27^{th} Computer Vision Winter Workshop
Terme Olimia, Slovenia, February 14–16, 2024

Detecting and Correcting Perceptual Artifacts in Synthetic Face Images

Adéla Šubrtová, Jan Čech
Faculty of Electrical Engineering,
Czech Technical University in Prague
subrtade@fel.cvut.cz

Akihiro Sugimoto
National Institute of Informatics
Tokyo, Japan

Abstract. *We propose a method for detecting and automatically correcting perceptual artifacts on synthetic face images. Recent generative models, such as diffusion models, can produce photorealistic images. However, these models often generate visual defects on the faces of people, especially at low resolutions, which impairs the quality of the images. We use a face detector and a binary classifier to identify perceptual artifacts. The classifier was trained on our dataset of manually annotated synthetic face images generated by a diffusion model, half of which contain perceptual artifacts. We compare our method with several baselines and show that it achieves superior accuracy of 93% on an independent test set. In addition, we propose a simple mechanism for automatically correcting the distorted faces using inpainting. For each face with artifact response, we generate several replacement candidates by inpainting and choose the best one by the lowest artifact score. The best candidate is then back-projected into to the image. Inpainting ensures a seamless connection between the corrected face and the original image. Our method improves the realism and quality of synthetic images.*

1. Introduction

Synthetic image generation has made a giant leap in recent years, thanks to the development of powerful generative models, such as generative adversarial networks (GANs) [10, 15] and diffusion models [24, 23]. These models generate photorealistic images that are often indistinguishable from real photographs by human observers. However, they also sometimes produce visually unpleasant and distracting artifacts, including distorted faces.

In this paper, we focus on detecting and correcting perceptual artifacts in synthetic face images. We use the Stable Diffusion – Realistic Vision model [5], which is a popular text-to-image model that can generate high-quality images from complex captions. We observe that, although this model can generate amazing images, it often produces artifacts on the faces of people, especially at low resolutions.

Unlike GANs, which have a known "truncation trick" [3] to avoid artifacts by restricting the latent codes to a narrow range (near the mean latent vector), diffusion models do not have such a simple technique to control the trade-off between the quality and the diversity of generated images. Therefore, we propose to train a detector to identify perceptual artifacts on synthetic face images, and use its output to automatically correct the generated faces. See Fig. 1 for an example. Our contributions are as follows.

- We trained a binary classifier to detect perceptual artifacts on face images generated by the diffusion model by learning on our dataset. We manually annotated a set of 1274 images where a half of the samples contained perceptual artifacts.
- We compared our method with several baselines, such as the size of the synthetic face, the response score of the face detector, the response of the LAION Aesthetics predictor [25], and a recent perceptual artifact detector PAL4VST [33], showing that our method achieves superior accuracy in detecting artifacts.
- We proposed a fully automatic method for fixing distorted faces generated by the diffusion model, using inpainting. For each face with the artifact response, we generate several replacement candidates by inpainting and choose the best one by the lowest artifact score.

