

FACE QUALITY TRANSFORMER: A FACE QUALITY ASSESSMENT AND ENHANCEMENT FRAMEWORK

Author(s) Name(s)

Author Affiliation(s)

ABSTRACT

The quality of image generation has reached impressive levels. Advanced text-to-image models have become amazingly good at creating objects, depicting actions with high precision. However, despite significant progress in image generation, the quality of generated faces remains a critical factor for users. Even the most advanced text-to-image diffusion models struggle to generate high quality faces consistently. This highlights the importance of estimation of face quality in generated images as one of the most important metric to assess. In this paper, we propose a hybrid architecture comprising of attention-based Vision-Transformer along with EfficientNet to capture intrinsic face deformations present in image generation models. We also conduct a comparative analysis of state-of-the-art diffusion models for face quality estimation including DALL-E, Flux, Stable Diffusion and Firefly. Furthermore, we show that our proposed pipeline can be plugged with image generation models to effectively correct the poor quality faces in generated images through automated re-generations.

Index Terms— Face Quality Estimation, Image Quality Assessment, Diffusion Models, Vision Transformers

1. INTRODUCTION

Advancements in text-to-image generation using diffusion models have made it possible for generated images to be used in multiple applications, e.g. in creative industries such as designing, advertising, media houses, social media banners, posts and virtual reality. With rapid advancements, the focus is shifting towards more practical use-cases of these models in a wide variety of applications. Even though these models are capable of generating unimaginable objects, beautiful creative images and compositions, generating high-quality human faces still remains a critical issue. Most advanced diffusion models do a decent job in generating large-size portrait images. However, these models fail miserably to generate multiple faces in group pictures or faces with appropriate expressions, people performing different actions, and faces with small face crops in generated images. Major problems associated with the faces generated by text-to-image diffusion models include pixelation, distortions, unnatural-look, faces

with exaggerated shape or size, facial features skewed or bent at unusual angles, asymmetrical facial features, unrealistic skin textures, inconsistent lighting and shadows, incomplete facial features and abnormal facial expressions. Observing facial features such as symmetry, skin texture, clear eyes, nose or lips positions can make it easier to distinguish between a bad and a good quality face. However, this manual assessment is highly subjective, as it may be influenced by variations in age (wrinkled faces), ethnicity, lighting conditions, environmental conditions, image source, surrounding noise or other unforeseen scenarios. With this work, our aim is to detect poor quality faces in generated images by calculating a Face Quality Score, thereby flagging images below a certain threshold score, which makes it possible to correct the faces automatically. Through this work, we propose:

1. A hybrid network 'FaceQ Transformer' utilizing pyramid VIT architecture along with EfficientNet to capture the global and local deformation patterns present in a face crop.
2. An automated pipeline to compare the quality of faces generated by various text-to-image diffusion models.
3. A dataset of 30,000 'bad quality' face-crops to advance the research in measuring the quality of faces generated with text-to-image diffusion models.

2. RELATED WORK

Image quality assessment has been at the center stage due to its importance in both image and video generation. There has been extensive research in estimating the overall quality of generated images [1, 2]. Inception score [3] and FID [4] are some of the popular objective metrics to measure the overall quality of generated images. Although these metrics focus on the global quality of an image, they overlook specific portions such as faces and hands [5] within the image. In the domain of image forensics and deepfakes, there has been significant research in detecting bad faces introduced by factors such as motion blur and image editing. Face quality estimation has also been a focus in applications such as face recognition, where the primary focus involved looking at the symmetry of detected faces, lighting conditions, and noise surrounding face crops in video frames or images [6, 7, 8]. Fur-

