Vision Transformers for Neural Human Rendering

Arda Arslan
ETH Zürich
aarslan@student.ethz.ch

Tancrède Guillou
ETH Zürich
tguillou@student.ethz.ch

Dhavan Atul Shah
University of Zürich
dhavanatul.shah@uzh.ch

Raphael Winkler
ETH Zürich
winklerr@student.ethz.ch

Abstract

In this paper, we introduce new methods for generating realistic human faces with Generative Adversarial Networks. Recently proposed neural network architectures have shown to be more successful than convolutional neural networks in several computer vision tasks. In this work, we address the challenge of synthesising novel faces from a black and white representation of basic face edges. Our method is built on top of Pix2Pix, a generative adversarial network working on edge images. We augment the original pipeline by replacing the convolutional discriminator with new architectures, namely Vision Transformer and MLP-Mixer. We create a flexible method that can use any of the aforementioned architectures as its discriminator, and use it to explore the qualities of each of these ones. We create a single-subject dataset for reconstruction task and use FaceForensics for novel synthesis experiments. Particularly, we show that using the new architectures achieves remarkable improvements over prior version on both datasets.

1. Introduction

Over the previous years, convolution has been the essential operation in computer vision architectures, and machine learning researchers have been attempting to replace them with other operations. Recently, Vision Transformer [6] and MLP-Mixer [33] are shown to be more successful than convolutional neural networks in computer vision tasks such as image segmentation, image classification and depth prediction [18].

Neural rendering is a computer vision task which makes use of ideas from computer graphics and machine learning to generate photo-realistic images or videos [32]. Many state-of-the-art neural rendering pipelines still include CNNs [25], and we believe the recently proposed computer vision architectures might replace CNNs in neural rendering pipelines as well.

In this work, we work on two types of neural human rendering tasks: the first one is novel human face synthesis from a given face edge map, which has been previously investigated in Vid2Vid [37] as part of a neural video to video translation work. Due to the limitations in computing power and time, we instead work on image to image translation, however use the same dataset which is used in Vid2Vid. Novel human face synthesis can be used in computer games and animations. The second one is generating a realistic face image from a given face edge map. For this task, we use a video from YouTube which is a recording of a single human face whose length is more than 2 hours. This task is also quite important because for face verification networks, we can increase number of training images using the resulting model of novel human face synthesis task. We can also frontalize face images during inference time in order to increase the accuracy of the face verification network [4].

As the baseline architecture, we use Pix2Pix [14] which is a conditional Generative Adversarial Network (GAN) [11] architecture, and whose main goal is to train a generator network which can map sketches to photo-realistic images.

In our work, we substitute the discriminator of the Pix2Pix architecture, which is a CNN, with a vision transformer or an MLP-Mixer. Our expectation is that this substitution will strengthen the discriminator network, and motivate the network to have a stronger generator. By this improvement, we aim to generate more realistic images.

Our contributions are as following:

- To the best of our knowledge, this work is the first time Vision Transformer and MLP-Mixer architectures are used for face reconstruction and novel face synthesis tasks.

- We provide three generator networks which can generate novel face images given a face edge map. These generator networks are trained together with a discriminator network which is whether a CNN, a Vision Transformer or an

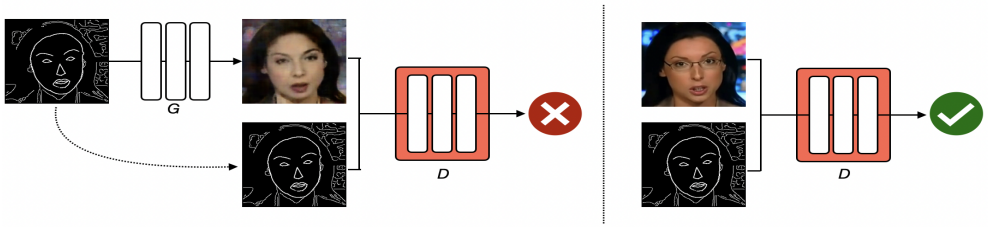


Figure 1. **Exploring new discriminator architectures in the Pix2Pix pipeline.** A typical GAN consists of a generator *G* and a discriminator *D*. The latter determines if a given image is true or if it is generated, while the generator tries to fool the discriminator. In this work, we are going to try two new types of discriminators : a (1) **Vision Transformer** and a (2) **MLP-Mixer**. The code is available at https://github.com/ardarslan/neural-human-rendering.