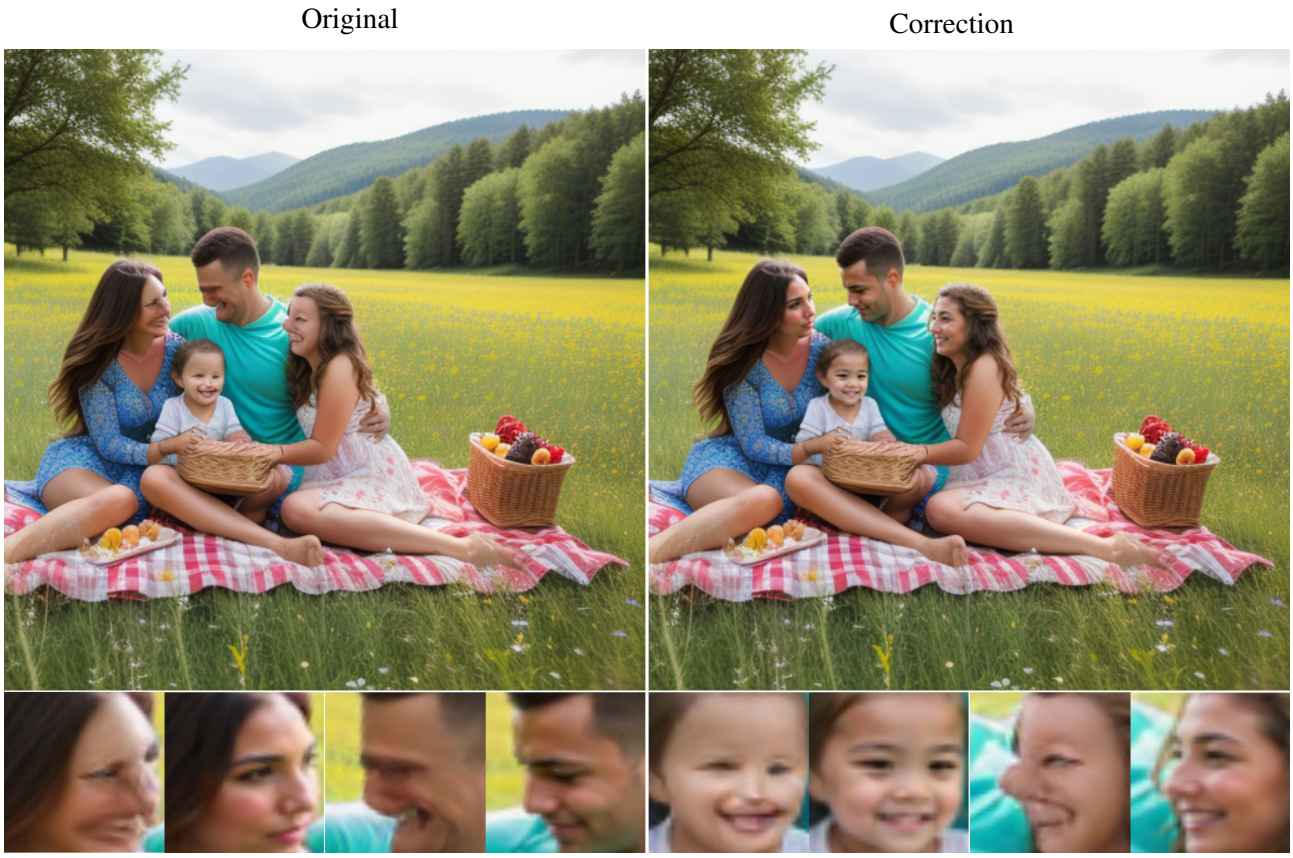

Figure 1: Detection and correction of perceptual artifacts on synthetic faces performed fully automatically by our method. Left image is the input, an original image generated by Realistic Vision model [5] with the prompt "A family enjoying a picnic in a vibrant, flower-filled meadow". Right image shows the result of our method. Bottom images are zoomed details of distorted/corected face pairs.

The rest of the paper is structured as follows. Related work is reviewed in Sec. 2, the method is presented in Sec. 3, experiments are given in Sec. 4 and finally, Sec. 5 concludes the paper.

2. Related work

Long before the availability of photo-realistic synthetic generators, researchers aimed to assess the quality of images rather from a technical perspective (for sharpness, noise, compression, etc.) [26, 18]. Early attempts to assess the perceptual image quality were made even before the boom of deep learning. Paper [28] classified photos taken by amateurs and professional photographers, or paper [7] learned a simple classifier on hand crafted features using a dataset from peer-rated photo website.

Recently, there have emerged many works on image aesthetics assessment. To name a few, the LAION Aestitics predictor [25] learns a simple multi-layer perceptron on CLIP embeddings [22], given crowd sourced aesthetics score. Paper [16] learns the aesthetic score indirectly from user comments of online images. 'Naturalness' of an image is learned in [4]. For a comprehensive review of these methods, we recommend surveys [8, 1].