thermore, there have been some notable efforts on quality assessment of faces and overall images generated using Generative Adversarial Networks (GANs) [6, 9, 10]. In prior works [11, 12], the authors have utilized the forward and backward processes of DDPMs to perturb facial images and quantify the impact of these perturbations on the corresponding image embeddings for purpose of quality prediction. Similarly in [13], the authors have measured the classifiability based on the allocation of the training sample feature representation in angular space with respect to its class center and the nearest negative class center. They utilize internal network observations during the training process to predict the quality of unseen samples. Although this work is closely related to our research, we have observed that the deformation textures or properties are different in text-to-image generation using diffusion models differ significantly from those found in poor-quality camera-captured images or images generated through GANs. Researchers have also attempted to solve the low-quality generation of faces using diffusion models in [14, 15, 16, 17]. For instance, in [17], the authors have introduced a Face Score based on in-painted facial regions within diffusion-generated-images, assuming that the in-painted face quality will be worse than the original image. We have experimented with the introduction of in-painted face crops as markers of low-quality face crops, but this approach does not scale well in terms of detection across all generative models since the noise and distortion patterns vary across different diffusion models. In [18], the authors have generated a dataset of faces for DALL-E, Midjourney and Stable Diffusion models, pointing out some of the key face distortions observed in diffusion models. It has become common to use negative prompts based on Classifier Free Guidance (CFG) technique[19] to limit poor-quality image generations. Although this approach is effective, it is limited by the model’s capability to generate high quality faces. Some work in estimating or verifying the quality of generated faces has been done using VLLMs (VQA) [20, 21, 22]. On doing a comprehensive analysis of measuring the subjective and objective quality of bad face crops with VLLMs, we found this technique to be ineffective (internVL, Table 1). Even the most advanced VLLM models such as GPT-4V, GPT-4o, Intern-VL, LLaVA with different backends (mistral, llama, phi, vicuna), fail miserably at estimating the quality of generated faces.

3. DATASET CREATION

3.1. Need for a new dataset

Dataset plays a crucial role in steering the output of a machine learning model. Although good datasets for faces covering diverse profiles, orientations and variations [23, 24, 25] are available, most of them are designed predominantly for the task of facial recognition and to assess the quality of camera-captured faces. It is important to note that for datasets involv-

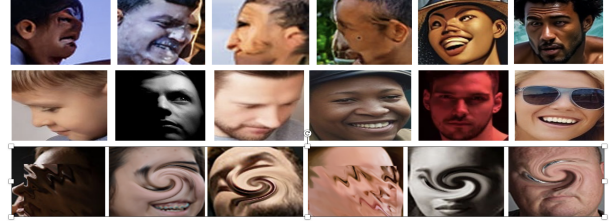


Fig. 1. First row : Examples of poor-quality crops generated by diffusion models. These are labeled as ‘bad face’ crops. Second row : ‘good face’ crops from Adobe Stock and images generated by Firefly. These good-quality face samples contain low-light, side-face profile, different facial orientation and occlusions on face making it different from datasets having Face Recognition task in focus. Third row : Distortions (Twirl, Zig-Zag, Ripple, Shear, Wave) generated using Adobe Photoshop on ‘good face’ crops (from Adobe Stock and FFHQ).

ing good-quality faces for face recognition and related tasks, side profiles or low-light face crops may be labeled as poor-quality face examples. However, for prompt-based image generation, users may specifically ask to generate side-profile faces or extreme scenario faces, making these face crops essential. In [26], the authors introduced a large-scale dataset of demographically annotated AI-generated faces, including real faces, faces from deepfake videos, and faces generated by GANs and diffusion models. With our work, we introduce a new dataset called ‘Labeled Bad Faces’ (LBF) capturing different kinds of distortions in faces generated by various state-of-the-art diffusion models.

3.2. Dataset Source

We collected two classes of face crops labeled as ‘bad faces’ and ‘good faces’. For ‘good faces’, we filtered approximately 60,000 samples sourced from Adobe Stock using labels ‘people’ and ‘human’, and extracted close to 15,000 samples from FFHQ dataset[27]. To cover extreme scenarios such as side-profile and low-light conditions, we also filtered an additional 10,000 samples sourced from Adobe Stock labeled specifically as ‘side faces’ and ‘low light faces’. This brought the total sample of ‘good faces’ to 85,000. Collecting ‘bad face’ samples was a critical task, requiring human annotators to invest significant time and resources in accurately labeling bad crops. We have employed three methods to collect ‘bad face’ samples:

1. We applied a synthetic data generation strategy to distort ‘good faces’ sourced from Adobe Stock and FFHQ dataset. We altered 45,000 ‘good face’ crops using different image distortion techniques such as Twirl, Zig-Zag, Ripple, Shear and Wave. These distortions were applied randomly to 30-70% area in the face crops, as

shown in Fig. 1.

