Recursive Copy and Paste GAN: Face Hallucination from Shaded Thumbnails

Yang Zhang, Ivor W. Tsang, *Senior Member, IEEE,* Yawei Luo, Changhui Hu, Xiaobo Lu, and Xin Yu

Abstract—Existing face hallucination methods based on convolutional neural networks (CNNs) have achieved impressive performance on low-resolution (LR) faces in a normal illumination condition. However, their performance degrades dramatically when LR faces are captured in non-uniform illumination conditions. This paper proposes a Recursive Copy and Paste Generative Adversarial Network (Re-CPGAN) to recover authentic high-resolution (HR) face images while compensating for non-uniform illumination. To this end, we develop two key components in our Re-CPGAN: internal and recursive external Copy and Paste networks (CPnets). Our internal CPnet exploits facial self-similarity information residing in the input image to enhance facial details; while our recursive external CPnet leverages an external guided face for illumination compensation. Specifically, our recursive external CPnet stacks multiple external Copy and Paste (EX-CP) units in a compact model to learn normal illumination and enhance facial details recursively. By doing so, our method offsets illumination and upsamples facial details progressively in a coarse-to-fine fashion, thus alleviating the ambiguity of correspondences between LR inputs and external guided inputs. Furthermore, a new illumination compensation loss is developed to capture illumination from the external guided face image effectively. Extensive experiments demonstrate that our method achieves authentic HR face images in a uniform illumination condition with a $16\times$ magnification factor and outperforms state-of-the-art methods qualitatively and quantitatively.

Index Terms—Face hallucination, super-resolution, illumination normalization, generative adversarial network.

1 Introduction

FACE hallucination, also known as face super-resolution (FSR), refers to generating high-resolution (HR) face images from their corresponding low-resolution (LR) inputs, has received significant attention in recent years. Existing face hallucination methods mainly focus on super-resolving face images with uniform illumination. However, due to the non-ideal imaging environments, the captured face images may be tiny and shaded. Hallucinating such shaded LR faces requires either face illumination normalization followed by face hallucination techniques, or face hallucination methods followed by illumination normalization. Note that in order to super-resolve shaded LR faces, existing FSR methods often rely on the availability of illumination-specific exemplar datasets. Nonetheless, both of these options are naturally very challenging.

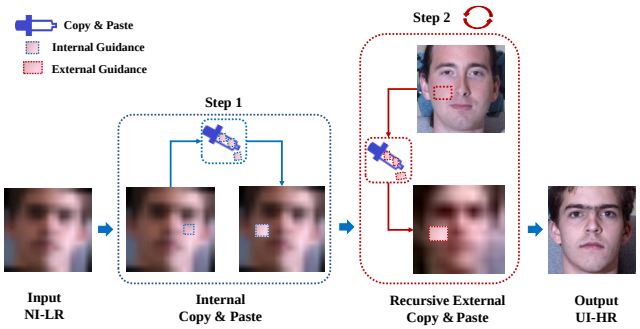


Fig. 1: Motivation of Re-CPGAN. Internal and recursive external CPnets are introduced to mimic the Clone Stamp Tool. Internal CPnet copies well-illuminated facial details and then paste them onto shadow regions. Recursive external CPnet further retouches the face using an external guided face from the UI-HR face database during the upsampling process to compensate for uneven illumination in the final HR face.

The state-of-the-art face illumination processing methods [1]–[3] usually fit the face region to a pre-aligned face template based on facial landmarks and then normalize illumination. However, these methods are unsuitable for tiny faces because accurate facial landmark detection requires the input image with a sufficient resolution. This leads to suboptimal illumination normalization results (see Fig. 2(d)). Moreover, the produced errors cannot be eliminated by the subsequent face hallucination process but are exaggerated (see Fig. 2(e)). Similarly, as shown Fig. 2(f), for shaded LR faces, applying illumination normalization followed by face hallucination also produces degraded results with obvious distortions and severe artifacts.

