Zusammenfassung

Highly descriptive text-to-face generation to sythesize authentic faces (photofits for criminology purposes) via a GAN. Main project of the teaching event Computer Vision and Deep Learning: Visual Synthesisïn the summer of 2022 at LMU Munich. We designed a framework that currently provides two GAN-Models – cDCGAN and TediGAN – that are easy to train, evaluate and use to generate photofits. It is implemented in PyTorch and highly configurable to accomodate many test cases. Due to its architecture other models, datasets and metrics can easily be added. We are looking forward to your pull requests.

[15,]

Inhaltsverzeichnis

1 Introduction

1.1 The Goal of Photofit Creation using GANs

Our proposed goal for the final project of the Computer Vision & Deep Learning: Visual Synthesislecture was to train existing Generative Adversarial Network (GAN) models and fine-tune their architectures in respect to generate authentic and unambiguous samples of real looking faces. These samples should be of such quality that they could be used as photofits (also phantom images, i.e. pictures representing a person's memory of a criminal's face, compare https://dictionary.cambridge.org/de/worterbuch/englisch/photofit-picture). Even if this common task is already done by phantom sketch artists working for the police or for lawyers, our assumption is that a GAN creating such photofits could easily outperform every sketch artist in terms of costs, speed, accuracy and photo-realism. Thus, such a network for the creation of photofits could be a very useful tool for criminology workers.

1.2 Text-to-Face Synthesis

Inspired by DALL-E and DALL-E 2 (see https://openai.com/dall-e-2/), our first intention was a text-to-image approach using two separate models – analogous to https://arxiv.org/abs/2012.03308 ??? TODO. On the one hand, we considered to use a text-encoder model for the embedding of a continuous

text describing a criminal's face into the latent space – such that the semantics of the textual description remain intact. On the other hand, we thought about a GAN or VAE model creating faces from those latent embeddings.

1.3 Vector-to-Face Synthesis

However, during the execution of the project we changed our plan to only focus on the generation part because of four crucial arguments. First, the lecture is about visual synthesis and not natural language processing, so our main focus should be on the creation of images and not on the semantic embedding of continuous text into the latent space. Second, training a text-encoder and a GAN respectively means twice as much calculation time which is inappropriate for the relatively short project time. Third, in respect to our described goal (see The Goal", TODO) we think that possible downstream applications would benefit more if the photofit creation is conditioned by vectors with values either 0 or 1 representing the truth value (0=False, 1=True) for each descriptive attribute of an image / a face. Fourth, the attempt of only generating phantom images based on attribute vectors is sufficient to get a proof of concept and to use such model for criminology purposes.

Therefore, to keep things simple and appropriate we decided to focus solely on the principle of using a GAN network – consisting of two separate models, i.e. a discriminator / encoder and a generator / decoder (TODO source GAN Godfellow).

We finalized a more stable adaption of a classical GAN, namely a Deep Convolutional GAN (DCGAN) which explicitly uses convolutional and convolutional-transpose layers in the discriminator and generator respectively (TODO source DCGAN). Since we do not just need to generate random images, but rather images that fit the vectorized description of a criminals face we conditioned our DCGAN to also use the attribute vector as input – then it is called a Conditional DCGAN (CDCGAN).

Moreover, we decided to re-implement another CDCGAN architecture: tediGAN (TODO source tediGAN). TODO @SchubertDaniel short description of tediGAN analogous to the paragraph before describing a CDCGAN

Due to time constraints and some other reasons (see TODO) we were not able to finalize the re-implementation of the tediGAN.

1.4 Related Work

The number of papers concering the same topic as ours, text-to-face generation from attributes, is small. We found 17 papers based on a broad range of keywords. The papers can be divided into two groups. One working with photofits/ forensic or composite sketches and the other with attribute guided face generation. The first group is mainly focused on generating images from those sketches [5,6,7]. This is not what we intended to do. The other group employs networks to generate faces from attributes which is a matches our goal [1,8-19]. We selected two papers which give relevant information for our project. First TediGAN: Text-Guided Diverse Face Image Generation and Manipulation"by Xia et al. published in 2021 [1]. The second paper is Ättribute-Guided Sketch Generation"by Tang et al. published in 2019 [4], which is the only paper bringing both aspects together.