2. We generated 15,000 samples of 'bad faces' by utilizing a custom face generation GAN trained on FFHQ face dataset with low number of epochs. Extracting outputs from different stages of training this GAN by early stopping produced the required low-quality faces.
3. We observed that although we can capture most of the high-frequency distortions with synthetic generation and GANs, a lot of low-frequency distortions cannot be captured by above strategies. To capture them, we have manually annotated and verified close to 30,000 samples of face crops generated by diffusion models such as Firefly, StableDiffusionXL[28], StableDiffusion 2.1, StableDiffusion 1.5, Bytedance SDXL and Segmind SSD-1b [29] trained on LORA optimization to extend bad textures of the face crops.

To ensure accurate face-cropping, we used RetinaFace [9] based on the Resnet50 architecture with a threshold of 0.7. This relatively low threshold was chosen to account for the extreme distortions present in some faces generated by diffusion models. We have also discarded blurred faces by filtering with the variance of the Laplacian, and excluded False Positives, such as crops having animals or other non human-face entities, using object detection models. To the best of our knowledge, no such dataset is available publicly which consists of a large number of human annotated bad-quality faces. With this work, we release a carefully-curated dataset of 30,000 human-annotated bad-quality faces (LBF).

3.3. Test Dataset

To test our model, we prepared a list of 189 generated images by each of Stable-Diffusion 3, Dall-E 3, FLux-Dev, Flux-schnell and Firefly, totaling to 945. These images were generated using prompts carefully designed to account for geographic, ethnic, racial, gender, and environmental diversity. This approach ensures comprehensive testing of the robustness of our model. These prompts were crafted using prompt engineering with ChatGPT, and the same prompt can be fed to GPT to scale to a larger number of prompts if needed. Using this initial set of prompts, we generated close to 2,000 face crop samples to test our model.

4. ARCHITECTURE DETAILS

To benchmark the results, we train a vanilla Resnet34 architecture to validate the initial results on our test dataset. We measure the accuracy for model trained on dataset without synthetically distorted samples and with synthetically distorted samples. We also fine-tune

1. EfficientNetB1 architecture,
2. ViTB16 to capture global patterns in the image,

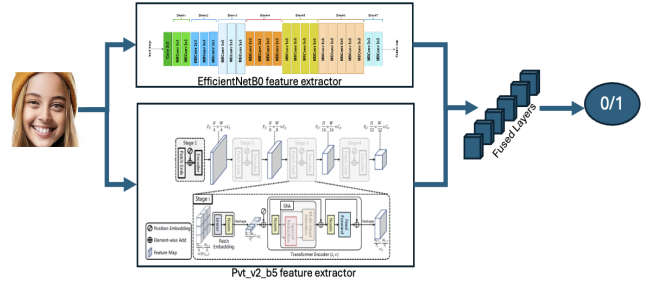


Fig. 2. Hybrid Pyramid-ViT and Efficient-net architecture with best results on most of diffusion models.

3. Pyramid ViT v2 with added design, including (a) linear complexity attention layer, (b) overlapping patch embedding, and (c) convolutional feed-forward network.

We introduce a new model architecture Face Quality Transformer (FaceQ Transformer, Fig. 2) which comprises of Pyramid ViT [30, 31] facial feature extractor fused with EfficientNet based local feature extractor to capture local and global distortion patterns present in the face crops. We show that this architecture outperforms all the previous methods and architectures. The illustrated results on various state-of-the-art diffusion model by our models are depicted in Table 1. Vanilla Resnet Shown in this table is trained without using 'bad faces' created by image distortions (Twirl, Zig-Zag, Wave, Ripple etc.). We observe that using the image distortions improves the result by 2% on our test dataset.