A standard approach to assess the quality of a generative model, is to use the Fréchet Inception Distance (FID) [11]. However, it assesses both the quality and diversity of generated images and is not defined for a single sample, but needs a large set of generated images.

The recent FreeU [27] promises a universal improvement of visual quality of diffusion models, without any additional training, by simply reweighting the skip connections in the denoising U-NET. However, the quality improvement seems to be at the cost of diversity and even prompt fidelity. A different approach [2] to improve the generator quality is to train the diffusion model by reinforcement learning, possibly using the aesthetic reward.

More closely related to our work are papers that learn perceptual artifacts in synthetic images. Paper [31] detects artifacts in super-resolution GANs, paper [34] detects artifacts in inpainting. The recent work [33] learns a predictor to localize the perceptual artifacts in images produced by recent synthetic generator models including the Stable Diffusion [24]. The paper also proposes a mechanism similar to ours to correct the artifacts. We compare with their results and show that our method has superior artifacts detection accuracy. Our automatic correction differs in the mechanism to select the best one out of several candidate replacements.

Our problem is indirectly related to out-of-distribution (OOD) [32] detection problem, where only the in-distribution samples are available for training. Although face images form a relatively compact domain, we observe that artifacts generated by the diffusion model are so specific that the supervised classification problem is more appropriate. Natural drawback of this choice is that we are model dependent and have to retrain for a new model.

Another related problem is forensic detection of synthetic images, a.k.a. 'deepfake' detection [20]. It might sound easy to train synthetic vs. real face image classifier and use it to spot images with artifacts. However, it is not true that this classifier will respond with higher synthetic score on images with obvious perceptual artifacts. We will show this experiment among our baselines. The reason is that the real vs. synthetic classifier learns low-level signal features (as reported e.g. by [30, 6]) and the higher-level content seems to be overlooked.

3. Method

Our aim is to develop a method to detect artifacts in synthetic images and correct them automatically. This work focuses on artifacts in the facial area, firstly, because human perception is very sensitive to faces and secondly, because a lot of artifacts in recent generative models are concentrated in the facial area. Specifically, our data-oriented method consists of two modules a detection module, see Sec. 3.1, and automatic face artifact removal module, see Sec. 3.2.

3.1. Artifact detection module

The artifact detection module consists of an off-the-shelf face detector [13] and a face artifact (binary) classifier. For the architecture we choose the powerful vision transformer for image classification [9]. The training is done in a supervised manner on our manually annotated synthetic dataset.

(a) A face framed by a hooded sweatshirt on a chilly day.

(b) A group of friends gathered around a bonfire, their faces illuminated by the flames.

(c) A farmer driving a tractor through a field of corn.

Figure 2: Examples of generated images alongside with their prompts.

Synthetic dataset Realistic Vision [5] is a popular text-to-image diffusion model. Each generated image requires as an input a Gaussian noise and textual prompt to guide the diffusion process.

To make the synthesis fully automatic, we generated random prompts using ChatGPT [21]. The queries for ChatGPT aimed to produce textual prompts describing images containing (1) people with focus on whole-body shots (e.g. Fig 2b) and (2) people's portraits (e.g. Fig 2a). We obtained 200 prompts in each of the queries, 400 in total[1]. During the dataset synthesis, we randomly sampled a prompt and an initial Gaussian noise to produce the images. We used the default negative prompt for the Realistic Vision model as recommended by its authors.

With this process, we synthesized a set of 3k images and manually separated the samples into two classes – with and without artifacts. The presence of artifacts is not a binary property in fact, as the boundary appears rather fuzzy, and for certain images it is very challenging and subjective to decide one of the two classes. Hence in our dataset, we include only the most severe and disturbing artifacts. Given the random nature of the generated prompts, some images had to be completely discarded, because they did not contain any visible face (See Fig. 2c).

