frontiers | Frontiers in Signal Processing

TYPE Original Research
PUBLISHED 02 May 2023
DOI 10.3389/frsip.2023.1106465

OPEN ACCESS

EDITED BY
Mohamed Abdel-Nasser,
Aswan University, Egypt

REVIEWED BY
Adel Saleh Ali Raimi,
Gaist Solutions Ltd., United Kingdom
Vivek Kumar Singh,
Queen's University Belfast,
United Kingdom

*CORRESPONDENCE
Ahmed Cheikh Sidiya,
ac0004@mix.wvu.edu

†PRESENT ADDRESS
Xuan Xu,
KLA Corporation, Milpitas, CA,
United States
Ning Xu,
Adeia Inc., San Jose, CA, United States

SPECIALTY SECTION
This article was submitted to Image Processing,
a section of the journal
Frontiers in Signal Processing

RECEIVED 02 December 2022
ACCEPTED 13 February 2023
PUBLISHED 02 May 2023

CITATION
Cheikh Sidiya A, Xu X, Xu N and Li X (2023), Degradation learning and Skip-Transformer for blind face restoration. *Front. Sig. Proc.* 3:1106465.
doi: 10.3389/frsip.2023.1106465

Degradation learning and Skip-Transformer for blind face restoration

Ahmed Cheikh Sidiya[1]*, Xuan Xu[2†], Ning Xu[2†] and Xin Li[1]

[1]West Virginia University, Lane Department of Computer Science and Electrical Engineering, Morgantown, United States, [2]Kwai Inc, Palo Alto, CA, United States

Blindrestoration of low-quality faces in the real world has advanced rapidly in recent years. The rich and diverse priors encapsulated by pre-trained face GAN have demonstrated their effectiveness in reconstructing high-quality faces from low-quality observations in the real world. However, the modeling of degradation in real-world face images remains poorly understood, affecting the property of *generalization* of existing methods. Inspired by the success of pre-trained models and transformers in recent years, we propose to solve the problem of blind restoration by jointly exploiting their power for degradation and prior learning, respectively. On the one hand, we train a two-generator architecture for degradation learning to transfer the style of low-quality real-world faces to the high-resolution output of pre-trained StyleGAN. On the other hand, we present a hybrid architecture, called Skip-Transformer (ST), which combines transformer encoder modules with a pre-trained StyleGAN-based decoder using skip layers. Such a hybrid design is innovative in that it represents the first attempt to jointly exploit the global attention mechanism of the transformer and pre-trained StyleGAN-based generative facial priors. We have compared our DL-ST model with the latest three benchmarks for blind image restoration (DFDNet, PSFRGAN, and GFP-GAN). Our experimental results have shown that this work outperforms all other competing methods, both subjectively and objectively (as measured by the Fréchet Inception Distance and NIQE metrics).

KEYWORDS

blind face restoration, degradation learning (DL), Skip-Transformer (ST), hybrid architecture design, face in the wild

1 Introduction

Blind face restoration is important for various vision applications, from face recognition in the wild (Ge et al., 2018) to the generation of 3D avatars (Ichim et al., 2015). Current state-of-the-art generative models for face images, such as StyleGAN (Karras et al., 2018) and its enhanced version (Karras et al., 2020) can produce synthetic images that are almost indistinguishable from real ones (except for certain artifact problems). The power of StyleGAN models has inspired the development of new face super-resolution (SR) techniques to take advantage of the prior information embedded into pre-trained generative models. One such method [e.g., GLEAN (Chan et al., 2021)] is capable of achieving nearly perfect SR results when trained and tested on artificial low-resolution images such as those generated with bicubic downsampling. However, the performance of GLEAN rapidly degrades when tested on low-quality real-world face images.

Blind restoration of low-quality (LQ) face images from the real world has remained an open problem due to the following challenges. First, the unknown degradation process in the

real world can be complicated and diverse, from blur kernels to compression artifacts. It is often difficult, if not impossible, to take into account various uncertainty factors of image degradation for likelihood modeling. Second, the effectiveness of the image prior, regardless of semantic-aware style transfer as in PSFRGAN (Chen C. et al., 2021) or a pre-trained face GAN as in GFP-GAN (Wang X. et al., 2021), is highly dependent on the choice of latent space for the disentangled representation of face information. When either likelihood or the previous model falls short, the reconstructed face images suffer from loss of facial details or undesirable artifacts.

The motivation behind our proposed solution is two-fold. On the one hand, we take the lesson from a previous work (Bulat et al., 2018) to carefully learn the degradation process from the real world. Unlike PSFRGAN (Chen C. et al., 2021) working with style transfer for the prior image, we argue that a more fruitful approach is to learn the real-world degradation process by style transfer. This unsupervised approach to degradation learning (DL) has been shown to be effective for both synthetic and real-world LQ face images. On the other hand, both GLEAN (Chan et al., 2021) and GFP-GAN (Wang X. et al., 2021) have used multiscale pre-trained GAN to facilitate the image reconstruction process. However, they differ in the way of latent space manipulation: GLEAN (Chan et al., 2021) uses a pre-trained generator as a latent bank, but requires a separate decoder to generate the output image; while GFP-GAN (Wang X. et al., 2021) introduces a spatial feature transform to modulate the extracted GAN features. Inspired by the complementary nature between GLEAN (Chan et al., 2021) and GFP-GAN (Wang X. et al., 2021), we advocate a hybrid approach of combining a transformer-based encoder with a pre-trained GAN-based decoder, leading to a novel Skip-Transformer (ST) design.

In this work, we propose using a coupled StyleGAN-based generator to learn the real-world degradation model from real-world LQ face images. Once learned, our degradation is combined with that of GFP-GAN (Wang X. et al., 2021) to create a *second order* degradation model, that is, GFP-GAN (Wang X. et al., 2021) serves as a second order degradation, improving the generalization property of our DL based on style transfer. We have also developed a new hybrid architecture design, named the *skip-transformer* (ST), based on powerful transformer encoder modules. By plugging the pre-trained face GAN as decoder modules, we can seamlessly integrate the representation power of the transformers with generative face priors (Wang X. et al., 2021). Furthermore, we have developed a novel extension of the skip connection to the *skip layers*, which facilitates the information flow between transformer-based encoders and StyleGAN-based decoders. A summary of our technical contributions is listed below.