- *This work was done when the first author visited the University of Technology Sydney. Y. Zhang is with the School of Automation, Southeast University, Nanjing 210096, China; the Australian Institute of Artificial Intelligence, University of Technology Sydney, Ultimo, NSW 2007, Australia; the Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing 210096, China, e-mail: zhangyang201703@126.com.*
- *I. W. Tsang and X. Yu are with the Australian Institute of Artificial Intelligence, University of Technology Sydney, Ultimo, NSW 2007, Australia, e-mail: ivor.tsang@uts.edu.au, xin.yu@uts.edu.au.*
- *X. B. Lu (corresponding author) is with the School of Automation, Southeast University, Nanjing 210096, China; Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing 210096, China, e-mail: xblu2013@126.com.*
- *Y. W. Luo is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310007, China, e-mail: yaweiluo329@gmail.com.*
- *C. H. Hu is with the College of Automation and the College of Artificial Intelligent, Nanjing University of Posts and Telecommunications, Nanjing 210023, China, e-mail: hchnjupt@126.com.*

(a) Input (b) Guided HR (c) GT (d) [3] (e) [3] + [4] (f) [4] + [3] (g) CPGAN (h) Re-CPGAN

Fig. 2: Comparison of our proposed Re-CPGAN with the state-of-the-art methods (16×16, $8\times$)[1]. (a) Interpolated unaligned NI-LR image. (b) Guided UI-HR image (128×128 pixels). (c) Ground-truth UI-HR image (128×128 pixels, **not available in training**). (d) Illumination normalization result of (a) by applying [3]. (e) Face hallucination result of (d) by applying [4]. (f) Result of face hallucination followed by illumination normalization by applying [4] and then [3] to the NI-LR face. (g) Result of our previous method CPGAN [5]. (h) Result of Re-CPGAN (128×128 pixels).

In this paper, we aim to hallucinate LR inputs under non-uniform illumination (NI-LR)[2] while achieving HR faces under uniform illumination (UI-HR)[3] in a unified framework. In particular, these two tasks (*i.e.,* face hallucination and illumination compensation) will be addressed simultaneously and mutually facilitate each other. To this end, we propose a Recursive Copy and Paste Generative Adversarial Network (Re-CPGAN), which adopts the internal and external guided illumination information to normalize the input NI-LR face progressively during the up-sampling procedure (as illustrated in Fig. 1).

Re-CPGAN: It consists of two components: a copy and paste based transformative upsampling network (CPUN) which embodies an internal CPnet, a recursive external CPnet and a face reconstruction net, as well as a discriminative network. We first design an internal CPnet to initially offset non-uniform illumination features and roughly enhance facial details by exploiting facial self-similarity information within an input NI-LR face. Then, we propose a recursive external CPnet to learn illumination patterns from a guided UI-HR face and upsample facial features. Our recursive external CPnet stacks External Copy and Paste (EX-CP) units for normalizing illumination and enhancing facial details alternatingly, and employs a global skip connection to pass low-frequency facial information to the output while mitigating the difficulty of training deep networks. Specifically, we employ recursive learning for the Ex-CP unit to make our model deep yet compact. In doing so, we normalize illumination and recover diverse characteristics of NI-LR inputs progressively. Afterwards, we transform the refined feature maps (multi-channel) back to the original image space (RGB-channel) via the face reconstruction net to generate the UI-HR face. Inspired by previous works [6]–[8], we employ the discriminative network to enforce the UI-HR output to resemble real human faces. Finally, we propose an illumination compensation loss to capture the normal illumination pattern and transfer the normal illumination to the inputs. As shown in Fig. 2(h), our hallucinated UI-HR face is realistic, and resembles the ground-truth with normal illumination.