Subsequently, we detected faces in the collected images using the YOLO v8 Face detector [13]. Faces with size smaller than 50 pixels were discarded. All

[1]Image dataset with the prompts will be released.

faces were aligned, so that the eye-keypoints line [2] was parallel with the horizontal axis.

In total, the dataset of 1274 images was randomly split, such that the training set consisted of 406 images for each class, validation set of 97 for each class and the test set of 134 faces for each class.

3.2. Automatic face artifact removal

We propose a simple mechanism to automatically and seamlessly rectify faces with artifacts in synthetic images.

The idea is to replace faces with detected artifacts by generative inpainting. Inpainting is a process used in image editing where unwanted parts of an image are filled in seamlessly to fit the overall context. We used the same generative model to do the inpainting [5]. Since the model struggles with generating faces at low resolution, we zoom in around the face bounding box to increase the likelihood that the inpainted face were artifact-free. Moreover, we always generate several inpainting candidates and decide the best one by our classifier response.

Our method consists of the following steps:

1. In the generated image, we find a face for which our classifier is positive for artifacts.

2. Using inpainting, we generate N candidates for replacement. Note that we zoom in, such that the face bounding box is magnified by factor m and inpaint the pixels inside the original bounding box.

3. For each of the N replacement candidates, we measure the response of our artifact classifier and choose the winner as one with the lowest score. See Fig. 3 for an example of replacement candidates sorted from highest to lowest artifact score.

4. The winning candidate is finally subsampled by $1/m$ to the original scale and projected back into the original image.

We cannot enlarge the face to the maximum possible size, because inpainting requires some context. If the context is insufficient, i.e., the area around the face region is too small and uninformative, the resulting inpainting does not match the original image (in terms of content, geometry, and lighting/shading). Therefore, we zoom in by factor $m = 2$, which is emprically found as a trade off between model realism and consitency with context. Inpainting itself ensures that the connection with the original image is seamless and no additional blending is needed.

We set the number of replacement candidates $N = 10$, as a trade-off between quality and computational time. More candidates increase the chance of finding a better candidate, but the system is less responsive.

This way we can effectively remove face artifacts and thus improve the perceptual quality and realism of generated images.

[2]Facial keypoints are also returned by the YOLO Face detector.

3.3. Implementation details

We initialized the classifier network with weights pretrained on the ImageNet dataset. The network was trained for 10 epochs with AdamW [19] optimizer and the initial learning rate of 5e-05. During training, we employed a linear learning rate scheduler and augmented our dataset by mirroring each example and adding it to the dataset. The images were resampled to ViT input resolution 224×224 pixels. Following the preprocessing of the pretrained ViT, we use normalization across the RGB channels with mean [0.5, 0.5, 0.5] and standard deviation [0.5, 0.5, 0.5].

For inpainting in the correction module, we used the same generative model and HuggingFace's diffusers library [12] (v0.17.1) with the following settings: num_inference_steps=200, strength=0.45, guidance_scale=15.5.

4. Experiments

To evaluate our method, we conducted number of experiments. Firstly, we report quantitative evaluation, comparing our classifier to other methods for artifact detection. To the best of our knowledge, there exists only one paper contributing directly on this topic, that is PAL4VST [33]. For that reason, we propose several additional baselines to compare our model with. Secondly, we present the qualitative evaluation of the baselines by ranking the test set according to responses of each classifier. Finally, we show results of the entire detection and automatic correction pipeline.

4.1. Baselines

Face-size based classifier. We observe high correlation between face size and the severity of artifacts. The size was determined from face detections found

Figure 3: Replacement candidate ranking. To find a replacement for the original face image with artifacts (left), we generate multiple candidates using inpainting and sort them based on the response of our artifact classifier. Subsequently, we select the one with the best response as a replacement for the original face.

by the YOLO v8 face detector [13]. For non-square bounding boxes, we took the longer side. The classification threshold that maximizes classification accuracy was determined on the validation set.