5. RESULT AND DISCUSSION

Table 1. F1 Score of FaceQ Transformer On Test dataset

Model	SD3	Dalle	Flux-s	Flux-d	Firefly	All
FaceQNet	0.54	0.64	0.76	0.85	0.87	0.73
FaceQan	0.63	0.70	0.71	0.75	0.77	0.71
CLIB-FIQA	0.56	0.70	0.79	0.92	0.89	0.77
InternVL-2.5 (26B)	0.64	0.65	0.64	0.65	0.79	0.69
Ours (Vanilla Resnet34)	0.82	0.86	0.92	0.92	0.80	0.86
Ours (Resnet34)	0.83	0.85	0.89	0.90	0.86	0.88
Ours (EfficientNetB1)	0.84	0.89	0.93	0.91	0.90	0.90
Ours (ViT B16)	0.86	0.90	0.88	0.89	0.87	0.88
Ours (PyramidViT)	0.86	0.87	0.91	0.92	0.84	0.89
Ours (FaceQTransformer)	0.90	0.91	0.91	0.93	0.88	0.91

5.1. Model performance

The plot in Fig. 5 represent the scores given by our FaceQ Transformer for 5 different state-of-the-art text-to-image dif-

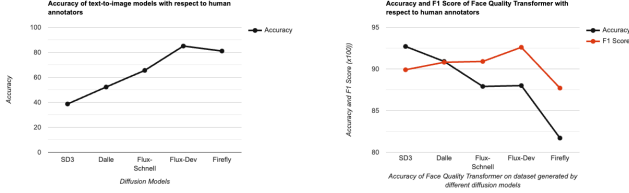


Fig. 3. (a) First figure (left) shows the accuracy of text-to-image diffusion models in generating high-quality faces (Flux-dev and Firefly generate more than 80% of 'good face' crops). (b) Accuracy and F1 Score of Face Quality Transformer Model on images generated by different diffusion models.

fusion models. These results closely align with human labeling since we observe that the quality of labels annotated by a team of three human annotators match with that of the result by our model. Fig. 3a Shows the percentage of good-quality face crops generated by various diffusion models. Fig. 3b shows the accuracy and F1-score of our pipeline considering human annotations as the ground truth. We achieve greater than 88% accuracy, 0.83 F1-score and greater than 0.91 AUC ROC for all the individual state-of-the-art text-to-image models. In our results we show that Firefly and Flux-dev are of superior quality for generating high-quality faces having accuracy of over 80%.

Some faces in images generated by diffusion models are ambiguous, making it difficult for even a team of three humans to determine their quality. We have omitted these highly ambiguous face-crops (less than 3% of the overall crops) for the face-quality models evaluation. We also observe that the texture of images generated by Dall-E 3 is very different from that of the other text-to-image models as Dall-E 3 generate more artistic faces making it hard to evaluate.

Some limitations of our model include estimating quality for very low-light or invisible face and heavily occluded face (more than 60-70% of occlusion). We observe that our model struggles to classify these cases with high accuracy.

5.2. Comparison with other models

To show the effectiveness of our approach we compare our pipeline with other state-of-the-art models for face quality estimation. We compare the results of FaceQNet[32], FaceQan[6] and CLIB-FIQA[33]. Although these models are mostly tailored specific to face recognition and similar utilities of face quality estimation but we wanted to compare the results to verify that these models are performing good for distorted and straight faces. We also show the face quality assessment by InternVL-2.5 26B parameter model using visual question answering. We show that our approach is best for classifying the quality of bad faces generated by

diffusion models. We also show the comparative results depicting F1-score of different face-quality estimation models on test dataset described in section 3.3. We have also fine-tune different different pre-trained models like efficient-net, ViT-B16, Pyramid-ViT v2 on our training dataset consisting of 'good quality' and 'bad quality' faces. We show that our proposed Face Quality Transformer (Fig 2) model consisting of a fusion of Pyramid-ViT architecture and efficient-net model outperforms all the previous models and is more robust across text-to-image diffusion models.