2 Main

2.1 Dataset

2.1.1 Suitable Datasets

During our research in the project planning phase we stumbled across various datasets that could be useful to us in terms of their properties. Some examples of datasets containing faces and corresponding descriptions or attributes are:

- celebA
- celebA HQ
- LFW
- MAAD-Face

Nevertheless, these datasets are not made for criminology purposes, and so they do not perfectly fit our approach of generating photofits. This has two main reasons.

First, baseline face datasets are mainly constructed for face recognition applications. On the one hand, as the authors of MAAD-Face outline, this leads to the consequence that datasets like celebA or LFW indeed contain a large amount of face images, but struggle with the overall annotation correctness and the total number of attributes. On the other hand, MAAD-Face aims to be better in those terms by merging face image datasets with their attribute annotations together and check their correctness by a human evaluation.

Second, as we already expected beforehand and was confirmed during the project execution, such relatively low numbers of distinctive attributes (compare table, TODO) would not fit the demand for accuracy needed for phantom image creation. Moreover, considering the 40 attributes of celebA one can see that there is some redundancy and incompleteness within – e.g. one extra attribute for each Black_Hair, Blond_Hair, Brown_Hair, and Gray_Hair, but there is no attribute like Red_Hair.

TODO table with dataset stats

2.1.2 Our Decision

Comparing the statistics of the datasets from above, we concluded to initially use a set that has a good trade-off between the total number of face images and the total number of distinctive attributes. Even if they are not perfectly fitted for criminology purposes, our assumption is that if the concept of attribute-conditioned face generation works on one of these datasets, it will also work on more accurate datasets that could be developed especially for the task of photofit creation in the future. And especially, it will work better on a better suited dataset.

Even if MAAD-Face aims to be better than celebA and LFW, we decided to use celebA. MAAD-Face has too many images as if we could manage to train our GAN on it in the give time frame of the project and LFW has too

few images.

2.2 Framework

2.2.1 Architecture / Structure

2.2.2 CDCGAN

2.2.3 TediGAN

TediGAN is a GAN and Framework proposed by Xia et al. To generate their images they use a inverted pretrained StyleGAN. The Framework includes multiple options to choose between layers and StyleGANs. Unfortunately neither the git-repository nor their paper provides clear information which was their final and best version. The framework also uses config files in an incomprehensible way. We tried to implement the network into our framework but couldn't get it to start training. Problems we encountered were the configs, which we replaced by one single choice. They also wrote certain layers in C++ and Cuda which we initially struggled with but in the end got to work. But then we encountered a dimension error which was weird due to all shapes matching one another. To resolve the dimension error we logged and followed the flow of the images in the fit() method. To get further information about the configs we contacted the authors but never got an answer. To check for implementation errors we also cloned their repository and tried executing their proposed way to train with their framework. It failed to start due to missing config options.

2.2.4 Metrics

Regarding the metrics we orientated us among the most frequently used ones from the papers we read and chose the four most relevant. In regard of image generation normal metrics, like accuracy, are very relative and should not be used due to their lack of information value. In most ML context this would not apply. When training a discriminator to decide if a picture is from a real dataset or generated from the GAN's generator, the scores most of the time

do not result in "goodäs in real images. A human could easily differentiate both. So for image generation and their realness one should use one of the following metrics for evaluating the new images on the overall similarity: Fréchet Inception Distance (FID), patch similarity (Learned Perceptual Image Patch Similarity, LPIPS), Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) or a metric using a discriminator specifically trained for this task, where you know the results are satisfactory. FID calculates the Fréchet distance is originally used for the distance between curves of functions but can also be used for the probability distribution, in our case two datasets [3].

$$d^2 = |\mu_X - \mu_Y|^2 + tr(\sum_X + \sum_Y - 2(\sum_X \sum_Y)^{\frac{1}{2}})$$

One way to check patch similiarity between two images is using LPIPS. This metric is based on comparing similarity of activations in a predefined network. A lower score is better [20]. To check overall spatial image quality one could use BRISQUE. It checks for e.g. noise or blurriness. Kynkäänniemi et al. propose an improved precision and recall framework, which can additionally to precision and recall scores also calculate a realism score. They use a StyleGAN to evaluate a set of images [2]. Some papers also use humans to give feedback on realness of generated images, which can be unreliable and the number of samples to review is limited [1].