Laion Aesthetics predictor. The Laion Aesthetics predictor [25] was trained to predict an aesthetics score in range $[0, 10]$ based on the visual appearance of an image, 10 being the best. The threshold was again found to maximize the validation accuracy. The model was trained on whole images, thus we asses this baseline in two modes, one with whole images as inputs and second mode with the face crops.

Face-detection-score based classifier. The YOLO v8 face detector [13] is our next choice for a baseline; specifically, the confidence score for each bounding box. Yet again, we determine the classification threshold on the validation set.

Perceptual artifact localisation (PAL4VST). Zhang et al. [33] train a segmentation transformer for artifact localization in synthetic images generated by multiple generative models (including the Stable Diffusion [24]). The output is a segmentation mask where active pixels mark the areas with artifacts. Since the method expects whole images, we test again two scenarios, face crops and whole images. To compare this method to our facial artifact detection, we inferred the classification labels as follows. We consider the prediction as "with artifacts" if at least one pixel in the output mask was active for the face crop or inside the face bounding box in case of whole images. Otherwise, the predicted label was "no artifacts".

Synthetic vs Real classification baseline. As next baseline, we consider a classifier between real and synthetic images. The classifier was trained in a supervised manner with 10k images in each class. The synthetic class was generated as described in Sec. 3.1 with the recommended negative prompt. As the real class, we used randomly selected subset of images of the FFHQ dataset [14] and cropped the faces in the same way as in the synthetic class.

We trained ViT, started from ImageNET model, but trained only the last layer and kept other weights frozen. This model achieved 99% accuracy in discriminating synthetic vs real images. We observed, that *artifact detection* accuracy was higher than when training the entire model. We hypothesize that the latter option learns the low-level signal features, as reported by [30, 6], and not the image content.

The threshold for artifact detection was again set on the validation set.

Model	Acc	AUC
Face-size	0.8731	0.9213
Laion Aesthetics [25]	0.8134	0.9420
Face detector score	0.5896	0.6475
PAL4VST [33] (face crops)	0.7164	0.7981
Synth/Real (last layer finetuned)	0.7761	0.8651
Ours	**0.9254**	**0.9678**
Laion Aesthetics [25] (whole images)	0.5633	0.5805
PAL4VST [33] (whole images)	0.6531	0.7766

Table 1: Quantitative results. Classification Accuracy (Acc) and Area under the precision-recall curve (AUC) calculated on our test set.

4.2. Quantitative results

The comparison between our artifact detector and the baselines is presented in Table 1. Namely, we

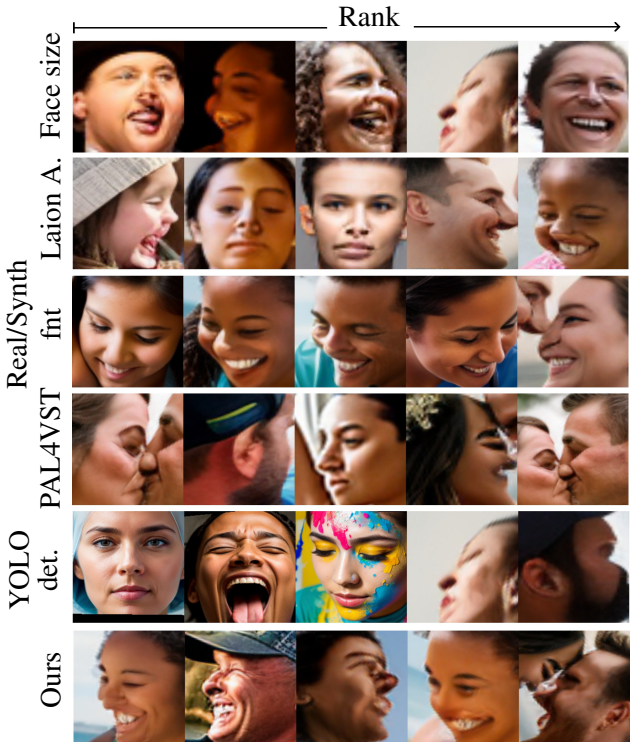

Figure 4: Image ranking – worst first. Each row depicts five worst images from the test set. Ranking is based on the response of each classifier.