Data Augmentation: Training a deep neural network requires very large datasets to prevent over-fitting. In our case, the existing public face datasets [9], [10] do not provide a sufficient number of NI/UI face pairs. For the training purpose, we propose a tailor-made Random Adaptive Instance Normalization (RaIN) model as our "illumination rendering engine". The proposed RaIN model adopts an encoder-decoder architecture and employs an Adaptive Instance Normalization (AdaIN) [11] layer and a Variational Auto-Encoder (VAE) [12] in the latent space. Specifically, we exploit AdaIN to normalize the features of a UI face image and then enforce the features to share the same channel-wise mean and standard deviation as those of a selected NI face image features. The VAE is inserted before the AdaIN layer to produce an unlimited number of plausible hypotheses for the feature statistics of the NI face image. Our RaIN model is able to transfer various illumination conditions to face images and thus generates sufficient NI face samples from UI-HR inputs. As a result, we construct a large corpus of NI/UI face pairs for training our Re-CPGAN.

Extension: In our previous work [5], we propose CPGAN which directly stacks multiple external CPnets and deconvolutional layers to offset non-uniform illumination and upsample facial details alternatingly. However, adding more convolutional layers introduces more parameters, and handicaps the model deployment in practice. As a notable extension of our previous work, we inherit the copy and paste strategy and design a more advanced network architecture from the perspective of recursive learning, thereby constructing a deep yet compact model with better hallucination performance. The major improvements lie in four-folds: (1) We design a recursive external CPnet to learn illumination and upsample facial details to improve the capability of our face hallucination network while achieving a deep yet compact model Re-CPGAN. (2) We provide analyses on our model with increasing recursions in terms of qualitative and quantitative performance as well as landmark estimation accuracy. (3) We further adopt three evaluation metrics, *i.e.,* face recognition rate, expression classification rate and facial landmark localization error, to evaluate our face hallucination performance and demonstrate the advantage of Re-CPGAN. (4) We evaluate our model on more challenging situations including the inputs under extremely low resolutions, large poses and complex expressions, and confirm that our Re-CPGAN outperforms the state-of-the-art in all cases.

The contributions of our work are summarized as follows:

1. (16×16, $8\times$): 16×16 represents the resolution of the original input NI-LR face; $8\times$ represents the magnification factor.
2. NI-LR faces: low-resolution faces under non-uniform illumination.
3. UI-HR faces: high-resolution faces under uniform illumination.

- We present a novel framework, dubbed Re-CPGAN, to address face hallucination and illumination compensation together in an end-to-end manner. Re-CPGAN is optimized by not only the conventional face hallucination losses but also a newly introduced illumination compensation loss.
- We design an internal CPnet to normalize illumination and enhance facial details coarsely, aiding subsequent illumination compensation and upsampling processes.
- We present a recursive external CPnet to learn illumination features from an external guided face. In this fashion, we are able to learn illumination explicitly rather than over-fitting to a certain illumination condition. With the recursive learning, the performance of our model can be significantly improved without introducing new parameters for additional layers.
- A tailor-made data augmentation model, namely RaIN, is proposed to generate sufficient NI/UI face pairs. Our constructed NI/UI face pair database will be publicly available for reproducibility.
- Our experiments demonstrate that Re-CPGAN is able to normalize and super-resolve (by a large upscaling factor of 16×) NI-LR face images (*e.g.,* 8×8 pixels) undergoing large poses (*e.g.,* 90^o) and complex facial expressions (*e.g.,* "disgust", "surprise"). Moreover, our Re-CPGAN is capable of providing superior hallucinated face images for downstream tasks, *i.e.,* face recognition and expression classification, in comparison to the state-of-the-art.

2 Related Work

2.1 Face Hallucination

Face hallucination methods aim at establishing the intensity relationships between input LR and output HR face images. The prior works can be categorized into three mainstreams: holistic-based, part-based, and deep learning-based methods.

The basic principle of holistic-based techniques is to upsample a whole LR face by a global face model. Wang *et al.* [13] formulate a linear mapping between LR and HR images to achieve face super-resolution based on an Eigen-transformation of LR faces. Liu *et al.* [14] incorporate a bilateral filtering to mitigate the ghosting artifacts. Kolouri and Rohde [15] morph HR faces from aligned LR ones based on optimal transport and subspace learning. However, they require LR inputs to be precisely aligned and reference HR faces to exhibit similar canonical poses and natural expressions.