MLP-Mixer.

- We do ablation studies for these three architectures where we compare the generated image qualities in terms of FID metric.

- We provide three generator networks which can generate face image of a certain person given a face edge map from different viewpoints. The training details of these generator networks are the same with the first bullet point.

- We do ablation studies for these three architectures where we compare the generated image qualities in terms of SSIM and LPIPS metrics.

2. Related Work

2.1. Neural Human Rendering

There has been a vast amount of work done on neural human rendering recently. [22], [23] work on generating realistic and controllable animations of human bodies from a motion sequence and template mesh of the human body. Moreover, [26] and [20] work on generating free-viewpoint video of a human performer using only a few camera views.

There are also many works which focus on generating novel human face images. By the time we write this paper, one of the state-of-the-art methods for novel face image synthesis on FFHQ 256 x 256 dataset [17] is [29]. This work uses strong neural network priors and a progressive growing method to train the most recent StyleGAN3 [16] generator on ImageNet [5] dataset.

Novel view synthesis is generating a 2D image of an object from an unobserved camera view. By the time we write this paper, one of the state-of-the-art methods on novel view synthesis for face images is [35]. In this work, the authors propose using a GAN to mix an image which is generated by NeRF [24] with an image which is rendered by a 3D Morphable Model.

2.2. Generative Adversarial Networks (GAN)

GANs are known to perform quite well on many computer vision tasks. Pix2Pix [14] is a GAN consisting of a U-Net generator and a CNN discriminator. It is well explored and adapted to countless tasks with good results. A notable alteration to Pix2Pix is Vid2Vid [38] which extends the problem of image to image translation to videos and introduces temporal consistency. This dramatically increases resource usage and training time. For our experiments we used Pix2Pix due to resource constraints, and leave the video to video task for future work.

2.3. Vision Transformer

In the recent years, the transformer architecture has shown great success in natural language processing tasks [36]. By taking inspiration from this achievement, in [6], the authors propose a transformer architecture which is proven to be successful in visual domain as well. This architecture does not use convolutions, but rather applies a transformer directly to sequences of image patch encodings. Vision transformers have been employed successfully in many computer vision tasks. They found initial success in image classification where many hybrid architectures were introduced to improve classification accuracy ([10], [12], [3]). Later on they were also used to improve the common CNN-based GAN architectures. [7] uses a Vision Generator and CNN discriminator, and ([15], [21]) use VIT's in both the Generator and Discriminator, presenting a GAN without any convolutions.

2.4. MLP-Mixer

MLP-Mixer is a recently proposed neural architecture used for computer vision tasks. It does not use convolutions or attention mechanism, it only uses multilayer perceptrons. The architecture uses two types of layers, one mixes the spatial information, the other one mixes the per-location features in images. In [33], authors claim that MLP-Mixer achieves competitive scores on image classification tasks when trained on sufficiently large datasets.

MLP architectures found similar success to the VIT architectures discussed before. [34] introduced an all-MLP architecture for image classification. [2] managed to create a

fast converging classification architecture using both multi-scale VIT's and MLP layers.

3. Method

3.1. Baseline Method: Original Pix2Pix

Pix2Pix architecture consists of a U-Net generator and a CNN discriminator. The architecture is illustrated in Figure 1.

For all experiments, we used the original U-Net generator of Pix2Pix.

For face reconstruction task, the discriminator of our baseline has three downsampling operations with 64, 128, 256 filters. Then there are two convolutional layers with 512 filters and 1 filter. This is the original Pix2Pix architecture.

For novel face synthesis task, the discriminator of our baseline has three downsampling operations with 32, 32 and 64 filters. Then there are two convolutional layers with 64 filters and 1 filter. We needed to make this change in the architecture because discriminator was too powerful for this task when we had more filters and this caused discriminator to win at the early iterations.