In our framework we implemented FID, LPIPS and BRISQUE. Kynkäänniemi's metric is implemented in an outdated version of TensorFlow. While implementing each metric a few key differences appeared which do not get clarified by any paper: e.g. LPIPS calculate the similarity between two pictures. But are they chosen at random or is an order selected in the beginning and then those images get compared? We choose to generate images based on the same attribute-vectors and compare those to one another.

2.2.5 Experiments

2.2.5.1 Configuration

2.2.5.2 Results

We planned the training process to run every model variation at least once for 100 epochs on a quarter of the celeba dataset. Each Training run took between 9 and 11 hours. After this preliminary phase we looked at the generated images, loss, accuracy and metrics of our networks. We then decided that a dropout of 0.3 or 0.5 and spectral convolution layer were beneficial. Thus we ran those on the entire dataset size. Those runs took 12 hours on our system. We also tried out running a model for 200 epochs but noticed mode collapse happened every time. Mode collapse also happend when restarting on an epoch without collapse.

Percieved realness is an intuitve score between 0 and 5: 0 just noise, 1 shape recognizeable, ge2 = can recognize faces, ge3= face with noise, ge4=face with small artifacts, 5=real faces without errors. DS_size means Dataset size.

When deciding which network was the best you can proceed based on statistics, on the proposed metrics or visually judgeing the generated images per epoch and foremost the last epoch. One could also proceed based on theoretically taught metrics, e.g. generator accuracy and discriminator accuracy should meet at 0.5 or at least converge against each other. In this case the network with spectral convolution and dropout value being 0.3 is supposed to be the best one. But even on first glance every human would be able to differentiate between real and fake images. Running metrics proposed in the metrics section on trained networks resulted in some mismatching images even compared to the generated ones in the last epoch. When going on the metrics FID is probably the most meaningful. The best network according to this metric would be with dropout 20% and without a spectral convolutional layer. Some generated images are just noise. Every now and then you can recognize parts of a face but the rest is still just random artifacts. So the results are only partly acceptable. The network with 0% dropout without a spectral

network	accuracy real accuracy fake before disc accuracy fake after disc	loss real loss fake loss gan	FID	LPIPS	BRISQUE	percieved realness (between 0 and 5)
dropout=0 spectral=False, DS_size=1/4	0.994 0.006 0.0007	0.0072 0.0074 13.868	339.957	0.379	16.946	1
dropout=0 spectral=True DS_size=1/4	0.968 0.031 0.009	0.046 0.411 6.815	139.774	0.161	34.917	2.5
dropout=0.2 spectral=False DS_size=1/4	0.894 0.105 0.062	0.210 0.211 6.640	124.138	0.156	31.376	2.5
dropout=0.2 spectral=True DS_size=1/4	0.969 0.028 0.062	0.0418 0.0356 6.6371	168.228	0.265	44.231	2
dropout=0.3 spectral=False DS_size=1/4	0.826 0.180 0.129	0.393 1.003 3.766	139.830	0.294	29.432	1
dropout=0.3 spectral=True DS_size=1/4	0.952 0.046 0.022	0.0612 0.0616 5.7318	173.991	0.205	41.228	1
dropout=0.3 spectral=True DS_size=1	0.999 7.308 7.210	1.12 0.0 44.171	218.512	0.271	28.791	1
dropout=0.5 spectral=False DS_size=1/4	0.506 0.501 0.496	0.750 1.737 0.768	145.371	0.327	73.911	1
dropout=0.5 spectral=True DS_size=1/4	0.821 0.179 0.131	0.263 0.262 3.003	141.477	0.353	45.244	2
dropout=0.5 spectral=True DS_size=1	0.929 0.070 0.048	0.108 0.107 5.587	135.339	0.213	64.889	1.5

Abbildung 1: Mutliple eyes in one face.

