VQFR: Blind Face Restoration with Vector-Quantized Dictionary and Parallel Decoder Supplementary Material

Yuchao Gu[1,2], Xintao Wang[2], Liangbin Xie[2,5], Chao Dong[4,5],
Gen Li[3], Ying Shan[2], and Ming-Ming Cheng[1] [0000−0001−5550−8758]

[1]TMCC, CS, Nankai University [2]ARC Lab, Tencent PCG
[3]Platform Technologies, Tencent Online Video [4]Shanghai AI Laboratory
[5]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
https://github.com/TencentARC/VQFR/

In this supplementary material, we first present the **architecture details of VQFR** in Sec. 1. Then we give more details about **evaluation metrics** in Sec. 2. The **limitations** of VQFR are discussed in Sec. 3. We then provide more **visual comparisons of ablation studies** to help better understand the VQFR designs in Sec. 4. More **visual comparisons with previous methods on the real-world datasets** are shown in Sec. 5.

1 Network Architectures

VQFR: The detailed architecture of VQFR is illustrated in Table. 1. There are six resolution levels, *i.e.*, $f = \{1, 2, 4, 8, 16, 32\}$, and the quantization operation is conducted on the feature level of $f32$. Each level of the encoder contains two residual blocks, and each level of the texture branch in the decoder contains three residual blocks. Each level of the main branch in the decoder contains one texture warping module and one residual block. We use a bilinear upsample/downsample followed by a 1×1 convolution to change the resolutions. VQFR has 76.3M params (1.07 TFlops) and takes 0.36s to process a 512^2 image on Nvidia A100.

Texture Warping Module (TWM): We use a 3×3 convolution with 32 output channels to extract input information of degraded faces. Then we resize the feature to match all resolution levels ($f = 1, 2, 4, 8, 16, 32$). The detailed architecture of TWM is shown in Table. 2. For each resolution level, the offset convolution is used to generate offsets from the concatenation of the texture feature and the input features of degraded faces. Then, the offsets and the texture features are fed into the deformable convolution, outputting the warped feature.

2 Evaluation Metrics

Our evaluation metrics contain two widely-used non-reference perceptual metrics: FID [7] and NIQE [11]. We also measure the pixel-wise metrics (PSNR and SSIM) and perceptual metric (LPIPS [16]) for benchmarking CelebA-Test with Ground-Truth (GT). However, as pointed out in [1], the distortion measure (*e.g.*, PSNR, SSIM) and perceptual quality are at odds with each other. Similar to GFP-GAN [13], we pursue perceptual quality in VQFR and provide PSNR/SSIM for reference only. In the Table. 1 of the

Table 1: The detailed architecture of VQFR. The residual block consists of 3×3 Conv-GN [15]-Swish [12]-3×3 Conv-GN-Swish. g: groups in GroupNorm (GN); c: channels; dg: deformable groups in deformable convolution [17]; f: compression patch size.

Input size	Encoder	Texture branch	Main branch
$f1: 512\times512$	{ Residual block, 128-c, 32-g } × 2 Bilinear downsampling 2× Conv 1 × 1, 128-c	{ Residual block, 128-c, 32-g } × 3	TWM, 128-c, 4-dg {Residual block, 128-c, 32-g} × 1
$f2: 256\times256$	{ Residual block, 128-c, 32-g } × 2 Bilinear downsampling 2× Conv 1 × 1, 128-c → 256-c	{ Residual block, 128-c, 32-g } × 3 Bilinear upsampling 2× Conv 1 × 1, 128-c	TWM, 128-c, 4-dg {Residual block, 128-c, 32-g} × 1
$f4: 128\times128$	{ Residual block, 256-c, 32-g } × 2 Bilinear downsampling 2× Conv 1 × 1, 256-c	{ Residual block, 256-c, 32-g } × 3 Bilinear upsampling 2× Conv 1 × 1, 256-c → 128-c	TWM, 256-c, 4-dg {Residual block, 256-c, 32-g} × 1
$f8: 64\times64$	{ Residual block, 256-c, 32-g } × 2 Bilinear downsampling 2× Conv 1 × 1, 256-c	{ Residual block, 256-c, 32-g } × 3 Bilinear upsampling 2× Conv 1 × 1, 256-c	TWM, 256-c, 4-dg {Residual block, 256-c, 32-g} × 1
$f16: 32\times32$	{ Residual block, 256-c, 32-g } × 2 Bilinear downsampling 2× Conv 1 × 1, 256-c → 512-c	{ Residual block, 256-c, 32-g } × 3 Bilinear upsampling 2× Conv 1 × 1, 256-c	TWM, 256-c, 4-dg {Residual block, 256-c, 32-g} × 1
$f32: 16\times16$	{ Residual block, 512-c, 32-g } × 2	{ Residual block, 512-c, 32-g } × 3 Bilinear upsampling 2× Conv 1 × 1, 512-c → 256-c	TWM, 512-c, 4-dg {Residual block, 512-c, 32-g} × 1