- **Degradation learning *via* style transfer.** We study how to train state-of-the-art methods (Chan et al., 2021; Wang X. et al., 2021) in our DL style transfer model. Our approach to DL exploits pre-trained models (i.e., StyleGAN) and style transfer in the latent space. To our knowledge, this work is the first approach to DL that has shown convincingly better performance than GAN-based (Bulat et al., 2018).
- **Hybrid architecture design.** Our ST network architecture combines transformer encoder modules with a pre-trained StyleGAN-based decoder using skip layers. Such a hybrid design is innovative in that it represents the first attempt to jointly exploit the global attention mechanism of the transformer and pre-trained StyleGAN-based generative facial priors.
- **Real-world blind face image restoration.** Extensive experimental results on real-world face datasets show that the proposed DL-ST method outperforms existing state-of-the-art blind restoration methods, including DFDNet (Li et al., 2020a), PSFRGAN (Chen C. et al., 2021) and GFP-GAN (Wang X. et al., 2021). The subjective evaluation of restored images is also convincingly in favor of our method.

2 Related work

2.1 Blind face restoration

Blind face restoration aims to recover high-quality face images from low-quality observations that suffer from various sources of degradation in the real world [e.g., noise (Anwar et al., 2017), blur (Shen et al., 2018; 2020), and compression artifacts (Yang et al., 2018)]. Early work assumes the availability of a high-quality reference [e.g., (Li et al., 2018)] or the feasibility of learning a degradation model [e.g., (Bulat et al., 2018)] to guide the image restoration process. These methods were further enhanced in (Li et al., 2020b) by using multiple-exemplar images and adaptive fusion of features from guidance and degraded images and in DFDNet (Li et al., 2020a) by learning facial component dictionaries. More recently, blind face restoration was formulated as a semantically guided generation problem in HiFaceGAN (Yang L. et al., 2020) and solved using a collaborative suppression and replenishment approach. The success of StyleGAN for the synthesis of facial images (Karras et al., 2020) has also inspired the development of GFP-GAN (Wang X. et al., 2021), which adopts a pre-trained GAN prior. Unlike GAN inversion methods that require image-specific optimization for inference, GFP-GAN can jointly restore facial details and enhance colors in a single forward pass. The most recent advances in blind image restoration include approaches based on building a 3D facial prior (Hu et al., 2021) and a restoration framework with memorized modification (RMM) (Li et al., 2021). However, those existing blind face restoration approaches face the fundamental barrier of domain shift, i.e., when the real-world degradation varies, the performance often degrades rapidly.

2.2 Face image super-resolution

A closely related problem to blind face restoration is single-image super-resolution (SISR) for face images. Early work such as FSRNet (Chen et al., 2018) has assumed that low-resolution (LR) face images are artificially generated by downsampling high-resolution (HR) face images. The key idea behind FSRNet is to use the geometry prior, such as facial landmark heatmaps and parsing maps, to superresolve LR face images without well-aligned requirement. To generate realistic faces, FSRNet has been extended to FSRGAN by incorporating adversarial loss. Inspired by the success of StyleGAN for the synthesis of face images (Karras et al., 2018; 2020), self-supervised photo-upsampling *via* latent space

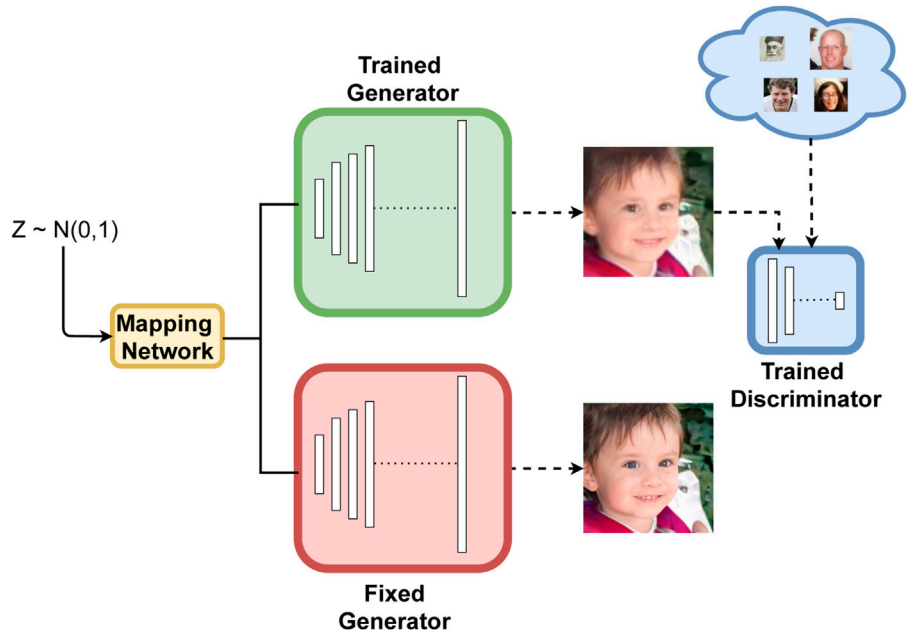

FIGURE 1
Overall architecture of the proposed DL networks. Red generator weights are kept fixed during training to generate the headquarters data. The output of the green generator is passed through a discriminator and compared with real-world LQ face images.

exploration of generative models (PULSE) was developed in (Menon et al. (2020) for SISR of large factor (up to × 64). Recognizing the weakness of estimating landmark and component maps from LR images, the FSR method based on iterative collaboration between two recurrent networks was proposed in Ma et al. (2020), focusing on facial image recovery and landmark estimation, respectively. In each step, the recovery branch uses the knowledge of the landmarks of the previous step to refine higher-quality images, which facilitates a more accurate landmark estimation in an iterative fashion. Most recently, the Generative Latent Bank (GLEAN) (Chan et al., 2021) went beyond current practices by directly leveraging rich and diverse priors encapsulated in a pre-trained GAN. By incorporating a simple encoder-bank-decoder architecture with multiresolution skip connections, GLEAN can easily handle images from diverse categories. Unfortunately, these existing face SR techniques mostly assume an over-simplified observation model for LR and cannot deliver satisfactory results on real-world scenarios.

2.3 Vision transformer

The field of vision transformers (Wang W. et al., 2021) has evolved rapidly in the last year. The great success of transformers in high-level vision, such as object detection [e.g., DETR (Carion et al., 2020)] and semantic segmentation [e.g., SWIN transformer (Liu et al., 2021)], has led to a flurry of low-level vision tasks, such as high-resolution image synthesis, for example, taming transformer (Esser et al., 2021), texture transformer (Yang F. et al., 2020), TransGAN (Jiang et al., 2021), and Ganformer (Hudson and Zitnick, 2021). So far, the study of vision transformers in image restoration has been scarce. The few exceptions in the open literature include the pre-trained image processing transformer (IPT) (Chen H. et al., 2021), the SWIN transformer for image restoration (SWINIR) (Liang et al., 2021) and the unpublished work of the U-shaped transformer (Uformer) (Wang Z. et al., 2021). Despite the promising experimental results reported on SISR and IPT image denoising, its network design consists of a standard transformer encoder/decoder pair originating from Vaswani et al. (2017). The design of novel transformer-based architectures to support low-level vision tasks such as blind image restoration serves as the primary motivation behind this work.