To address pose and expression variations, part-based methods are proposed to make use of exemplar facial patches to upsample local facial regions instead of imposing global constraints. The approaches [16]–[18] super-resolve local LR patches based on a weighted sum of exemplar facial patches in reference HR database. Liu *et al.* [19] develops a locality-constrained bi-layer network to jointly super-resolve LR faces as well as eliminate noise and outliers. Moreover, SIFT flow [20] and facial landmarks [21] are introduced to locate facial components for further super-resolution. Since these techniques need to localize facial components in LR inputs preciously, they may fail to process very LR faces.

Recently, deep learning based face hallucination methods have been actively explored and achieved superior performance compared to traditional methods. Yu *et al.* [22] exploit deconvolutional layers to upsample LR faces and employ unsharp filtering to enhance image sharpness [23]. Later, Yu *et al.* [24], [25] develop GAN-based models to hallucinate very LR face images. Huang *et al.* [26] incorporate the wavelet coefficients into deep convolutional networks to super-resolve LR inputs with multiple upscaling factors. Cao *et al.* [27] design an attention-aware mechanism and a local enhancement network to alternately enhance facial regions in super-resolution. Xu *et al.* [28] jointly super-resolve and deblur face and text images with a multi-class adversarial loss. Dahl *et al.* [29] present an autoregressive PixelRNN [30] to hallucinate pre-aligned LR faces. Yu *et al.* [6] present a multiscale transformative discriminative network to hallucinate unaligned input LR face images with different resolutions. Menon *et al.* [31] present a Photo Upsampling via Latent Space Exploration (PULSE) algorithm to generate high-quality frontal face images at large resolutions. Zhang *et al.* [32] develop a two-branch super-resolution network to compensate and upsample ill-illuminated LR face images. However, these methods focus on super-resolving near-frontal LR faces. Consequently, they are restricted to the inputs under small pose variations.

Several face hallucination techniques have been proposed to super-resolve LR faces under large pose variations by introducing facial prior information [4], [33], [34]. Chen *et al.* [33] incorporate facial geometry priors into their hallucination model to super-resolve LR faces. Bulat *et al.* [34] propose a method to learn not only the mappings between LR and HR faces but also the real-world degeneration process. Yu *et al.* [4] exploit the facial component information from the intermediate upsampled features to encourage the upsampling stream to produce photo-realistic HR faces. However, these approaches only hallucinate tiny face images with normal illumination.

2.2 Face Illumination Compensation

Face illumination compensation methods focus on compensating for uneven illumination in face images. Conventional and emerging researches on face illumination compensation can be grouped into two classes: illumination normalization and illumination synthesis methods.

Illumination normalization methods usually obtain the normal illumination face image by directly manipulating or modifying the input face. Shashua *et al.* [35] propose a quotient image technique for face illumination normalization, where an input face image is normalized by multiplying it with the ratio of a reference uniform illumination face and the input one. Quotient images have also been used to transfer subtle shading effects caused by expression variations [36] and match illumination for face swapping [37]. [38] presents a novel image processing chain to calculate illumination-insensitive features. Chen *et al.* [1] design detailed illumination layers based on edge-preserving filters to modify non-uniform illumination in input images. However, these methods require an input face to be strict pre-aligned, which is impractical in serve illumination and low-resolution cases.

Recently, image-to-image translation algorithms have been proposed to tackle the face illumination transfer problem including illumination normalization. Tran *et al.* [39] propose a GAN-based model to learn disentangled representations of input faces, and then generate label-assisted face images. Yang *et al.* [40] design an IL-GAN model to produce face images with desired illumination styles by injecting illumination codes. Zhu *et al.* [41] propose a cycle consistent network to render a content image to new images with different styles. However, these multi-domain