Table 2: The detailed architecture of the texture warping module (TWM). OConv: convolution for generating offsets; DConv: deformable convolution; c: channels; dg: deformable groups.

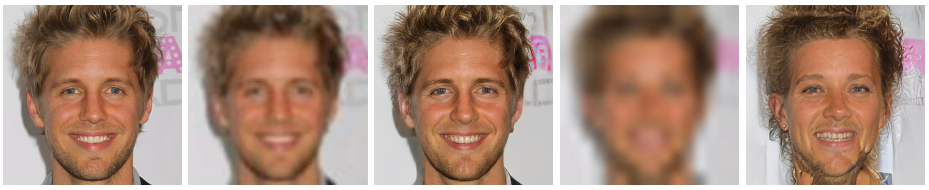
Input size	$f1: 512\times512$	$f2: 256\times256$	$f4: 128\times128$
TWM	OConv: { Conv 1×1, (128+32)-c → 128-c Depthwise Conv 7×7, 128-c Conv 1×1, 128-c } DConv: { Deformable Conv 3×3, 128-c, 4-dg }	OConv: { Conv 1×1, (128+32)-c → 128-c Depthwise Conv 7×7, 128-c Conv 1×1, 256-c } DConv: { Deformable Conv 3×3, 128-c, 4-dg }	OConv: { Conv 1×1, (256+32)-c → 256-c Depthwise Conv 7×7, 256-c Conv 1×1, 256-c } DConv: { Deformable Conv 3×3, 256-c, 4-dg }
Input size	$f8: 64\times64$	$f16: 32\times32$	$f32: 16\times16$
TWM	OConv: { Conv 1×1, (256+32)-c → 256-c Depthwise Conv 7×7, 256-c Conv 1×1, 256-c } DConv: { Deformable Conv 3×3, 256-c, 4-dg }	OConv: { Conv 1×1, (256+32)-c → 256-c Depthwise Conv 7×7, 256-c Conv 1×1, 256-c } DConv: { Deformable Conv 3×3, 256-c, 4-dg }	OConv: { Conv 1×1, (512+32)-c → 512-c Depthwise Conv 7×7, 512-c Conv 1×1, 512-c } DConv: { Deformable Conv 3×3, 512-c, 4-dg }

main manuscript, the best PSNR and SSIM are achieved by degraded inputs, as all other methods are optimized for the perceptual quality instead of the distortion measures.

For fidelity measurement, we follow previous work [13] to use the embedding angle of ArcFace [3] as the identity metric, which is denoted by 'Deg.'. However, this Deg. metric actually cannot well reflect the fidelity due to the following reasons. 1) The ArcFace model downsamples the face images into 128×128 during inference, which loses the spatial dimension. Thus, it cannot evaluate detailed facial positions. 2) The ArcFace is designed for the recognition task and is trained with the invariance to expressions. While the expression is important for measuring fidelity in face restoration. In order to better measure the fidelity with accurate detailed facial positions and expressions, we further adopt landmark distance (LMD) as the fidelity metric. Specially, we use AWing [14] to obtain 98 landmarks for both the restored face and the ground-truth face.

Input VQFR Input VQFR

(a) Results on faces of extremely poses.



GT Common Case Extreme Case

(b) Results on extremely less informative faces.

Fig. 1: Limitations of VQFR.

Then we calculate the L2 distance for each landmark and average the distance as the final score of the LMD metric.