3 Methodology

3.1 Overview of the proposed approach

In the following sections, we first describe the general framework of our DL-ST method and then elaborate on the individual components. Given a low-quality (LQ) face image X suffering from unknown degradation, the goal is to restore an image Y with high quality (HQ). The proposed method consists of two basic modules: 1) DL: degradation learning *via* style transfer; and 2) ST: transformer-based image restoration with skip layers.

Our DL method is based on the training of a pair of pre-trained models [StyleGAN2 generators (Karras et al., 2020)] in a coupled manner to approximate real-world degradation. The network design is inspired by previous work on style transfer (Pinkney and Adler, 2020; Richardson et al., 2021) in latent space, but our targeted application is degradation learning instead of image manipulation (Wang et al., 2020). In fact, modeling image degradation in the real world as style transfer is supported by its application in SISR

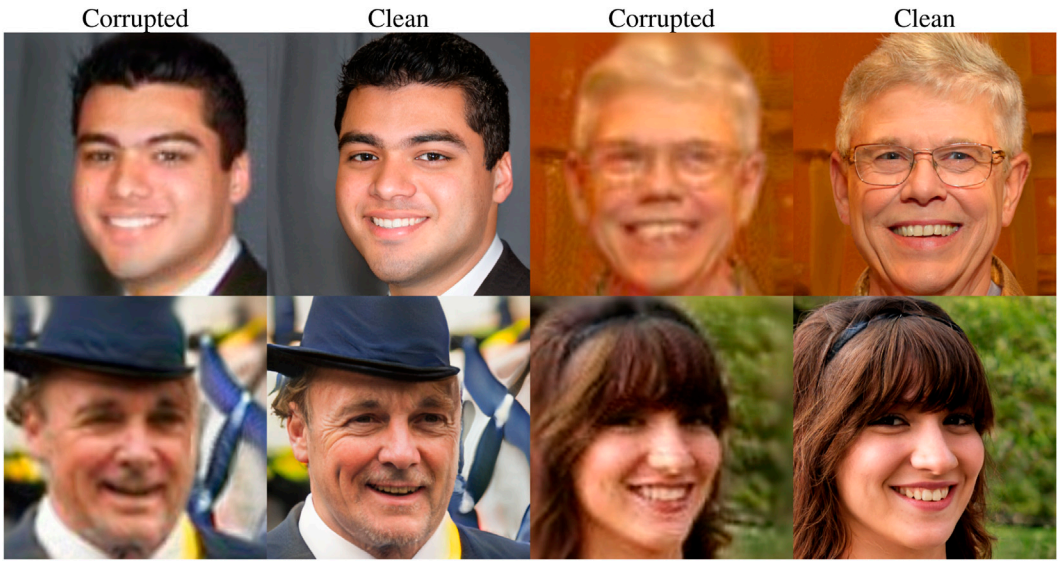

FIGURE 2
Examples of paired (Corrupted/Clean) randomly generated with the trained generators.

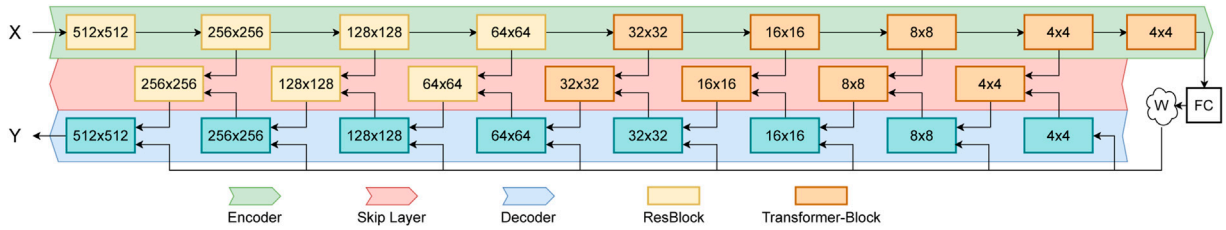

FIGURE 3
Overview of the proposed Skip-Transformer (ST) network architecture. The size numbers in the rectangle box stand for the output height and weight size of the corresponding boxes.

(Johnson et al., 2016). Unlike the Unet-based degradation removal module (Wang X. et al., 2021), both modules in our DL network (red and green generators in Figure 1) can be used to generate *unlimited* number of LQ and HQ face image pairs once trained. Compared to previous work on learning real-world degradation (Bulat et al., 2018), our approach is more powerful because it leverages the power of a style-based generator to represent more realistic image degradation.

Based on the learned degradation model, we will train a hybrid network architecture for face image restoration. The newly designed restoration network, called the Skip-Transformer (ST), consists of three parts: a transformer-based encoder, a skip layer, and a StyleGAN-based decoder. The encoder consists of transformer blocks for every resolution level between 512×512 and 4×4, as well as a fully connected layer that outputs the latent vector w. The skip layer is responsible for connecting the output of the encoder layer with that of the decoder. Finally, the decoder network borrows the architecture of the StyelGAN2 (Karras et al., 2020) generator as a pre-trained model. Although a similar idea of using a pre-trained GAN existed in Chan et al. (2021), (Wang X. et al. (2021), the combination with the transformer encoder and the introduction of skip layers differ our network design from others. To our knowledge, this hybrid design (transformer + StyleGAN) represents the first attempt to jointly exploit the global attention mechanism of transform-based and StyleGAN-based generative facial priors. Next, we will discuss these two modules in detail.

3.2 Degradation learning *via* style transfer

In blind face restoration, the challenge with DL lies in the lack of paired HQ/LQ images in the wild. To overcome this barrier, we propose here a new style transfer-based approach to DL that works for unpaired HQ/LQ images.

3.2.1 Real-world LQ data collection

To collect LQ images with real-world degradation, we used the Wider Face dataset (Yang et al., 2016) as a starting point. First, we crop and detect the face region using Zhang et al. (2017) and keep the cropped regions with a resolution threshold greater than or equal to 20×20. The cropped images are resized to 1024×1024 resolution

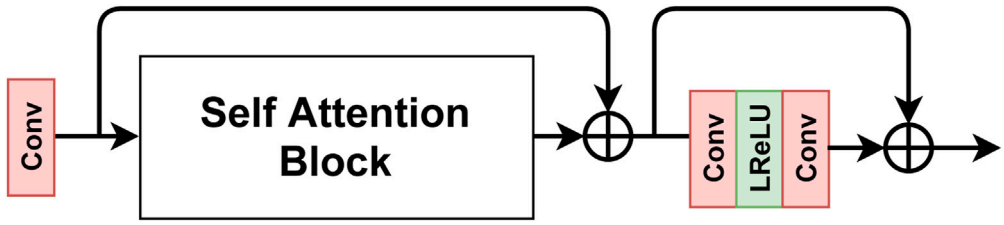

FIGURE 4
Architecture of the Transformer Block. For input resolution greater than 32 × 32, the self-atention block and the first residual skip connection are eliminated.