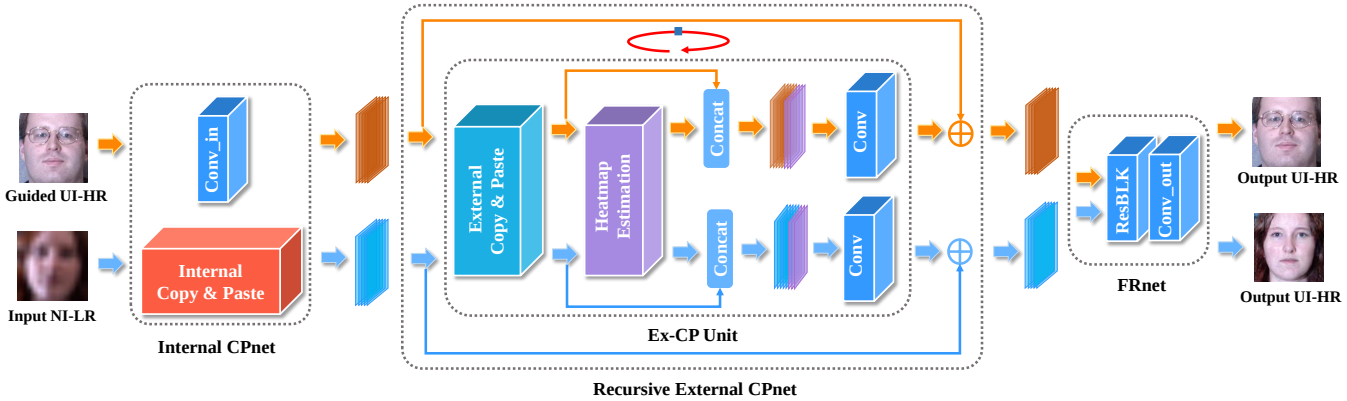


Fig. 3: The pipeline of the copy and paste upsampling network (CPUN) in our proposed Re-CPGAN.

transfer techniques require explicit domain labels. FUNIT [42] is proposed to generate the target domain images with only a few examples. HiDT [43] is a recent advanced image-to-image translation method that does not rely on domain labels during either training or inference. It is able to re-render an image with different illumination conditions in a continuous space. However, without imposing facial priors to their framework, these methods would reconstruct inferior facial details, especially when the resolutions of input images are low.

Illumination synthesis methods infer intrinsic face properties, material properties and illumination separately based on the physical lighting model [44]–[46]. Blanz *et al.* [47] first propose the 3D Morphable Model (3DMM) to estimate and synthesize lighting conditions by a linear combination of prototype models. Then, Wang *et al.* [48] design a 3D spherical harmonic basis morphable model (SHBMM), fusing 3DMM and spherical harmonic illumination representation. However, since existing 3DMMs are always built with face images captured in controlled environments, these 3DMM-based methods only work well in under-controlled scenarios. Then, Barron *et al.* [49] define a simple optimization problem in which they recover the reasonable intrinsic scene properties including shape, reflectance, and illumination under the guidance of image priors. [50] enforces illumination representations to be aware of face geometry by employing a generic three-dimensional morphable face model, where the spherical harmonics coefficients and the standard color histogram matching are used to model the illumination. Saito *et al.* [51] synthesize a photo-realistic albedo from a partial albedo based on traditional methods. Wang *et al.* [46] decompose illumination into different channels, including specular and shadows, and exploit a network to learn such decomposition. However, those methods often resort to graphic rendering and thus are very time-consuming to obtain a large number of images. Recently, some deep learning approaches are proposed to disentangle real-life face images. Zhou *et al.* [**?**] present a Label Denoising Adversarial Network (LDAN) for lighting regression on real face images. Shu *et al.* [52] propose a physically grounded rendering-based disentangling network to render in-the-wild faces with real and arbitrary backgrounds. SfSNet [3] is inspired by a physical rendering model and disentangles normal and albedo into separate subspaces.

2.3 Recursive Neural Network

Recursive neural network has been proposed to address various tasks, such as semantic segmentation [53], object classification [54], and image super-resolution [55] and image deraining [56].