Figure 5: Image ranking – best first. Each row depicts five best images from the test set. Ranking is based on the response of each classifier.

report classification accuracy (Acc) and the area under the precision-recall curve (AUC). Our method achieves superior results for both metrics.

As expected, the simple face-size based classifier is a strong baseline. It confirms the artifacts are most common in faces in low resolution, but might be present in higher resolution, too.

Laion Aesthetics predictor in the whole image setting is weaker. Likely, the mismatch between detecting artifacts and predicting aesthetic quality is significant. Ranking in Fig. 5 suggests that the most aesthetics of an image reside in colorfulness and not in structural correctness. We also observe that the version with cropped faces is significantly more accurate, probably because the artifacts are more prominent.

Face detector score is a surprisingly weak baseline. We hypothesize that unlike classical scannig Viola-Jones [29] detector, YOLO [13] decides on a larger context (i.e., a human body), the distorted faces do not impact the score much. Faces with severe artifacts were confidently detected on our dataset.

PAL4VST [33] does not perform very well to detect face artifacts either on the face crops or whole images, despite it is a recent method trained on a much larger dataset of generated images including Stable Diffusion.

Synth/Real classifier is another rather weak baseline. It is a proxy problem that does not solve the target artifact detection task very well.

4.3. Ranking experiment

To qualitatively compare all the models, we conduct ranking experiment on held-out test set. Each test image is evaluated using each model and ranked by its response; in the case of PAL4VST, we rank by the size of the region with artifacts, i.e., the number of active pixels in the segmentation mask. Images with the most severe artifacts are depicted in Fig. 4, the cleanest or the most photo-realistic are in Fig. 5.

We can see that different baselines returned different ranking, which indicates each model focus on different features. Laion Aesthetics predictor returned rather visually pleasant (colorful) images as the best. PAL4VST returned very distorted images as the best ones, YOLO detect response returns several good images among the worst ones. The ranking confirms quantitative results in Tab. 1.