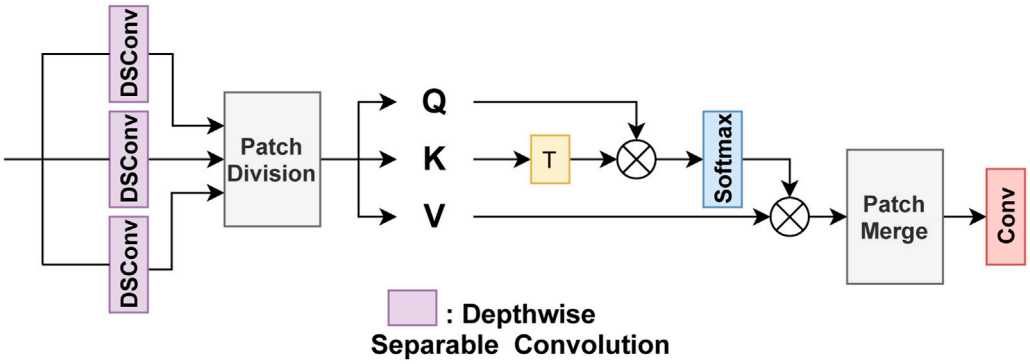

FIGURE 5
Architecture of the Self-Attention Block. The input is passed through three separable depth-wise convolution layers (Howard et al., 2017) to generate the features *Q*, *K*, and *V*.

using bicubic interpolation as the LQ input. Finally, all images are aligned using the facial landmark method of Kazemi and Sullivan (2014), which generates a total of 30,000 LQ images.

3.2.2 Degradation learning *via* style transfer

Our architecture for learning real-world face image degradation can be interpreted as a style transfer method between unpaired LQ and HQ face images (in our case, the HQ comes from the output of a pre-trained StyleGan generator). It consists of two generator networks (marked by the red and green colors in Figure 1) initialized with the weights of the pre-trained StyleGAN2 (Karras et al., 2020). The mapping network that projects the noise vector from the *Z* space to the *W* space is shared between the two networks and kept fixed [also initialized with the weights of the StyleGAN2 (Karras et al., 2020) mapping network]. The red generator that is responsible for outputting the high-resolution face image is also fixed during training. The green generator is responsible for the generation of the degraded face images. Two types of losses are applied during training: GAN loss is responsible for pushing its output to a style similar to the real-world LQ data (Yang et al., 2016), and the loss of content is used to keep the content similar to the output of the fixed red generator. Our framework shares similarities with previous work (Pinkney and Adler, 2020; Richardson et al., 2021), the main difference being that we make a conscious choice to avoid the use of projection algorithms in the latent space of StyleGAN2 (Karras et al., 2020) because they often distort the projected image and slow down the training.

3.2.3 Loss functions

During training, a random noise vector Z is sampled from the normal distribution and inputted into the mapping network before going through both generators. The mapping network and the red generator are kept fixed and the green generator is trained according to the objective function of (1); where $\hat{y}$ and y refer to the output of the green and red generators, respectively. It consists of L_2 loss, *VGG* loss, and adversarial loss. *VGG* loss is used to preserve the color information and the average downsampling is applied with a factor of 64 to its input. The weights α, β and γ are 1, 0.02 and 0.01, respectively. $\phi(\cdot)$ gives the output of the $18th$ feature layer of the pre-trained *VGG*19 (Simonyan and Zisserman, 2015) network, $\downarrow_s$ means the down-scale operator.

$$L_{DL} = \alpha L_{l_2} + \beta L_{VGG} + \gamma L_{adv}, \tag{1}$$

$$L_{l_2} = \|\hat{y} - y\|_2, \tag{2}$$

$$L_{VGG} = \|\phi(\hat{y}\downarrow_s) - \phi(y\downarrow_s)\|_2 \tag{3}$$

the introduction of L_{adv} is to transfer the low-resolution real-world face style to the output of the green generator. Similarly to Karras et al. (2020), the logistic loss is adopted:

$$L_{adv} = -\gamma \mathbb{E}_{\hat{y}} Softplus(D(\hat{y})) \tag{4}$$

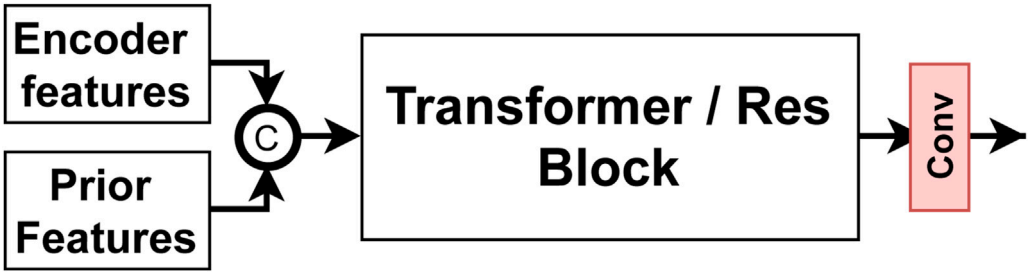

FIGURE 6
Design of the Skip Layer. The goal is to merge the features from both sources: Encoder and Decoder, by concatenation and passing the result through a transformer block similar to the one in Figure 4.

where $Softplus(x) = log(1 + exp(x))$ is a smooth approximation of the ReLU activation function.

The discriminator network is trained to distinguish between the output of the green generator and the Wider Face data (Yang et al., 2016). It has the same architecture and is initialized with the trained weights of StyleGAN2 (Karras et al., 2020). Following Mo et al. (2020), the early layers of the discriminator are frozen during training. The learning rate is 1×10^{-4} for both the green generator and the discriminator and is kept constant during training. We use Adam optimization with beta parameters equal to (0, 0.9). The architecture is trained for 90,000 iterations with a batch size of 12.

3.2.4 Second-order training procedure

Once trained, the green and red networks are used to produce paired clean and degraded face images. Similarly to the training procedure, a random vector is passed through the mapping network and both generators to output the pair of face images. Figure 2 shows examples of paired generated images. To further improve the generalization performance of our DL network, we propose the following second-order training procedure.