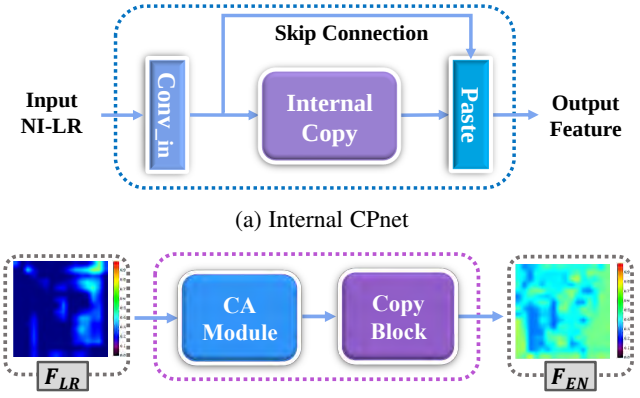
Socher *et al.* [57] propose a model based on a combination of convolutional and recursive neural networks to learn features from RGB-D data effectively for 3D object classification. Liang *et al.* [54] propose a recurrent CNN for object recognition by incorporating recurrent connections into each convolutional layer of the feed-forward CNN. Kim *et al.* [58] present a very deep recursive convolutional network using a chain structure, namely DRCN, for image super-resolution. To mitigate the training difficulty, DRCN uses recursive-supervision and skip-connections to promote gradient back-propagation, and adopts an ensemble strategy to further improve the model performance. Similarly, Ying *et al.* [55] design a recursive block consisting of several residual units and construct a very deep CNN model with the recursive blocks. Benefiting from the recursive fashion, models can achieve better performance while maintaining the parameter sizes.

3 Hallucination with "Copy" and "Paste"

To reduce the ambiguous mapping from NI-LR to UI-HR faces caused by non-uniform illumination, we present an Re-CPGAN framework that takes a NI-LR face as the input and an external HR face with normal illumination as a guidance to hallucinate a UI-HR one. Our Re-CPGAN consists of a copy and paste upsampling network and a discriminative network. The copy and paste upsampling network introduces external guided illumination information into the hallucination process to normalize the lighting conditions of input NI-LR faces. The discriminative network is used to enforce the generated UI-HR faces to lie on the manifold of real face images.

3.1 Copy and Paste Upsampling Network (CPUN)

Our CPUN is composed of three parts: an internal CPnet, a recursive external CPnet and a face reconstruction net, as shown in Fig. 3. First, we design the internal CPnet to enhance facial details and normalize illumination coarsely for an input NI-LR face by exploiting facial self-similarity information. Meanwhile, we employ a convolutional layer to extract features of a guided UI-HR face. Note that, our guided face is different from the ground-truth of the NI-LR input. Second, we propose a recursive external CPnet that resorts to the external guided features for further illumination compensation during the upsampling process. Finally, we use the face reconstruction net to transform the refined features to the RGB image.



(a) Internal CPnet

(b) Internal Copy Block

Fig. 4: The architecture of the internal CPnet. Copy block here treats the output features of CA module as the both content features and guided features. Paste block here represents the additive operation functionally. Here, we conduct the channel-wise average operation on the feature maps.