3.2. Vision Transformer (VIT) as Discriminator

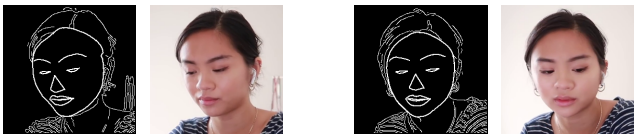
We used a standard multi-head Vision Transformer available from the official TensorFlow [1] website to replace the CNN discriminator in Pix2Pix. We configured it to use 3 heads, a projection dimension of 32 channels, and 3 transformer layers. This is much smaller than the default settings for most Vision Discriminators, but it is more in line with the size of the replaced CNN discriminator. Making the VIT stronger results in a disproportionately strong discriminator which hinders convergence of the GAN overall. While the standard Pix2Pix network was hard to train in our experiments, we observed that the GAN with a VIT discriminator converges much more reliably. A balanced training occurs much more frequently than with the CNN discriminator.

3.3. MLP-Mixer as Discriminator

Our MLP-Mixer pipeline is composed of multiple Mixer Blocks (two or four). Each Mixer Block contains exactly two MLP Blocks which consist in two fully-connected layers. The MLP Blocks are applied independently to each row of the input data tensor. A Mixer Block takes as input a sequence of S non-overlapping image patches, each projected to a hidden dimension C, which yields a real-valued input table $X \in \mathbb{R}^{S \times C}$. Using an input image of size $H \times W$, and patches of resolution $R \times R$, the number of patches we need simply becomes

$$\mathcal{S} = HW/P^2$$

The two MLP Blocks inside every Mixer Block are :



Input Ground truth Input Ground truth

Figure 2. **Face Reconstruction dataset.** Two examples of $(input - ground\_truth)$ pairs for the face reconstruction dataset. The input is a black and white image showing edges of the face and some details. The ground truth is the true image of the subject.

- **Token Mixing Block**. This block acts along the columns of the input table $X \in \mathbb{R}^{S \times C}$ and performs a $\mathbb{R}^S \mapsto \mathbb{R}^S$ mapping.

- **Channel Mixing Block.** This block acts along the rows of the input table $X \in \mathbb{R}^{S \times C}$ and performs a $\mathbb{R}^C \mapsto \mathbb{R}^C$ mapping.

Since the embedding dimensions of the MLP blocks are set independently of the number of input patches, the computational complexity of the network is linear in the number of input patches. This makes the Mixer model asymptotically better than a quadratic Vision transformer, and compares to the Convolutional Neural Network's complexity.

This block duality make the classifier adapt better to different data and is more accurate on edge patches, where the discriminator might struggle during GAN training.

4. Experiments

4.1. Datasets

4.1.1 Face reconstruction dataset

We first evaluate our method on a face reconstruction dataset created for the purpose of this project. We selected a YouTube video of a single subject [8], facing the camera with white unicoloured background. We divide the frames linearly into training, validation and test set. Finally we use Vid2Vid preprocessing method to create $(input - ground\_truth)$ pairs as shown in Figure 2. The goal of our model is to recreate the subject face from unseen edge image input.

4.1.2 Novel face synthesis dataset

We use FaceForensics [28] dataset for novel face synthesis task. This dataset consists of 1000 videos of news reporters. We use the train/validation/test split which is the default split in the download script. We extract frames from each video and create face edge maps using the preprocessing method of Vid2Vid.

Figure 3. **Qualitative Comparison of novel pose face reconstruction.** Comparison of images generated by three pipelines with different discriminators on the face reconstruction dataset.

4.2. Evaluation Metrics

4.2.1 LPIPS

The Learned Perceptual Image Patch Similarity (LPIPS) [41] is for measuring the perceptual similarity between two input images. It is the similarity between activations for these two images from a deep layer of a neural network [9]. We use AlexNet [19] and VGG [30] as the neural network for our evaluations.

4.2.2 SSIM

Structural Similarity Index Measure (SSIM) [39] is a weighted combination of three comparison measurements between two input images: contrast, luminance and structure. SSIM values are in the range of 0 and 1. When SSIM is 1, this means the two images we are comparing are exactly the same with each other. [40]

4.2.3 Fréchet Inception Distance (FID)

The *Fréchet Inception Distance* [13] or FID is a method for comparing the statistics of two distributions by computing the distance between them. In GANs, the FID method is used for computing how much the distribution of the Generator looks like the distribution of the Discriminator. By consequence, it is a metric of GAN performance - the lower the FID, the better the GAN. This metric is known to correlate well with the visual quality of generated samples. It summarizes how similar the two groups are in terms of statistics on computer vision features of the raw images calculated using the Inception v3 model [31]. Since we are not doing reconstruction anymore we are not comparing similarity between ground truth and generated image. That's why we don't use LPIPS and SSIM, but rather FID which is a good metric for generated images.