Our learned degradation is combined with the degradation of GFP-GAN (Wang X. et al., 2021) as a second-order refinement in the following way: we use 35,000 pairs of clean and degraded faces generated by trained green and red generators. Other high-resolution images 35,000 are randomly selected from the FFHQ dataset (Karras et al., 2018). The same degradation as in Wang X. et al. (2021) is applied to both 35,000 degraded faces and the HQ FFHQ dataset. In total, we have 70,000 training pairs of images that cover both synthetic and real-world LQ images. The inclusion of FFHQ high-resolution data enables us to have HQ training face images with more diverse poses and fewer artifacts than the ones generated by pre-trained models. Meanwhile, the use of GFP-GAN degradation as a second-order degradation in generated LQ images covers a wider range of LQ images, such as a mixture of low-resolution, blur, noise, and JPEG artifacts (Wang X. et al., 2021).

3.3 Face restoration *via* Skip-Transformer

Training data collected based on the procedure in 3.2 are used to optimize our ST image restoration network. The overall architecture of the ST network is shown in Figure 3. It can be interpreted as a hybrid encoder-decoder architecture [as in GLEAN (Chan et al., 2021)] and a progressive style transformation [as in PSFRGAN (Chen C. et al., 2021)], which connects the transformer-based encoder with the StyleGAN-based decoder through skip layers. The key novelty of our design lies in 1) the joint exploitation of the global attention mechanism [*via* multiscale/pyramid transformer (Wang W. et al., 2021)] and pre-trained models [*via* StyleGAN (Karras et al., 2020)]; 2) the introduction of skip layers, which generalizes the existing skip connection (Huang et al., 2017), facilitates the information flow across the hybrid network.

3.3.1 Encoder

The encoder is made up of nine blocks in total, the first four consists of one or multiple residual blocks, the next four are called transformer blocks, and the last block is a fully connected layer. Every block starts with a convolution layer that is used to downsample features by setting the convolution stride as 2 (see the supplemental materials for more details). In the transformer block, we follow O'Shea and Nash (2015) and use depth-wise separable convolution (DSConv) (Howard et al., 2017) to project the input features to query, key, and value (Q, K, V). To reduce training and inference time, we divide the output of DSConv into smaller patches as in Eq. 5 and flatten them as in Eq. 6 where p^2 means how many patches to divide. Both the Q and K features are multiplied to generate the attention mask that is multiplied by the V feature map, as in Eq. 7 where d is the number of channels on the input feature map, to obtain the output. Inverse operations of flattening and patch division, called Patch Merge module, are performed to give the final output the same dimension as the input. Finally, a convolution layer is applied. Figures 4, 5 detail the components of the transformer block.

The self-attention layer is not used for the first four blocks due to computational cost, and instead we have only the residual block. Inspired by GLEAN (Chan et al., 2021), the first block of resolution 512×512 does not down-sample its input and contains four successive residual blocks. The last block is a fully connected layer that outputs the latent vector w from the $W \in R^{512}$ space of StyelGAN2 (Karras et al., 2020).

$$\text{Patch Division: } \mathbb{R}^{(C,W,H)} \rightarrow \mathbb{R}^{\left(Cp^2, \frac{W}{P}, \frac{H}{P}\right)} \tag{5}$$

$$\text{Flatten: } \mathbb{R}^{\left(Cp^2, \frac{W}{P}, \frac{H}{P}\right)} \rightarrow \mathbb{R}^{\left(\frac{WH}{p^2}, Cp^2\right)} \tag{6}$$

$$\text{Output} = \frac{softmax(QK^T)V}{\sqrt{d}} \tag{7}$$

TABLE 1 Comparison of degradation methods (lower FID is better).

Degradation type → networks type ↓	GFP-GAN (Wang et al., 2021b) FID ↓	Ours FID ↓
GFPGAN (Wang et al., 2021b)	59.20	**58.93**
GLEAN (Chan et al., 2021)	60.76	**58.53**
Skip-Transformer (**ours**)	55.23	**53.08**

The bold values denote the best performance.

FIGURE 7
Visual comparison among different approaches trained on the two types of degradation. Column **(A)** refers to the output of networks trained in GFP-GAN degradation (Wang X. et al., 2021) and column **(B)** refers to those trained in our degradation. Each row corresponds to a specific network architecture: GFP-GAN (Wang X. et al., 2021), GLEAN (Chan et al., 2021) and Skip-Transformer (ours) respectively.

3.3.2 Skip layers

Skip layers are used to combine encoder features with decoder features and can be thought of as an extension of skip connections (Huang et al., 2017). In total, we used six skip layers for the output resolution between 256 × 256 and 4 × 4; the multiscale concatenation of the skip layer is similar to that of the encoder block. For a hybrid design, we have used residual blocks for a resolution greater than 32 × 32 and a transformer block for other resolutions. The encoder and decoder features are concatenated, passed through the skip layer, and passed through a convolution layer to restore the original feature dimension of the decoder layer. For a better illustration, Figure 6 shows an example of how the skip layer works (for a resolution greater than 32 × 32, the self-attention block is not used).

TABLE 2 Comparison with state-of-the-art methods. ↓ means that a lower value is better.

Our test data			PSFR-RealTest			WebPhoto-test		
Method	FID ↓	NIQE ↓	Method	FID ↓	NIQE ↓	Method	FID ↓	NIQE ↓
DFDNet	109.67	5.80	DFDNet	76.15	5.24	DFDNet	101.28	5.22
GLEAN	72.05	4.91	GLEAN	51.40	4.47	GLEAN	90.85	5.02
PSFRGAN	62.28	4.14	PSFRGAN	52.28	4.09	PSFRGAN	85.32	4.06
GFP-GAN	54.78	3.99	GFP-GAN	44.41	4.02	GFP-GAN	87.76	4.21
DL-ST (Ours)	**53.05**	**3.63**	**DL-ST (Ours)**	**43.19**	**3.73**	**DL-ST (Ours)**	**81.96**	**3.76**

The bold values denote the best performance.

FIGURE 8
Qualitative comparison of our DL-ST and three competing methods (note that the eyes and ear regions were better restored than others).

3.3.3 Decoder

The decoder architecture is borrowed from the StyleGAN2 generator (Karras et al., 2020), and we use pre-trained weights to initialize the decoder network. Each layer takes the latent vector w (output from the fully connected encoder layer) and the output of the previous skip layer as input. Progressively, the decoder combines the corresponding encoder and decoder outputs into a pyramid reconstruction of the target face images. A similar architecture was used in both GFP-GAN (Wang X. et al., 2021) where a pre-trained GAN was used *a priori* and GauGAN (Park et al., 2019) where a series of residue blocks is used for semantic image synthesis.