3.1.1 Internal CPnet

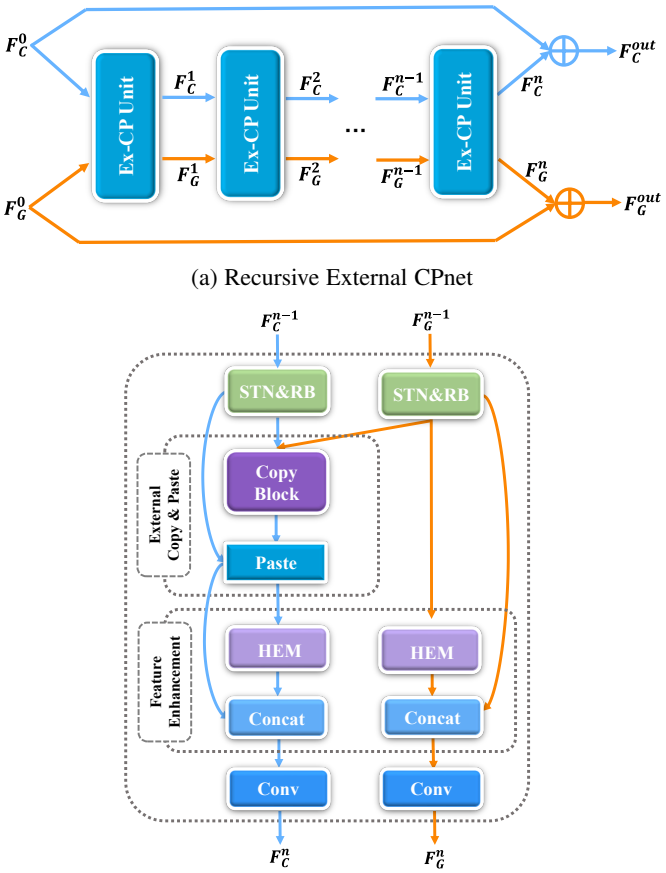
Our internal CPnet consists of an input convolutional layer, an internal copy block, a paste block and a skip-connection (see Fig. 4(a)). Since an input NI-LR face often contains few facial details, we design the internal copy block (see Fig. 4(b)), which adopts a Channel-wise Attention (CA) module [59] and a copy block, to enhance high-frequency facial details and normalize illumination coarsely. The CA module [59] is employed to model inter-dependencies among channels, and thus enhances channel-wise high-frequency features. Meanwhile, the copy block (see Fig. 6(a)) is designed to capture spatial dependencies between any two positions within the input feature maps, and then enhances spatial-wise high-frequency features. Note that, our copy block here takes the output features of the CA module as both content features (F_C) and guided features (F_G), as shown in Fig. 6(a). Then, the high-frequency facial details can be effectively recovered by the internal copy block.

To demonstrate the effect of our proposed internal copy block, we visualize the changes between the input and output feature maps (see Fig. 4(b)). Since the input NI-LR face does not contain discriminative facial features, the input features F_{LR} mainly reside in the low-frequency band (in blue color). After our internal copy block, the output features F_{EN} spread in the direction of high-frequency band (in red color), and span the whole band. Therefore, we use the name "internal copy block" because its functionality resembles an operation that "copies" the high-frequency features to the low-frequency parts.

As shown in Fig. 7(c), the Re-CPGAN variant without the internal CPnet produces inferior results. This also indicates that our internal CPnet initially refines the input NI-LR face at the feature level and thus benefits subsequent face hallucination and illumination compensation processes.

3.1.2 Recursive External CPnet

Our recursive external CPnet is composed of two parts: cascaded Ex-CP units that offset non-uniform illumination and enhance facial details recursively, and a global skip connection that helps gradient back-propagation during training (see Fig. 5(a)).



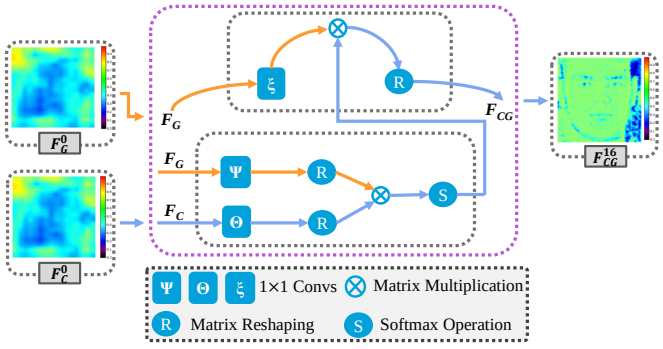
(a) Recursive External CPnet

(b) EX-CP Unit

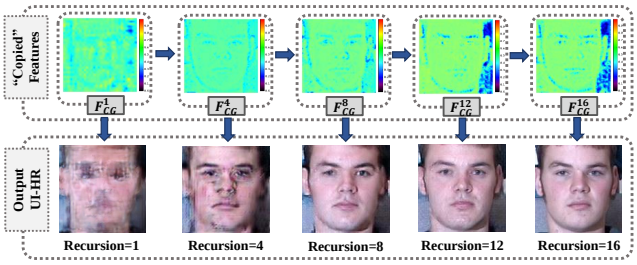
Fig. 5: The architecture of the recursive external CPnet. Here, STN, RB and HEM represent the spatial transformer network, the residual block, and the heatmap estimation module, respectively.