4.3. Results

4.3.1 Results on the face reconstruction dataset

The quantitative results in Table 1 gather the setups that yielded the best LPIPS (VGG) value for each architecture. These results show that both Vision Transformer [6] and MLP Mixer [33] outperform original Pix2Pix [14] in all metrics. In particular, MLP Mixer achieves remarkable results with significantly better results than both other concurrent versions. These results are corroborated by the reconstruction images displayed in Figure 3. All three methods produce reasonable results, however the original Pix2Pix produces some artifacts in the image. Moreover, we can see it struggles to properly reconstruct details of the head when the orientation is too complex, as we can see on the second image. On the other hand, both Vision Transformer and MLP Mixer achieve very realistic results and close to no artifacts. We also note that MLP Mixer results are slightly more accurate in the details, and seem to deal better with lighting and shadows.

4.3.2 Results on the FaceForensics dataset

Table 4 shows the *Fréchet Inception Distance* [13] results on the FaceForensics dataset for each architecture. Once again, MLP Mixer achieves the best performance. However, Vision Transformer achieves poorly on this metric. Although the images generated by Vision Transformer

	Pix2Pix	VIT	MLP-Mixer
SSIM ↑	0.7754	0.7952	**0.8157**
LPIPS*(Alex) ↓	18.65	17.41	**16.68**
LPIPS*(VGG) ↓	31.91	29.83	**28.17**

Table 1. **Benchmark comparison on the face reconstruction task.** ($LPIPS^* = LPIPS * 10^2$)

Disc. lr	FID
0.000075	150.13
0.000100	**127.58**
0.000125	135.44
0.000150	133.37

Table 2. **FID scores of original Pix2Pix models trained with different discriminator learning rates on novel face synthesis task.**

were generally more realistic than the images generated by Pix2Pix, the high FID score can be explained by the fact that the images it generates are not diverse enough (it is one of the key aspects of the FID metric). Figure 4 shows a comparisons between the architectures for several different subjects. We take the subjects that yield the best results for original Pix2Pix, since the other, mostly bad, images are not very relevant for comparisons. The results show that, while original Pix2Pix is able to generate good images, the images can still produce little artifacts on some specific details of the head. The results on ViT and MLP Mixer are usually smoother, with MLP Mixer achieving a nice blend between perceptual similarity and details.
It is worth noting that we explicitly chose to display the best original Pix2Pix results to see the differences in how these discriminators operate. In general, Vision Transformer and especially MLP Mixer, achieve much better results on average, and adapt better to complex face orientation.

4.4. Ablation studies

4.4.1 Ablation Studies on Original Pix2Pix

Face reconstruction task. In all our experiments, we used a generator learning rate 0.0001 and batch size 32. We varied the discriminator learning rate, and observed that we never had a converged training. We provide some of the results in Table 2.

Face synthesis task. In all our experiments, we used a generator learning rate 0.0002 and batch size 32. We varied the discriminator learning rate, and observed that again, we never had a converged training. We provide some of the results in Table 5.

	Recon.	Synthesis	
Disc. lr	FID	SSIM	LPIPS (VGG)
0.00005	192.6	0.7880	0.3002
0.00020	**184.25**	**0.7952**	**0.2983**
0.00050	188.5	0.7942	0.2991
0.00100	199.1	0.7801	0.3044

Table 3. **Comparison of metrics for the Vision Transformer models.** We tested various models trained with different discriminator learning rates on the reconstruction task (left, FID metric) and the novel face synthesis task (right, SSIM and LPIPS metrics). All ViT experiments used a batch size of 4, and use a similar amount of GPU memory than the CNN version with batch size of 32.