3.3.4 Loss functions

ST network is optimized with the loss in Eq. 8. It consists of L_2 loss and loss of characteristics, where the pre-trained network ϕ ($VGG16$) is used to extract deep features from the input and the ground truth. The weights α, β and γ are 1, 0.02 and 0.02, respectively.

$$L_{ST} = \alpha\|\hat{y} - y\|_2 + \beta\|\phi(\hat{y}) - \phi(y)\|_2 + \gamma L_{adv} \tag{8}$$

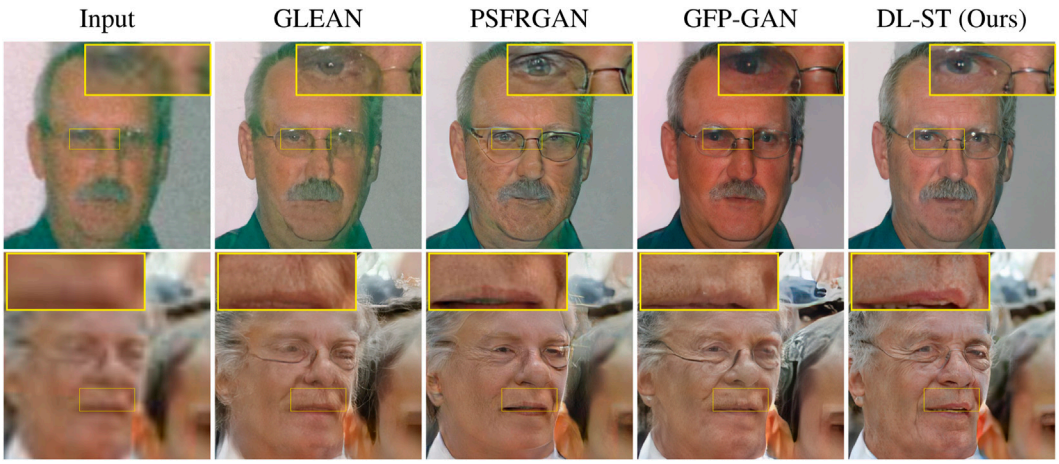

FIGURE 9
Qualitative comparison of ours and three competing methods on WebPhoto-Test data (top row) and PSFR-RealTest data (bottom row).

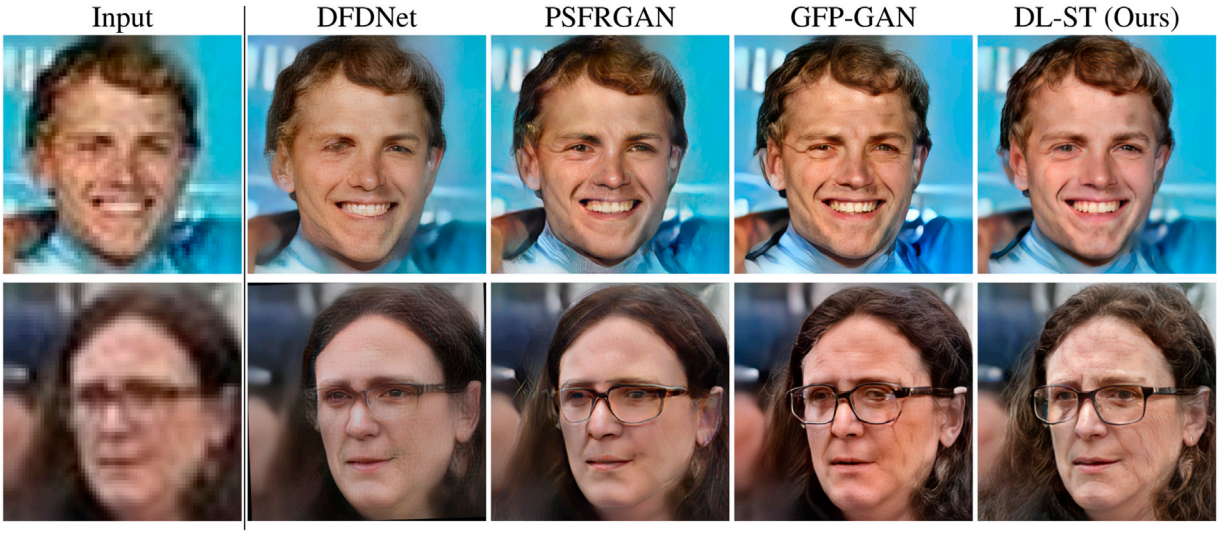

FIGURE 10
Our approach can restore faces with richer details and fewer artifacts than DFDNet (Wang X. et al., 2021), PSFRGAN (Chen C. et al., 2021) and GFP-GAN (Wang X. et al., 20210.

we have used cross-entropy loss for adversarial loss (L_{adv}). The discriminator network has the same architecture and is initialized with the trained weights of StyleGAN2 (Karras et al., 2020).

3.3.5 Training procedure

The learning rate is 5×10^{-5} for the encoder and skip layers of the generator, as well as for the discriminator. For the decoder (the prior network), the learning is 5×10^{-6}. We use cosine annealing with warm restarts (Loshchilov and Hutter, 2016) as a learning rate scheduler. Adam optimizer with beta parameters equal to (0.5, 0.99) and a batch size of eight is applied. To stop early, our network is trained for 220,000 iterations.

4 Experimental results

In this section, we show that 1) our training data generation procedure can improve the results of two other state-of-the-art architectures as well as ours, thanks to the improved DL model; 2) the proposed DL-ST network trained with our generated data outperforms previous state-of-the-art methods both visually and in terms of objective metrics, especially in the situation of extreme poses; 3) both second-order training and the self-attention mechanism contribute to the good performance of DL-ST.

FIGURE 11
Qualitative comparison of results for input with challenging pose.

TABLE 3 Comparison between the different training data generation approaches: (1) is the generation of data without two-order degradation; (2) is the proposed generation of data.

Dataset →	Our test data	PSFR-RealTest	WebPhoto-test
Approach ↓	FID ↓ NIQE ↓	FID ↓ NIQE ↓	FID ↓ NIQE ↓
Method (1)	58.96 3.92	43.82 3.88	84.29 4.01
Method (2)	**53.05 3.63**	**43.19 3.73**	**81.96 3.76**

The bold values denote the best performance.

TABLE 4 Comparison between the two network architectures with and without Self-Attention.

Dataset →	Our test data	PSFR-RealTest	WebPhoto-test
-	FID ↓ NIQE ↓	FID ↓NIQE ↓	FID ↓ NIQE ↓
w/o attention	55.33 3.76	44.23 3.82	85.43 4.05
w attention	**53.05 3.63**	**43.19 3.73**	**81.96 3.76**

The bold values denote the best performance.