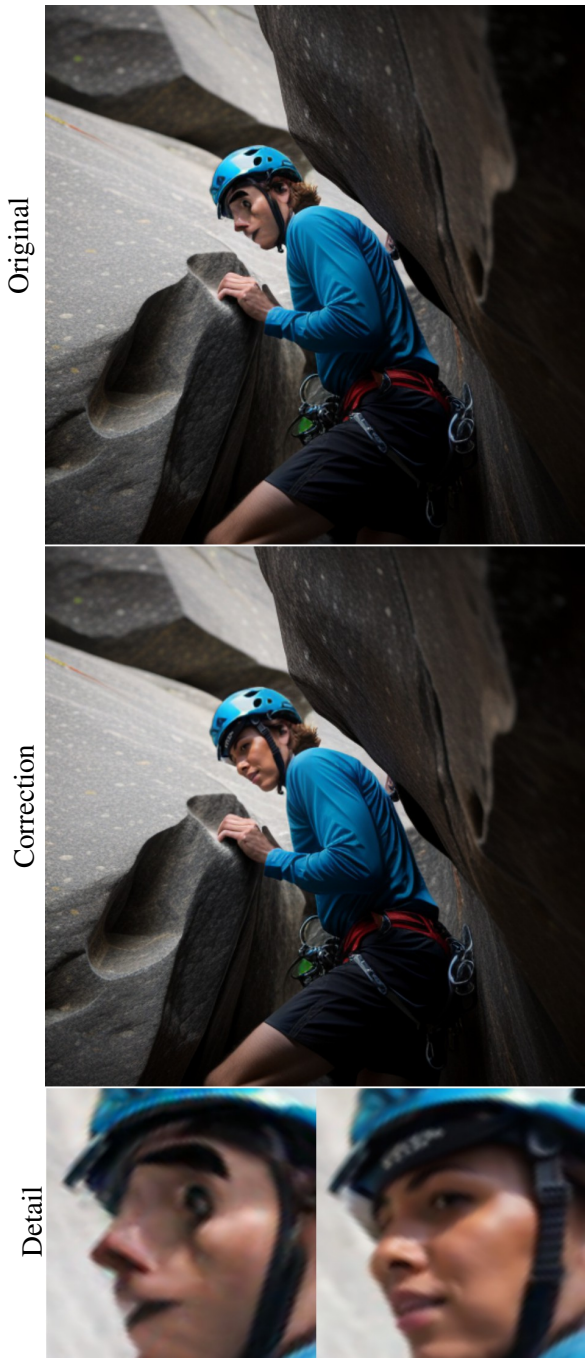

Figure 6: Example of the application of our method. Original image (top) contains severe artifacts in the facial area. Artifacts are discovered using our pre-trained classifier and multiple candidates for replacement are generated using inpainting. The candidates are again evaluated by our classifier and ranked according to its response. The one with the best score is selected as the replacement. The corrected images are shown in the middle row, while the details of the faces are depicted in the bottom row.

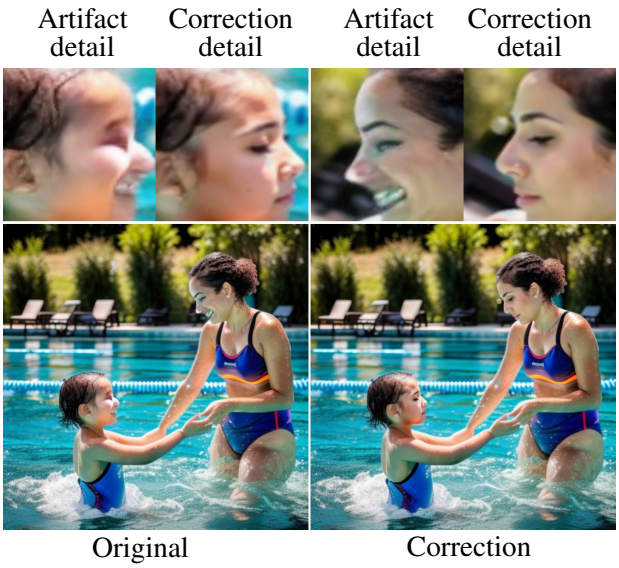

Figure 7: An example of automatic correction of face artifacts in a synthetic image. The original image contains unnatural facial features. The newly-generated faces look much more realistic.

4.4. Results of the entire pipeline

Finally, we show results of the entire pipeline (detection and correction) on several images. See Figs. 1, 6, 7, 8 for examples. Our detector finds distorted faces and correctly selects a good replacement candidate. The result is a seamless correction of faces with artifacts.

5. Conclusion

In this work we propose an artifact classifier for synthetic face images trained on our manually annotated dataset. We provide comparison with several baselines such as face-size based classifier, LAION Aesthetics predictor or the recent perceptual artifact detector [33], showing that our method achieves superior classification metrics in face artifact detection.

Furthermore, we demonstrate that our method is applicable in automatic correction of the facial artifacts caused by recent diffusion models. Specifically, we generate multiple replacement candidates of the face with artifacts using standard inpainting. Subsequently, we evaluate the new face candidates with our classifier and, in the end, we select the candidate with the lowest artifact score as the replacement.

Limitations and future work. One of the weaknesses of our method is the fact that during the automatic artifact correction, we use quite an ambigu-

Figure 8: Example of automatically rectified face artifacts, produced by our method.

ous prompt "face" to regenerate the image. Due to this fact, we do not have any guarantee that the corrected face will be of the same age or gender, we only rely on the context. In minor cases, semantically incompatible faces are found. That might be avoided by keeping the original prompt if available or estimate the prompt with off-the-shelf image captioning model such as BLIP [17].