	Pix2Pix	VIT	MLP-Mixer
FID ↓	127.58	184.25	**126.23**

Table 4. **Benchmark comparison on the novel face synthesis task.**

Disc. lr	SSIM	LPIPS (AlexNet)	LPIPS (VGG)
0.00015	**0.7754**	**0.1865**	**0.3191**
0.00020	0.7716	0.1837	0.3221
0.00025	0.7698	0.1839	0.3234
0.00030	0.7682	0.1843	0.3249
0.00035	0.7676	0.1848	0.3265

Table 5. **Scores of Original Pix2Pix models trained with different discriminator learning rates on face reconstruction task.**

4.4.2 Ablation Studies on Vision Transformer

The original implementation of Vision Transformer uses eight attention layers. With this setup, the ViT discriminator was significantly and consistently too strong for the Pix2Pix convolutional generator. We thus created a reduced ViT discriminator using three attention layers only, in order to reach convergence to a meaningful optimum and achieve a stable GAN training.
Table 3 shows the quantitative results of the Vision Transformer discriminator pipeline. Once using the right amount of attention layers, the training is stable over all learning rates, and altering it has little impact on the performance. However, it is interesting noting that the performance is stable over the different experiments, with 0.0002 discriminator learning rate achieving best results on both the reconstruction and the synthesis tasks.
We show some qualitative results in Figure 4. The generated faces are realistic and consistent over face expressions and orientations. We can however distinguish some overall lack of detail and sharpness, which characterizes the difference with MLP Mixer results.

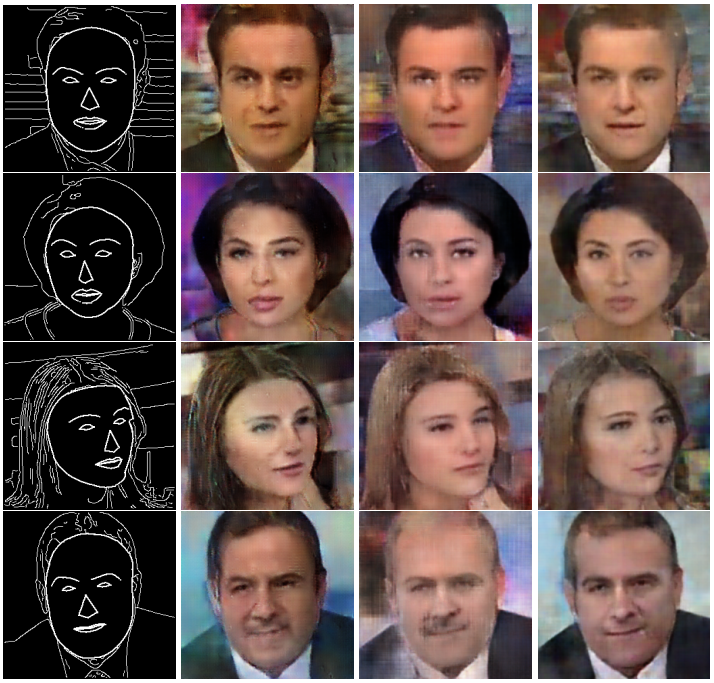


Figure 4. **Comparisons on Face Forensics dataset.** Four example images generated by different architectures

Emb. dims → ↓ *Num. blocks*	256	128
4	128.02	127.64
2	**126.23**	126.89

Table 6. **FID scores of MLP-Mixer models trained with different settings on novel face synthesis task.**

4.4.3 Ablation Studies on MLP-Mixer

For the MLP-Mixer experiments, we conduct ablation studies on the FaceForensics dataset. We explore the performances of our model trained with different numbers of embedding dimensions and Mixer blocks. Table 6 shows the quantitative results, which indicate that training on a lower number of Mixer blocks improves the synthesis of new faces. The discriminator is too powerful with four Mixer blocks, however it becomes too weak if we set both *emb_dims* and *num_blocks* to lower values.

These results also show that using an MLP-Mixer as a discriminator still outperforms original Pix2Pix in all settings.

5. Discussion

Limitations:

- We trained the original Pix2Pix architecture with varying number of convolutional layers and varying number of filters for each layer. We also varied both the discriminator and generator learning rates. We never observed a converged training in these experiments both for the novel face synthesis, and face reconstruction tasks.
- In one of our experiments, we tried to replace the generator of the original Pix2Pix architecture with a Vision Transformer, however due to GPU memory constraints, we were never able to train it.

Possible feature work:

- Using more advanced GAN training techniques to train the original Pix2Pix architecture for a more fair comparison.
- Fine-tuning the CLIP Vision encoder [27] discriminator to analyse and compare the results.
- Trying using all the aforementioned architectures as generators in the Pix2Pix [14] pipeline.
- Augmenting the pipeline for full body image synthesis.
- Using more powerful GPU setup to try Vision Transformer [6] at full scale.