6. ABLATION STUDY

6.1. Improving the existing DDPM model with Face classification network

We demonstrate the effectiveness of our approach by feeding the face quality estimation network as a loss function to a face generation diffusion model with a simple attention-based Unet architecture. We observe that using our face classification network as a feature loss with DDPM for face generation improves the face generation diffusion model by 6% on a sample set of 512 generated face crops. We setup a diffusion network for unconditional face generation from the 'good quality' faces available with us (close to 85,000). We first train a model with simple MSE loss, and for the second experiment, we feed our face quality estimation classifier as a loss function to the DDPM model. The updated loss function is shown below. To evaluate the quality of faces generated by diffusion model we generate random 512 images by DDPM trained with MSE loss and DDPM trained with face quality loss. We show that DDPM with face quality loss has significant improvement in the quality of generated faces Fig.4 (left vs right).

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda \cdot \mathcal{L}_{\text{feature}} \quad (1)$$

where:

- $\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$ is the Mean Squared Error loss,
- $\mathcal{L}_{\text{feature}}$ is the Face Quality Network based feature loss (we feed Face Quality Network with Resnet34 architecture to extract face quality features),
- The weighting parameter λ has been experimentally set to 0.1 for optimal performance.

The feature loss $\mathcal{L}_{\text{feature}}$ is defined as:

$$\mathcal{L}_{\text{feature}} = \frac{1}{N} \sum_{i=1}^N \|\phi(y_i) - \phi(\hat{y}_i)\|_2^2 \quad (2)$$

where: $\phi(\cdot)$ represents the feature extraction function of a Face Quality Estimation network, and $\|\cdot\|_2^2$ denotes the squared ℓ_2 -norm (Euclidean distance) between the feature representations.

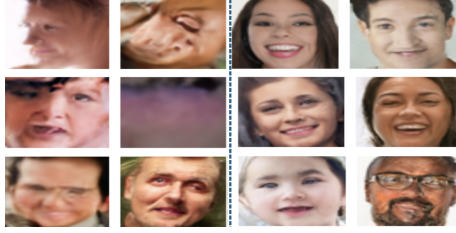


Fig. 4. Sample Distribution of faces generated by Vanilla face Diffusion Model with MSE loss (Left) and with face quality loss (Right). Observe the quality difference separated by dotted line.

6.2. Correction Pipeline and User acceptance

We conducted an additional experiment to iteratively re-generate images k number of times if the images generated were of poor-quality. This experiment involved 100 prompts for image generation and 3 annotators to annotate the quality of images based on faces-generated. We observed an improvement of the perceived quality of generated images by 11.7% for second automated re-generation ($k=2$) and 14% by third automated re-generation ($k=3$). This experiment was done for Firefly images where the initial acceptance of images by annotators was approximately 80%.

7. CONCLUSION

In this work we propose Face Quality Transformer, an approach to address one of the most critical problems to classify the poor-quality of faces generated by text-to-image diffusion models. We demonstrate that our algorithm is able to detect poor-quality faces across different open and proprietary text-to-image diffusion models. We also propose a dataset "Labeled Bad Faces" which will help advance the research in classifying the bad quality of faces generated by diffusion and non-diffusion models having structurally distorted faces. We show that this face quality estimation network can be plugged as a loss function to improve existing face generation models. We also highlight the effectiveness of our approach by improving the quality of bad faces by 14% in just two more re-generation steps. With our work, we can also objectively rank various text-to-image models on their ability to generate good-quality images containing high-quality faces.

8. REFERENCES

- [1] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong, "Imagereward: Learning and evaluating human preferences for text-to-image generation," 2023.
- [2] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li, "Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis," *arXiv preprint arXiv:2306.09341*, 2023.
- [3] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, "Improved techniques for training gans," 2016.
- [4] 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," 2018.
- [5] Wenquan Lu, Yufei Xu, Jing Zhang, Chaoyue Wang, and Dacheng Tao, "Handrefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting," *arXiv preprint arXiv:2311.17957*, 2023.
- [6] Žiga Babnik, Peter Peer, and Vitomir Štruc, "FaceQAN: Face Image Quality Assessment Through Adversarial Noise Exploration," *Proceedings of the IAPR International Conference on Pattern Recognition (ICPR)*, pp. 234–778, 2022.
- [7] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang, "SDD-FIQA: Unsupervised face image quality assessment with similarity distribution distance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [8] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper, "SER-FIQ: unsupervised estimation of face image quality based on stochastic embedding robustness," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020, pp. 5650–5659, IEEE.
- [9] Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen, "Giga: Generated image quality assessment," *arXiv preprint arXiv:2003.08932*, 2020.
- [10] Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger, "An empirical study on evaluation metrics of generative adversarial networks," 2018.
- [11] Žiga Babnik, Peter Peer, and Vitomir Štruc, "DiffIQA: Face Image Quality Assessment Using Denoising Diffusion Probabilistic Models," in *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, 2023.
- [12] Ali Borji, "Pros and cons of gan evaluation measures: New developments," 2021.