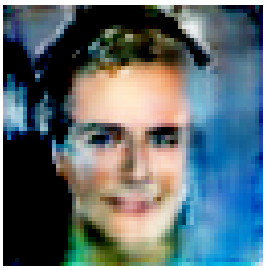

Abbildung 2: Network struggles to make entire head of hair in blonde.

layer is correctly the worst variant with the highest FID score of 339.957 and images looking like weird color sprinkles 6 When looking through all images of every last epoch the best network is tied between (dropout=50%, spectral=True, DS_size=1/4) and (dropout=0%, spectral=True, DS_size=1/4). But even these images still have some artifacts in the image or on the face, the images are highly noisy around the face. Sometimes the network tries to generate two faces into one 1. Also our networks seemed unable to learn the difference between certain attributes. E.g. they are not able to generate blonde people 2 or they put sunglasses on people 3. The sunglasses phenomenon happend with a constant c-vector which didn't specify it and the glasses would appear and disappear epoch-wise. A relative good result is 4 and 5. But even those fake images are distinguishable compared to the used dataset.

Abbildung 3: Network puts glasses on face.

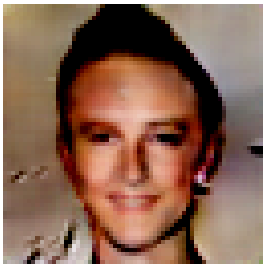

Abbildung 4: Relatively good picture after 100 epochs. Dropout=0.5, spectral=True

Abbildung 5: Relatively good picture after 100 epochs. Dropout=0.5, spectral=True

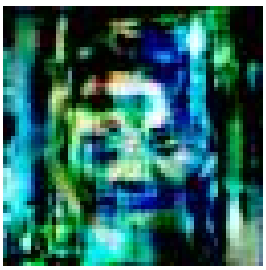

Abbildung 6: Really bad result.

Abbildung 7: Example image from dataset with artifacts.

3 Conclusion

In this chapter we want to conclude an reflect on the results and things we noticed while implementing or experimenting.

3.1 Datasets

First, we observed some defects in the training images that some faces were streched or had some artifacts e.g 7 8 9. Second, the attributes are redunand, incomplete. E.g. 4 attributes for hair color, Red_Hair is completly missing. Third it would be better if the direction in which a person looks, would only be straight ahead or be labeled. 10 11 Four, due to the small time frame we were unable to run our experiments with the other datasets mentioned above 2.1.1 Fifth, often it is the case that many faces match to one and the same attribute-vector so for criminology purposes one should consider to use a dataset with a much larger attribute-vector.

3.2 Mode Collapse

Mode Collapse is a common problem when working with GANs reference. Normally a GAN is consider successful if it's samples can fool a discriminator and the generator samples diverse images with a distribution like in the real world. This mean that given an attribute-vector c the GAN should sample different images. Mode collapse happens if the GAN starts to produce the

Abbildung 8: Example image from dataset with artifacts.

Abbildung 9: Example image from dataset with artifacts.

Abbildung 10: Example image from dataset: person looking down and face hidden behind hair and hat.

Abbildung 11: Example image from dataset: person from the side.

Abbildung 12: Example for mode collapse after 100 epochs.

same image again and again for the same c because it successfully fools the discriminator. 12 The image is a result after 100 epochs with spectral convolutions and 0% dropout. As you can see the images all look alike and there is no real difference between them. Also we have checked some attribute-vectors and as one might assume there are multiple individuals annotated with the same vector. This seems pretty plausible with a dataset of more than 200 thousand images and a relatively small attribute-vector size of 40. Therefore we also expected our GAN to sample different images for the same c. However thinking of photofits it could be useful to run into mode collapse, here a specific vector should always lead to the same result.

3.3 Imagesize

For time and hardware reasons we opted to scale the dataset images to size 64 by 64. We assume that we can achieve better results with bigger sized images due to a higher detail level.

3.4 More time + GPU-power

With more time and better hardware (maybe even multiple GPUs) we would have done more extensive testing and more training runs. And prove our hypothesis that a more detailed dataset and larger images lead to a general improvement.

4 Future Work

As described in the conclusion we would like to train our GAN on other datasets(referenceToDatasets) and with a larger image size(referenceToImageSize). Also the development of a new dataset specialized for criminology purposes would be helpful for us and downstream applications. Such a dataset should contain much more than fourty attributes, we assume about 200 would be a good starting point. Moreover the images of the dataset should not include bad samples as pointed out above but only passport photo should be included. Regarding mode collapse it would be interesting if someone could use this phenomenon for photofit generation.