4.1 Test dataset

To generate challenging test data, we collected low-quality face images from several public datasets: CelebA (Liu et al., 2015), VGGface2 (Cao et al., 2017), AFLW (Koestinger et al., 2011), LS3D-W (Bulat and Tzimiropoulos, 2017). We keep images with a face region of size smaller than or equal to 40 × 40. In total, the 653 facial images were aligned well before testing.

4.2 Data generation methods

To verify the effectiveness of our data generation approach, we train two state-of-the-art network architectures: GFP-GAN (Wang X. et al., 2021) and GLEAN (Chan et al., 2021). Both were created from scratch using our generated training dataset and the FFHQ dataset (Karras et al., 2018) with the degradation implemented in GFP-GAN (Wang X. et al., 2021).

For a fair comparison, all networks are trained in 200,000 iterations, and the losses and hyperparameters are kept the same as in the original paper. Fréchet Inception Distance (FID) scores were calculated. Table 1 shows that, when tested on our test dataset, the proposed data generation method leads to a lower FID metric for the three architectures. The visual comparison in Figure 7 confirms the superiority of networks trained on our degradation.

4.3 Comparison with state-of-the-arts methods

We compare our DL-ST method with four other state-of-the-art approaches: DFDNet (Li et al., 2020a), GLEAN (Chan et al., 2021, PSFRGAN (Chen C. et al., 2021), and GFP-GAN (Wang X. et al., 2021). Three datasets are used to evaluate network performance: our

test data described in Section 4.1, WebPhoto-Test dataset from GFP-GAN (Wang X. et al., 2021), and the PSFR-RealTest dataset from PSFRGAN (Chen C. et al., 2021). Table 2 shows that our DL-ST outperforms the others in both the FID and the NIQE metrics. Subjective comparisons of the qualitative results shown in Figures 8–10 also indicate that our method can arguably preserve more details and suffers from fewer artifacts than state-of-the-art methods. Note that the visual quality improvements are mostly visible around facial landmarks (e.g., eyes and mouth). This is because DL-ST has adopted a pre-trained StyleGAN-based decoder in our hybrid design.

4.3.1 Challenging poses

To further verify the validity of our DL-ST, we compare our proposed method with GFP-GAN (Wang X. et al., 2021) in the case of input faces with challenging poses. Figure 11 shows a visual comparison where both methods cannot perfectly reconstruct low-quality faces with challenging poses, but our proposed method can better reconstruct more details, especially for eyes and mouth with fewer artifacts. This improvement is directly related to the two-generator architecture for degradation learning. Compared to GFP-GAN, DL-ST has a better generalization property because it attempts to simulate a wide range of real-world degradations by style transfer (refer to Section 3.2).

4.4 Ablation study

4.4.1 Training Data Generation

To investigate the effectiveness of our proposed data generation method, we used 50% of the HQ data generated by the red generator (Figure 1) applied the degradation of GFP-GAN and 50% of the LQ data generated by the green generator (Figure 1) without the degradation of GFP-GAN applied to train our ST network [dubbed Method (1)]. Table 3 shows that with our proposed second-order data generation strategy [dubbed Method (2)], both the FID and the NIQE scores have been improved. Unlike the combination method we have been using so far [or what we call (2) in Table 3], GFP-GAN's degradation is not applied on top of the learned one; instead, both degradation sources are used 50% of the time. Furthermore, only high-quality synthetic images are used.

4.4.2 Importance of self attention

To demonstrate the validity of the attention mechanism in helping with the restoration task, we carried out another experiment that removed all self-attention modules from our proposed ST network. Table 4 shows that the ST network with the attention mechanism leads to a reduction in the FID and NIQE scores.

4.4.3 Failure cases

Despite the improvement over the current state-of-the-art, we recognize that DL-ST has its own limitations. The experimental results reported in this paper have only shown an improved generalization on a few popular test datasets. The generalization performance of DL-ST in more challenging real-world scenarios (e.g., extreme pose) remains to be further explored. Meanwhile, our experimental findings have shown that DL-ST is also susceptible to undesirable artifacts, such as wrinkles and specularities around the cheek regions. The question of how to address these weaknesses has been left for future research.

5 Conclusion

In this work, we proposed a state-of-the-art DL-ST framework for real-world face restoration tasks. The DL method combined the learned degradation style from real-world LQ face images and hand-made degradation to create a robust training dataset. Furthermore, a newly designed ST network with transformer and skip layers allows us to better use the generative priors for LQ face reconstruction. Our extensive experimental results convincingly outperform previous state-of-the-art methods, both subjectively and objectively. Future work includes the investigation of the explainability of the proposed DL-ST model and its further optimization in more challenging situations (e.g., extreme poses). How to suppress undesirable artifacts in reconstructed HR face images also deserves further study.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/NVlabs/ffhq-dataset.

Ethics statement

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

XX is AC's research mentor during the summer internship of 2021. XX helps AC with the experiments and paper preparation NX is AC's research manager and directly responsible for the problem formulation, resource support, and paper finalization.

Conflict of interest

This work was partially conducted when AC was doing a summer internship at Kwai Inc. Authors XX and NX were employed by Kwai Inc. at the time that this research was conducted.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Anwar, S., Porikli, F., and Huynh, C. P. (2017). Category-specific object image denoising. *IEEE Trans. Image Process.* 26, 5506–5518. doi:10.1109/tip.2017.2733739

Bulat, A., and Tzimiropoulos, G. (2017). *How far are we from solving the 2d and 3d face alignment problem? (and a dataset of 230, 000 3d facial landmarks). CoRR* abs/1703.07332.

Bulat, A., Yang, J., and Tzimiropoulos, G. (2018). "To learn image super-resolution, use a gan to learn how to do image degradation first," in Proceedings of the European conference on computer vision (ECCV) (Cham: Springer), 185–200.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2017). *Vggface2: A dataset for recognising faces across pose and age. CoRR* abs/1710.08092.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). "End-to-end object detection with transformers," in European Conference on Computer Vision (Berlin, Germany: Springer), 213–229.

Chan, K. C., Wang, X., Xu, X., Gu, J., and Loy, C. C. (2021). "Glean: Generative latent bank for large-factor image super-resolution," in Proceedings of the IEEE CVPR.

Chen, C., Li, X., Lingbo, Y., Lin, X., Zhang, L., and Wong, K. (2021a). *Progressive semantic-aware style transformation for blind face restoration*.

Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., et al. (2021b). "Pre-trained image processing transformer," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12299–12310.