- [13] Fadi Boutros, Meiling Fang, Marcel Klemmt, Biying Fu, and Naser Damer, “Cr-fiq: Face image quality assessment by learning sample relative classifiability,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 5836–5845.
- [14] Y. Zhao, T. Hou, Y. Su, X. Jia, Y. Li, and M. Grundmann, “Towards authentic face restoration with iterative diffusion models and beyond,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Los Alamitos, CA, USA, oct 2023, pp. 7278–7288, IEEE Computer Society.
- [15] Yun Pang, Jiawei Mao, Libo He, Hong Lin, and Zhenping Qiang, “An improved face image restoration method based on denoising diffusion probabilistic models,” *IEEE Access*, vol. 12, pp. 3581–3596, 2024.
- [16] Michail Tarasiou, Stylianos Moschoglou, Jiankang Deng, and Stefanos Zafeiriou, “Improving face generation quality and prompt following with synthetic captions,” 2024.
- [17] Zhenyi Liao, Qingsong Xie, Chen Chen, and Haonan Lu, “Fine-tuning diffusion models for enhancing face quality in text-to-image generation,” 2024.
- [18] Ali Borji, “Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2,” 2023.
- [19] Jonathan Ho, “Classifier-free diffusion guidance,” *ArXiv*, vol. abs/2207.12598, 2022.
- [20] Zhihao Chen, Bin Hu, Chuang Niu, Yuxin Li, Hongming Shan, and Ge Wang, “Iqagpt: Image quality assessment with vision-language and chatgpt models,” 2023.
- [21] Xiao Cui, Qi Sun, Wengang Zhou, and Houqiang Li, “Exploring GPT-4 vision for text-to-image synthesis evaluation,” in *The Second Tiny Papers Track at ICLR 2024*, 2024.
- [22] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold, “Gpt-4v(ision) as a generalist evaluator for vision-language tasks,” 2023.
- [23] 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,” Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.
- [24] C D Castillo V M Patel R Chellappa D W Jacobs S Sengupta, J C Cheng, “Frontal to profile face verification in the wild,” in *IEEE Conference on Applications of Computer Vision*, February 2016.
- [25] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn C. Adams, Tim Miller, Nathan D. Kalka, Anil K. Jain, James A. Duncan, Kristen E Allen, Jordan Cheney, and Patrick Grother, “Iarpa janus benchmark-b face dataset,” *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 592–600, 2017.
- [26] Li Lin, Santosh, Xin Wang, and Shu Hu, “Ai-face: A million-scale demographically annotated ai-generated face dataset and fairness benchmark,” *arXiv preprint arXiv:2406.00783*, 2024.
- [27] Timo Aila Tero Karras, Samuli Laine, “A style-based generator architecture for generative adversarial networks,” *IEEE[Online]*. Available: <https://ieeexplore.ieee.org/document/8953766>, vol. 3, 2019.
- [28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, “High-resolution image synthesis with latent diffusion models,” 2021.
- [29] Yatharth Gupta, Vishnu V. Jaddipal, Harish Prabhala, Sayak Paul, and Patrick Von Platen, “Progressive knowledge distillation of stable diffusion xl using layer level loss,” 2024.
- [30] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao, “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.
- [31] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao, “Pvtv2: Improved baselines with pyramid vision transformer,” *Computational Visual Media*, vol. 8, no. 3, pp. 1–10, 2022.
- [32] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay, “Faceqnet: Quality assessment for face recognition based on deep learning,” 2019.
- [33] Fu-Zhao Ou, Chongyi Li, Shiqi Wang, and Sam Kwong, “Clib-fiq: Face image quality assessment with confidence calibration,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 1694–1704.