3 Limitation

The limitations of VQFR are two-folds. 1) As shown in Fig. 1(a), faces in extreme poses lead to poor restoration results, since the codebook is built from the training dataset, in which most samples are frontal faces. One potential solution is to increase the dataset diversity and codebook size, which will help build a more comprehensive dictionary. 2) As shown in Fig. 1(b), the restoration from extremely less informative faces is far from satisfactory, since the VQFR does not build upon a generative model. Moreover, VQ may further lead to divergent codebook quantization due to the less informative inputs. One promising direction to improve is to equip VQFR with generation ability. For example, when the input faces contain extreme low-information and thus the code mapping is ambiguous, the auto-regressive [5] or bi-directional [4] transformer can help model the code selection.

4 More Visualizations of Ablation Study

Importance of input features of degraded faces. Input features of degraded faces play an important role in preserving fidelity. We compare the SimVQFR without input features and our VQFR with input features. As shown in Fig. 2, with the input features of degraded faces, VQFR could generate more faithful expressions (the first and fourth

row), more faithful facial lines (the second and forth row), and facial components (the third row) than SimVQFR. The facial lines and components can be roughly recovered from the input LQ faces but can be easily changed by the discrete quantization, thus influencing the final recovered expressions and identity. Our VQFR incorporates the input features from degraded faces at different spatial levels and preserve better fidelity.

Importance of the parallel decoder. The parallel decoder is the key design of VQFR to preserve high-quality facial details when fusing texture features of the VQ codebook and input features from degraded faces. We compare the variant-1 (single branch) and variant-2 (parallel decoder) in Fig. 3. With the parallel decoder, variant-2 could generate high-quality facial components (the first row), realistic hairs (the first row) and skins (the second row). In Fig. 4, we provide more visual examples to show the importance of the parallel decoder design in generating realistic skins (the first and second rows) high-quality hairs and eyes (the third and fourth rows).

Influence of dual discriminators. We adopt dual discriminators to remove the regular pattern when utilizing facial textures of the VQ codebook in VQFR. We adopt style-based wavelet-driven discriminator [6] as the global discriminator and adopt PatchGAN discriminator [8] as the local discriminator. We show the influence of dual discriminators in Fig. 5. When we only use the global discriminator (the second column), we can find that there are regular patterns on skin and hair. When adding the patch discriminator as the local discriminator, the regular patterns are removed (the third column).