5 Collaboration

We started our project commonly by exchanging our ideas, thoughts and how to approach our topic. We did most of the implementation of our framework by pair programming. Certain parts which only one of us did are marked in the table below.

Implementation	Person
main.py error.py log.py	Max
CDCGAN.py	70-30 Max-Daniel
TediGAN.py	30-70 Max-Daniel
Metrics.py	Daniel
Config.py	Max
Dataset.py	Max
Training.py	80-20 Max-Daniel
Experiments	Daniel

Report-chapter	Person
1.1	Max
1.2	Daniel
2.1	Max
2.2.1	Max
2.2.2	Max
2.2.3	Daniel
2.2.4	Daniel
2.2.5.1	Max
2.2.5.1	Daniel
3	Together
4	Together
5	Together

Literatur

[1] Navaneeth Bodla, Gang Hua, and Rama Chellappa. Semi-supervised fusedgan for conditional image generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[2] Zhu-Liang Chen, Qian-Hua He, Wen-Feng Pang, and Yan-Xiong Li. Frontal face generation from multiple pose-variant faces with cgan in real-world surveillance scene. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1308–1312, 2018.

[3] Jia Deng, Gaoyang Pang, Zhiyu Zhang, Zhibo Pang, Huayong Yang, and Geng Yang. cgan based facial expression recognition for human-robot interaction. *IEEE Access*, 7:9848–9859, 2019.

[4] Prithviraj Dhar, Ankan Bansal, Carlos D. Castillo, Joshua Gleason, P. Jonathon Phillips, and Rama Chellappa. How are attributes expressed in face dcnns? In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 85–92, 2020.

[5] Cambridge dictionary. photofit (picture). https://dictionary.cambridge.org/de/worterbuch/englisch/photofit-picture. Accessed: 2021-08-11.

[6] Jon Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class project for Stanford CS231N: convolutional neural networks for visual recognition, Winter semester*, 2014(5):2, 2014.

[7] Baris Gecer, Binod Bhattarai, Josef Kittler, and Tae-Kyun Kim. Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3d morphable model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[8] Hu Han, Brendan F. Klare, Kathryn Bonnen, and Anil K. Jain. Matching composite sketches to face photos: A component-based approach. *IEEE Transactions on Information Forensics and Security*, 8(1):191–204, 2013.

[9] 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 I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[10] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[11] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[12] 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.

[13] Yingtao Lei, Weiwei Du, and Qinghua Hu. Face sketch-to-photo transformation with multi-scale self-attention gan. *Neurocomputing*, 396:13–23, 2020.

[14] Yaoyao Liu, Qianru Sun, Xiangnan He, An-An Liu, Yuting Su, and Tat-Seng Chua. Generating face images with attributes for free. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6):2733–2743, 2021.

[15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[16] Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. Attribute-guided face generation using conditional cyclegan. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[17] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[18] openai. Dall-e. `https://openai.com/dall-e-2/`. Accessed: 2021-08-11.

[19] Naima Otberdout, Mohamed Daoudi, Anis Kacem, Lahoucine Ballihi, and Stefano Berretti. Dynamic facial expression generation on hilbert hypersphere with conditional wasserstein generative adversarial nets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):848–863, 2022.

[20] M.S. Sannidhan, G. Ananth Prabhu, David E. Robbins, and Charles Shasky. Evaluating the performance of face sketch generation using

generative adversarial networks. *Pattern Recognition Letters*, 128:452–458, 2019.

[21] Hao Tang, Xinya Chen, Wei Wang, Dan Xu, Jason J. Corso, Nicu Sebe, and Yan Yan. Attribute-guided sketch generation. 4:1–7, 2019.

[22] Philipp Terhörst, Daniel Fährmann, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Maad-face: A massively annotated attribute dataset for face images. *CoRR*, abs/2012.01030, 2020.

[23] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[24] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2256–2265, June 2021.

[25] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[26] Zheng Yuan, Jie Zhang, Shiguang Shan, and Xilin Chen. Attributes aware face generation with generative adversarial networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 1657–1664, 2021.

[27] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CoRR*, abs/1801.03924, 2018.