Conclusion. We explored the use of several architectures as a discriminator for Pix2Pix face synthesis tasks. In particular, we show that Vision Transformer or MLP

Mixer [33] can produce a remarkable improvement on the results. We performed several ablation studies on the different architectures and settings to get a quantifiable understanding of the specificity of each architecture. We hope the results show the potential of such architectures and point toward expanding their use to other similar neural rendering pipelines.

6. Contributions of team members

- Arda: Implemented the initial version of the experimentation pipeline which includes the original Pix2Pix architecture, implemented CLIP encoder discriminator*, the preprocessing script for novel face synthesis dataset, SSIM and LPIPS evaluation metrics. Trained the original Pix2Pix architecture for face reconstruction and novel face synthesis tasks.

- Dhavan: Trained the original Pix2Pix for face reconstruction and novel face synthesis tasks.

- Raphael: Implemented and trained Vision Transformer for face recognition and novel face synthesis tasks.

- Tancrede: Implemented and trained MLP-Mixer for face recognition and novel face synthesis tasks. Implemented the preprocessing script for the face reconstruction task. Also implemented FID evaluation metric.

* We only experimented with pre-trained CLIP discriminator, and observed that it always predicted zero from the beginning of each training, therefore we did not include results of these experiments.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 3

[2] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *CoRR*, abs/2103.14899, 2021. 2

[3] Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *CoRR*, abs/2106.04803, 2021. 2

[4] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition, 2017. 1

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 1, 2, 4, 6

[7] Ricard Durall, Stanislav Frolov, Andreas Dengel, and Janis Keuper. Combining transformer generators with convolutional discriminators. *CoRR*, abs/2105.10189, 2021. 2

[8] elloitsangela. Real time study with me (no music): 2.25 hour pomodoro session with breaks (background noise), 2022. 3

[9] PyTorchLightning et al. Learned perceptual image patch similarity (lpips), 2022. 4

[10] Behnaz Gheflati and Hassan Rivaz. Vision transformer for classification of breast ultrasound images. *CoRR*, abs/2110.14731, 2021. 2

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 1

[12] Jianyuan Guo, Kai Han, Han Wu, Chang Xu, Yehui Tang, Chunjing Xu, and Yunhe Wang. CMT: convolutional neural networks meet vision transformers. *CoRR*, abs/2107.06263, 2021. 2

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *CV*, 2018. 4

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 1, 2, 4, 6

[15] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two transformers can make one strong GAN. *CoRR*, abs/2102.07074, 2021. 2

[16] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks, 2021. 2

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

[18] Salman H. Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *CoRR*, abs/2101.01169, 2021. 1

[19] Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks, 2014. 4

[20] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering, 2021. 2

[21] Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. Vitgan: Training gans with vision transformers. *CoRR*, abs/2107.04589, 2021. 2

[22] Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeongwoo Kim, Wenping Wang, and Christian Theobalt. Neural human video rendering by learning dynamic textures and rendering-to-video translation. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 05 2020. 2

[23] Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. Neural rendering and reenactment of human actor videos, 2018. 2

[24] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020. 2

[25] NVlabs. imaginaire, 2022. 1

[26] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans, 2020. 2

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 6

[28] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces, 2018. 3

[29] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets, 2022. 2

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. 4

[31] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CV*, 2015. 4

[32] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul P. Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, Tomas Simon, Christian Theobalt, Matthias Nießner, Jonathan T. Barron, Gordon Wetzstein, Michael Zollhöfer, and Vladislav Golyanik. Advances in neural rendering. *CoRR*, abs/2111.05849, 2021. 1

[33] Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. *CoRR*, abs/2105.01601, 2021. 1, 2, 4, 7

[34] Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. *CoRR*, abs/2105.01601, 2021. 2

[35] Satoshi Tsutsui, Weijia Mao, Sijing Lin, Yunyi Zhu, Murong Ma, and Mike Zheng Shou. Novel view synthesis for high-fidelity headshot scenes, 2022. 2

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 2

[37] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *CoRR*, abs/1808.06601, 2018. 1

[38] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[39] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 4

[40] Wikipedia. Structural similarity, 2022. 4

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 4