y.C., Shen, X., Jia, J.: Makeup-go: Blind reversion of portrait edit. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4501–4509 (2017)
11. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
12. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. arXiv preprint arXiv:1912.01865 (2019)
13. Dawid, A.P., Skene, A.M.: Maximum likelihood estimation of observer error-rates using the em algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics) **28**(1), 20–28 (1979)
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009)
15. French, G., Mackiewicz, M., Fisher, M.: Self-ensembling for visual domain adaptation. arXiv preprint arXiv:1706.05208 (2017)
16. Gabbay, A., Hoshen, Y.: Style generator inversion for image enhancement and animation. arXiv preprint arXiv:1906.11880 (2019)
17. Goetschalckx, L., Andonian, A., Oliva, A., Isola, P.: Ganalyze: Toward visual definitions of cognitive image properties. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
18. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
19. He, Z., Kan, M., Shan, S., Chen, X.: S2gan: Share aging factors across ages and share aging trends among individuals. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9440–9449 (2019)
20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems. pp. 6626–6637 (2017)
21. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
22. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. arXiv preprint arXiv:1711.03213 (2017)
23. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1501–1510 (2017)
24. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 172–189 (2018)
25. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV (2018)

26. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
27. Jahanian, A., Chai, L., Isola, P.: On the "steerability" of generative adversarial networks. arXiv preprint arXiv:1907.07171 (2019)
28. Kar, A., Prakash, A., Liu, M.Y., Cameracci, E., Yuan, J., Rusiniak, M., Acuna, D., Torralba, A., Fidler, S.: Meta-sim: Learning to generate synthetic datasets. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4551–4560 (2019)
29. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
30. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
31. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. arXiv preprint arXiv:1912.04958 (2019)
32. Khosla, A., Raju, A.S., Torralba, A., Oliva, A.: Understanding and predicting image memorability at a large scale. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2390–2398 (2015)
33. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M.K., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: European Conference on Computer Vision (2018)
34. Lee, H.Y., Tseng, H.Y., Mao, Q., Huang, J.B., Lu, Y.D., Singh, M.K., Yang, M.H.: Drit++: Diverse image-to-image translation viadisentangled representations. arXiv preprint arXiv:1905.01270 (2019)
35. Li, P., Hu, Y., Li, Q., He, R., Sun, Z.: Global and local consistent age generative adversarial networks. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 1073–1078. IEEE (2018)
36. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in neural information processing systems. pp. 700–708 (2017)
37. Liu, M.Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., Kautz, J.: Few-shot unsupervised image-to-image translation. arXiv preprint arXiv:1905.01723 (2019)
38. Liu, Y., Li, Q., Sun, Z.: Attribute-aware face aging with wavelet-based generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11877–11886 (2019)
39. Mirzadeh, S.I., Farajtabar, M., Li, A., Ghasemzadeh, H.: Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher. arXiv preprint arXiv:1902.03393 (2019)
40. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 51–59 (2017)
41. Nikitko, D.: Stylegan – encoder for official tensorflow implementation. https://github.com/Puzer/stylegan-encoder (2019)
42. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2337–2346 (2019)
43. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)

44. Ravuri, S., Vinyals, O.: Classification accuracy score for conditional generative models. In: Advances in Neural Information Processing Systems. pp. 12247–12258 (2019)
45. Ravuri, S., Vinyals, O.: Seeing is not necessarily believing: Limitations of biggans for data augmentation (2019)
46. Ruiz, N., Schulter, S., Chandraker, M.: Learning to simulate. arXiv preprint arXiv:1810.02513 (2018)
47. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9243–9252 (2020)
48. Shi, T., Yuan, Y., Fan, C., Zou, Z., Shi, Z., Liu, Y.: Face-to-parameter translation for game character auto-creation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 161–170 (2019)
49. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2107–2116 (2017)
50. Song, J., Zhang, J., Gao, L., Liu, X., Shen, H.T.: Dual conditional gans for face aging and rejuvenation. In: IJCAI. pp. 899–905 (2018)
51. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 23–30. IEEE (2017)
52. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7472–7481 (2018)
53. Wang, T.C., Liu, M.Y., Tao, A., Liu, G., Kautz, J., Catanzaro, B.: Few-shot video-to-video synthesis. In: Conference on Neural Information Processing Systems (NeurIPS) (2019)
54. Wang, T.C., Liu, M.Y., Zhu, J.Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. arXiv preprint arXiv:1808.06601 (2018)
55. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
56. Wang, Z., Tang, X., Luo, W., Gao, S.: Face aging with identity-preserved conditional generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7939–7947 (2018)
57. Xie, Q., Hovy, E., Luong, M.T., Le, Q.V.: Self-training with noisy student improves imagenet classification. arXiv preprint arXiv:1911.04252 (2019)
58. Yang, H., Huang, D., Wang, Y., Jain, A.K.: Learning continuous face age progression: A pyramid of gans. IEEE transactions on pattern analysis and machine intelligence (2019)
59. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
60. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5810–5818 (2017)

61. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
62. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: Advances in Neural Information Processing Systems (2017)