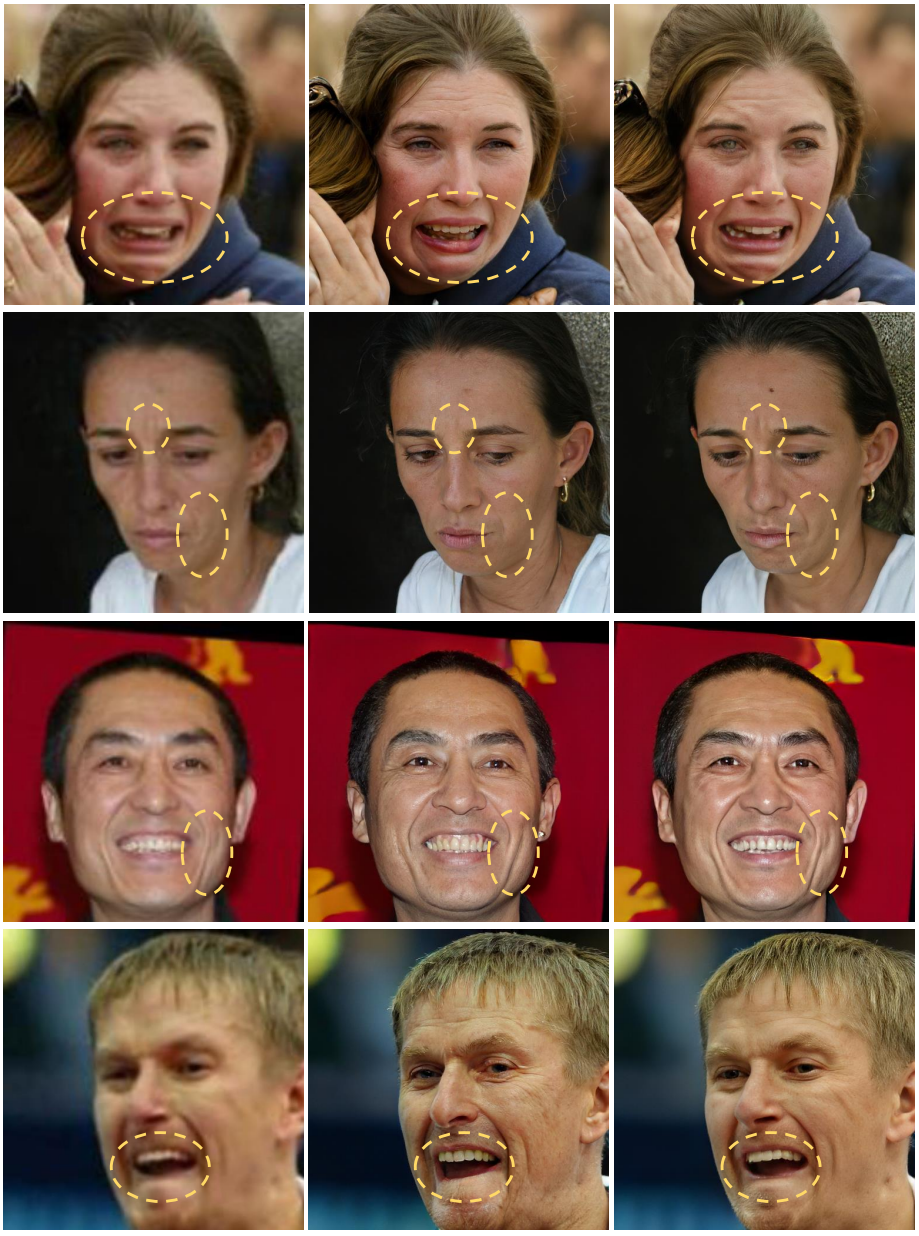


Fig. 2: Comparisons between the SimVQFR (without input features) and our VQFR (with input features). With input features of degraded faces, our VQFR could generate more faithful expressions (the first and third row), facial lines (the second and fourth row) and facial components (the fourth row) than SimVQFR.

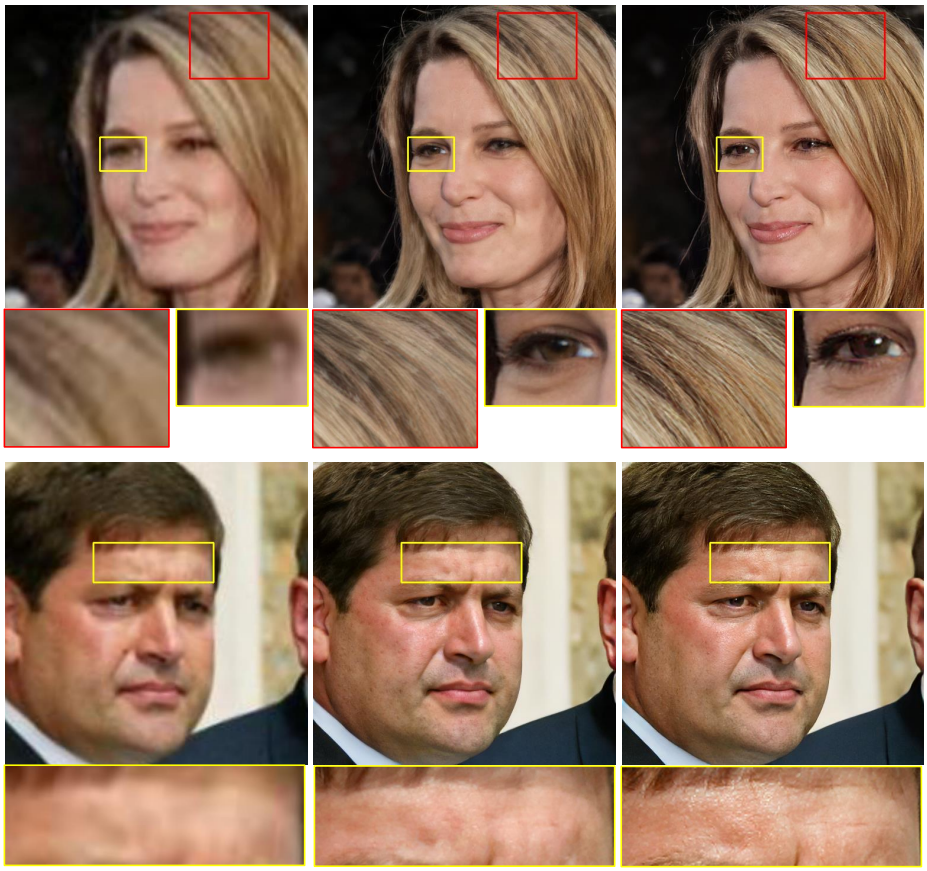


Fig. 3: Comparisons between the Variant-1 (single branch) and Variant-2 (parallel decoder). With the proposed parallel decoder, high-quality facial details from the VQ codebook could be preserved. Therefore, Variant-2 could generate better facial components (the first row), more realistic hairs (the first row) and skin (the second row) than Variant-1.