Acknowledgements

This work was supported by the NII international internship program and by the CTU Study Grant SGS23/173/OHK3/3T/13.

References

[1] A. Anwar, S. Kanwal, M. Tahir, M. Saqib, M. Uzair, M. K. I. Rahmani, and H. Ullah. A survey on image aesthetic assessment. *arXiv preprint arXiv:2103.11616*, 2021. 2

[2] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 2

[3] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 1

[4] Z. Chen, W. Sun, H. Wu, Z. Zhang, J. Jia, X. Min, G. Zhai, and W. Zhang. Exploring the naturalness of ai-generated images. *arXiv preprint arXiv:2312.05476*, 2023. 2

[5] CivitAI. Realistic vision, v5.1, 2023. `https://civitai.com/models/4201/realistic-vision`. 1, 2, 3, 4

[6] R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, and L. Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proc. CVPR*, pages 973–982, 2023. 3, 5

[7] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part III 9*, pages 288–301. Springer, 2006. 2

[8] Y. Deng, C. C. Loy, and X. Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4):80–106, 2017. 2

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and

N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[11] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2

[12] Hugging Face. Diffusers – Inpainting. 2023. `https://huggingface.co/docs/diffusers/using-diffusers/inpaint`. 4

[13] A. Kanametov. Yolo v8 face detector, 2023. `https://github.com/akanametov/yolov8-face`. 3, 5, 6

[14] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proc. CVPR*, 2019. 5

[15] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proc. CVPR*, 2020. 1

[16] J. Ke, K. Ye, J. Yu, Y. Wu, P. Milanfar, and F. Yang. VILA: learning image aesthetics from user comments with vision-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10041–10051, 2023. 2

[17] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 8

[18] Q. Li and Z. Wang. Reduced-reference image quality assessment using divisive normalization-based image representation. *IEEE journal of selected topics in signal processing*, 3(2):202–211, 2009. 2

[19] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 4

[20] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, 223:103525, 2022. 3

[21] OpenAI. Chatgpt v3.5, 2023. `https://chat.openai.com/chat`. 3

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[23] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[24] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 2022. 1, 3, 5

[25] C. Schuhmann and LAION team. LAION-AESTHETICS, 2022. `https://laion.ai/blog/laion-aesthetics/`. 1, 2, 5

[26] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006. 2

[27] C. Si, Z. Huang, Y. Jiang, and Z. Liu. FreeU: Free lunch in diffusion u-net. *arXiv preprint arXiv:2309.11497*, 2023. 2

[28] H. Tong, M. Li, H.-J. Zhang, J. He, and C. Zhang. Classification of digital photos taken by photographers or home users. In *Advances in Multimedia Information Processing-PCM 2004: 5th Pacific Rim Conference on Multimedia, Tokyo, Japan, November 30-December 3, 2004. Proceedings, Part I 5*, pages 198–205. Springer, 2005. 2

[29] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, 2001. 6

[30] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros. CNN-generated images are surprisingly easy to spot...for now. In *Proc. CVPR*, 2020. 3, 5

[31] L. Xie, X. Wang, X. Chen, G. Li, Y. Shan, J. Zhou, and C. Dong. DeSRA: Detect and delete the artifacts of gan-based real-world super-resolution models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, 2023. 3

[32] J. Yang, K. Zhou, Y. Li, and Z. Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 3

[33] L. Zhang, Z. Xu, C. Barnes, Y. Zhou, Q. Liu, H. Zhang, S. Amirghodsi, Z. Lin, E. Shechtman, and J. Shi. Perceptual artifacts localization for image synthesis tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7579–7590, 2023. 1, 3, 4, 5, 6, 7

[34] L. Zhang, Y. Zhou, C. Barnes, S. Amirghodsi, Z. Lin, E. Shechtman, and J. Shi. Perceptual artifacts localization for inpainting. In *European Conference on Computer Vision*, pages 146–164. Springer, 2022. 3