Chen, Y., Tai, Y., Liu, X., Shen, C., and Yang, J. (2018). "Fsrnet: End-to-end learning face super-resolution with facial priors," in Proceedings of the IEEE Conference on CVPR, 2492–2501.

Esser, P., Rombach, R., and Ommer, B. (2021). "Taming transformers for high-resolution image synthesis," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12873–12883.

Ge, S., Zhao, S., Li, C., and Li, J. (2018). Low-resolution face recognition in the wild via selective knowledge distillation. *IEEE Trans. Image Process.* 28, 2051–2062. doi:10.1109/tip.2018.2883743

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). *Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR* abs/1704.04861.

Hu, X., Ren, W., Yang, J., Cao, X., Wipf, D. P., Menze, B., et al. (2021). Face restoration via plug-and-play 3d facial priors. *IEEE Trans. Pattern Analysis Mach. Intell.* 44, 8910. doi:10.1109/TPAMI.2021.3123085

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in Proceedings of the IEEE CVPR, 4700–4708.

Hudson, D. A., and Zitnick, C. L. (2021). *Generative adversarial transformers. arXiv preprint arXiv:2103.01209*.

Ichim, A. E., Bouaziz, S., and Pauly, M. (2015). Dynamic 3d avatar creation from hand-held video input. *ACM Trans. Graph.* 34, 1–14. doi:10.1145/2766974

Jiang, Y., Chang, S., and Wang, Z. (2021). *Transgan: Two transformers can make one strong gan. arXiv preprint arXiv:2102.07074* 1.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). "Perceptual losses for real-time style transfer and super-resolution," in European conference on computer vision (Berlin, Germany: Springer), 694–711.

Karras, T., Laine, S., and Aila, T. (2018). *A style-based generator architecture for generative adversarial networks. CoRR* abs/1812.04948.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). "Analyzing and improving the image quality of StyleGAN," in Proc. CVPR.

Kazemi, V., and Sullivan, J. (2014). "One millisecond face alignment with an ensemble of regression trees," in Proceedings of the IEEE conference on CVPR, 1867–1874.

Koestinger, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2011). "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops) (Barcelona, Spain: IEEE). doi:10.1109/ICCVW.2011.6130513

Li, J., Huang, H., Jia, X., and He, R. (2021). *Universal face restoration with memorized modulation. arXiv preprint arXiv:2110.01033*.

Li, X., Chen, C., Zhou, S., Lin, X., Zuo, W., and Zhang, L. (2020a). "Blind face restoration via deep multi-scale component dictionaries," in European Conference on Computer Vision (Berlin, Germany: Springer), 399–415.

Li, X., Li, W., Ren, D., Zhang, H., Wang, M., and Zuo, W. (2020b). "Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2706–2715.

Li, X., Liu, M., Ye, Y., Zuo, W., Lin, L., and Yang, R. (2018). "Learning warped guidance for blind face restoration," in Proceedings of the European conference on computer vision (ECCV), 272–289.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., and Timofte, R. (2021). "Swinir: Image restoration using swin transformer," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 1833–1844.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 10012–10022.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). "Deep learning face attributes in the wild," in Proceedings of International Conference on Computer Vision (ICCV).

Loshchilov, I., and Hutter, F. (2016). *Sgdr: Stochastic gradient descent with restarts. CoRR* abs/1608.03983.

Ma, C., Jiang, Z., Rao, Y., Lu, J., and Zhou, J. (2020). "Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation," in Proceedings of the IEEE/CVF conference on CVPR, 5569–5578.

Menon, S., Damian, A., Hu, S., Ravi, N., and Rudin, C. (2020). "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, 2437–2445.

Mo, S., Cho, M., and Shin, J. (2020). *Freeze discriminator: A simple baseline for fine-tuning gans. CoRR* abs/2002.10964.

O'Shea, K., and Nash, R. (2015). *An introduction to convolutional neural networks. CoRR* abs/1511.08458.

Park, T., Liu, M. Y., Wang, T. C., and Zhu, J.-Y. (2019). "Semantic image synthesis with spatially-adaptive normalization," in Proceedings of the IEEE/CVF conference on CVPR, 2337–2346.

Pinkney, J. N. M., and Adler, D. (2020). *Resolution dependent GAN interpolation for controllable image synthesis between domains. CoRR* abs/2010.05334.

Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., et al. (2021). "Encoding in style: A stylegan encoder for image-to-image translation," in IEEE/CVF Conference on CVPR.

Shen, Z., Lai, W.-S., Xu, T., Kautz, J., and Yang, M.-H. (2018). "Deep semantic face deblurring," in Proceedings of the IEEE Conference on CVPR, 8260–8269.

Shen, Z., Lai, W.-S., Xu, T., Kautz, J., and Yang, M.-H. (2020). Exploiting semantics for face image deblurring. *Inter. J. Comput. Vis.* 128, 1829–1846. doi:10.1007/s11263-019-01288-9

Simonyan, K., and Zisserman, A. (2015). "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in Advances in neural information processing systems, 5998–6008.

Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., et al. (2021a). *Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. arXiv preprint arXiv:2102.12122*.

Wang, X., Li, Y., Zhang, H., and Shan, Y. (2021b). "Towards real-world blind face restoration with generative facial prior," in The IEEE Conference on CVPR.

Wang, Z., Cun, X., Bao, J., and Liu, J. (2021c). *Uformer: A general u-shaped transformer for image restoration. arXiv preprint arXiv:2106.03106*.

Wang, Z., Zhao, L., Chen, H., Qiu, L., Mo, Q., Lin, S., et al. (2020). "Diversified arbitrary style transfer via deep feature perturbation," in Proc. of the IEEE/CVF Conf. on CVPR, 7789–7798.

Yang, C. Y., Liu, S., and Yang, M. H. (2018). Hallucinating compressed face images. *Inter. J. Comput. Vis.* 126, 597–614. doi:10.1007/s11263-017-1044-4

Yang, F., Yang, H., Fu, J., Lu, H., and Guo, B. (2020a). "Learning texture transformer network for image super-resolution," in Proceedings of the IEEE/CVF Conference on CVPR, 5791–5800.

Yang, L., Wang, S., Ma, S., Gao, W., Liu, C., Wang, P., et al. (2020b). "Hifacegan: Face renovation via collaborative suppression and replenishment," in Proceedings of the 28th ACM International Conference on Multimedia, 1551–1560.

Yang, S., Luo, P., Loy, C.-C., and Tang, X. (2016). "Wider face: A face detection benchmark," in Proceedings of the IEEE conference on CVPR, 5525–5533.

Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., and Li, S. Z. (2017). *S^3fd: Single shot scale-invariant face detector. CoRR* abs/1708.05237.