Input Variant-1 (Single Branch) Variant-2 (Parallel Decoder)

Fig. 4: Comparisons between the Variant-1 (single branch) and Variant-2 (parallel decoder). With the parallel decoder, Variant-2 could generate more realistic skins (the first and second rows), eyes (the third and fourth rows) and hairs (the third and fourth rows) than Variant-1.

Fig. 5: Influence of dual discriminators. With only the global discriminator (the second column), there are regular patterns on hairs and skins. When adding the patch discriminator as a local discriminator, the regular patterns are removed (the third column).

5 More Qualitative Results on Real-World Data

We show more qualitative results on the real-world dataset, *i.e.*, *LFW-Test*, *CelebChild* and *WebPhoto*. We compare our VQFR with several state-of-the-art face restoration methods: DFDNet [9], PSFRGAN [2], PULSE [10] and GFPGAN [13].

The qualitative comparisons on the *WebPhoto* are shown in Fig. 6, Fig. 7 and Fig. 8. The qualitative comparisons on the *CelebChild* are present in Fig. 9 and Fig. 10. Moreover, qualitative comparisons on *LFW-Test* are shown in Fig. 11, Fig. 12 and Fig. 13. Our VQFR produces high-quality facial components and more realistic hairs and skins than previous methods.

Fig. 6: Qualitative comparison on the real-world *WebPhoto* dataset. Our VQFR could restore more realistic facial components (eyes and ears) than previous methods. (**Zoom in for best view**).

Fig. 7: Qualitative comparison on the real-world *WebPhoto* dataset. Our VQFR could restore more realistic facial components (eyes and ears) and more realistic skins than previous methods. (**Zoom in for best view**).

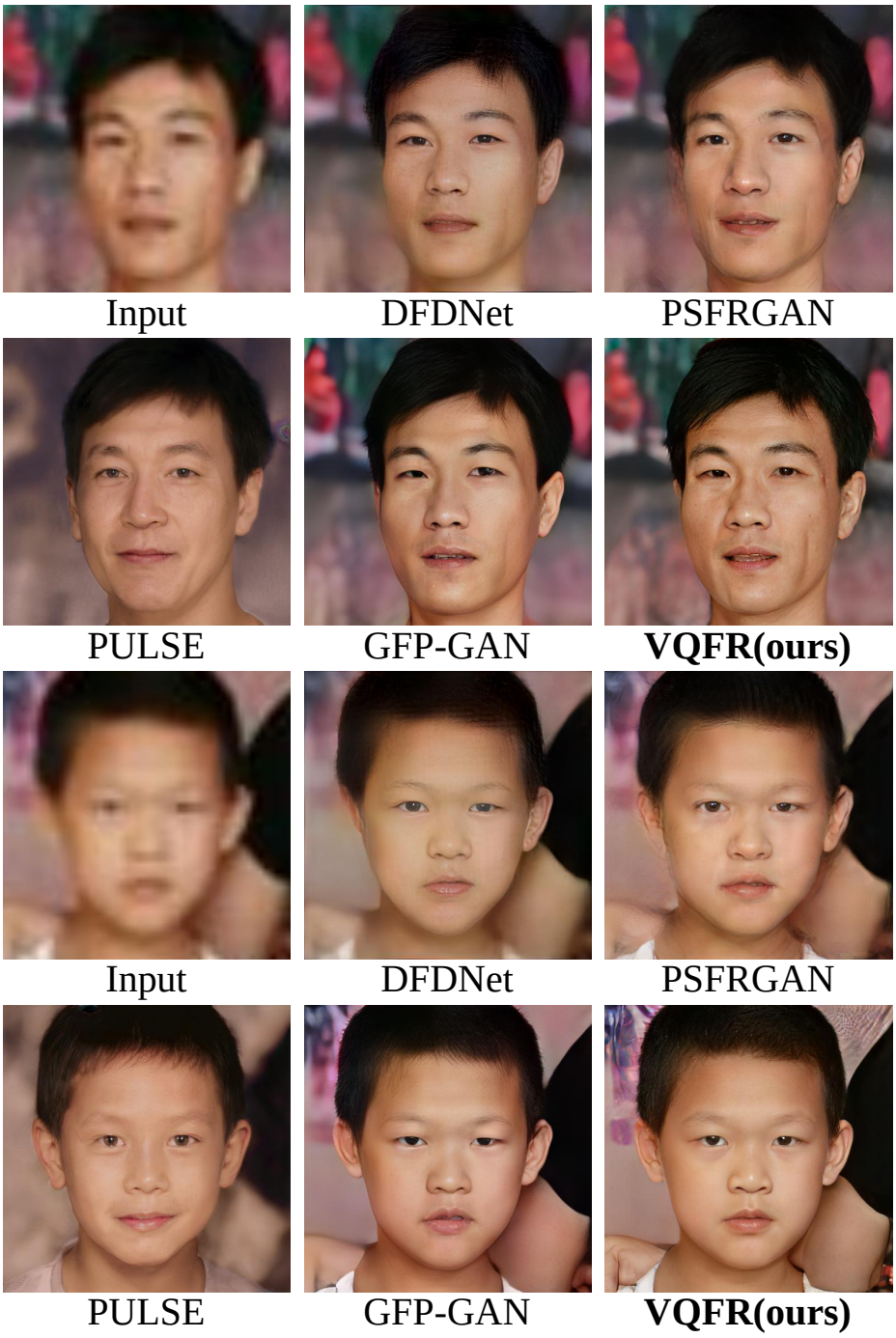


Fig. 8: Qualitative comparison on the real-world *WebPhoto* dataset. Our VQFR could restore more realistic facial components (eyes and ears) and more realistic skins and hairs than previous methods. (**Zoom in for best view**).

Fig. 9: Qualitative comparison on the real-world *Celeb-Child* dataset. Our VQFR could restore more realistic eyes and hairs than previous methods. (**Zoom in for best view**).

Fig. 10: Qualitative comparison on the real-world *Celeb-Child* dataset. Our VQFR could restore more realistic eyes and hairs than previous methods. (**Zoom in for best view**).

Fig. 11: Qualitative comparison on the real-world *LFW-Test* dataset. Our VQFR could restore high-quality facial components (eyes and hairs) and more realistic skins than previous methods. (**Zoom in for best view**).

Fig. 12: Qualitative comparison on the real-world *LFW-Test* dataset. Our VQFR could restore high-quality facial components (eyes) and more realistic skins than previous methods. (**Zoom in for best view**).

Fig. 13: Qualitative comparison on the real-world *LFW-Test* dataset. Our VQFR could restore high-quality facial components (eyes) and more realistic skins and hairs than previous methods. (**Zoom in for best view**).

References

1. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6228–6237 (2018) 1
2. Chen, C., Li, X., Yang, L., Lin, X., Zhang, L., Wong, K.Y.K.: Progressive semantic-aware style transformation for blind face restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11896–11905 (2021) 9
3. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019) 2
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 3
5. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12873–12883 (2021) 3
6. Gal, R., Hochberg, D.C., Bermano, A., Cohen-Or, D.: Swagan: A style-based wavelet-driven generative model. ACM Transactions on Graphics (TOG) **40**(4), 1–11 (2021) 4
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017) 1
8. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017) 4
9. Li, X., Chen, C., Zhou, S., Lin, X., Zuo, W., Zhang, L.: Blind face restoration via deep multi-scale component dictionaries. In: European Conference on Computer Vision. pp. 399–415. Springer (2020) 9
10. Menon, S., Damian, A., Hu, S., Ravi, N., Rudin, C.: Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 2437–2445 (2020) 9
11. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012) 1
12. Ramachandran, P., Zoph, B., Le, Q.V.: Searching for activation functions. arXiv preprint arXiv:1710.05941 (2017) 2
13. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9168–9178 (2021) 1, 2, 9
14. Wang, X., Bo, L., Fuxin, L.: Adaptive wing loss for robust face alignment via heatmap regression. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6971–6981 (2019) 2
15. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018) 2
16. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 1
17. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9308–9316 (2019) 2