The Ex-CP unit consists of a spatial transformer network (STN) [60], a residual block, a copy block, a paste block, a heatmap estimation module (HEM) and an output convolutional layer (see Fig. 5(b)). Since the input NI-LR faces may undergo misalignment, such as in-plane rotations, translations and scale changes, STN is employed to compensate for misalignment [8], [61], as shown in the cyan blocks in Fig. 5(b). Meanwhile, inspired by SRGAN [62], the residual block is adopted to recover photo-realistic textures from LR features. Then, we design the copy block (see Fig. 6(a)) to explicitly learn the illumination pattern from the external guided UI-HR face. Afterwards, inspired by [63], we employ the stacked hourglass networks [64] as our heatmap estimation module. It estimates facial structure priors, *i.e.,* facial landmark heatmaps, from intermediate facial features to preserve facial structure. Finally, we concatenate the estimated priors with the refined facial features. In this fashion, we exploit not only low-level information (*i.e.,* intensity similarity) but also middle-level information (*i.e.,* facial structure) to achieve accurate hallucination results. Therefore, the Ex-CP unit streamlines the process of alignment, illumination normalization, structure estimation and super-resolution on NI-LR faces.

Furthermore, we apply a recursive mechanism to the Ex-CP unit to increase the depth of our model without introducing extra parameters. In this way, we can progressively adjust alignment, normalize illumination and recover facial characteristics for the NI-LR inputs. This distinctive design also alleviates the ambiguity



(a) Copy Block



(b) Impacts of Recursion Depths on the "Copied" Features.

Fig. 6: The diagram of the copy block. The "copied" features F_{CG}^{N} represent the output features of the copy block in the Re-CPGAN variant with N EX-CP units. Here, we conduct the channel-wise average operation on the feature maps.

of correspondences between NI-LR inputs and external UI-HR ones. As seen in Figs. 7(k) and (l)), our hallucinated faces become more visually appealing as the recursion depth increases.

3.1.3 Copy Block

Fig. 6(a) depicts the "copy" procedure of our copy block. The guided features F_G^0 and content features F_C^0 are extracted from the external guided UI-HR image and the input NI-LR image respectively. First, the guided features F_G and content features F_C are normalized and transformed into a common feature space by applying two mappings ψ and θ for feature similarity measurement. Then, the "copied" features F_{CG} can be formulated as a weighted sum of the guided features F_G that are similar to the content features F_C at different positions. The i-th output response is expressed as:

$$F_{CG^i} = \frac{1}{M(F)} \sum_{\forall j} \left\{ \exp\left(W_\theta^T \left(\overline{F_{C^i}}\right)^T \overline{F_{G^j}} W_\psi \right) F_{G^j} W_\zeta \right\}, \quad (1)$$

where $M(F) = \sum_{\forall j} \exp\left(W_\theta^T \left(\overline{F_{C^i}}\right)^T \overline{F_{G^j}} W_\psi \right)$ is the sum of all output responses over all positions. $\overline{F}$ is a transform on F by applying the mean-variance channel-wise normalization. Here, the embedded transformations W_θ, W_ψ and W_ζ are learnt during training.

Although our copy block shares a similar network architecture to the existing non-local module [65], it differs from non-local module since our copy block focuses on addressing the fusion of the guided and content features. As a consequence, the copy block can integrate the illumination of guided features into content features as well as enhance high-frequency facial details (see F_{CG}^{16} in Fig. 6(a)). Furthermore, as shown in Fig. 6(b), the facial details in the generated "copied" features F_{CG} become much clearer as recursions increase. This indicates that the recursive structure helps the "copy" operation in a coarse-to-fine manner and enhances facial details progressively.

(a) (b) (c) (d)
(e) (f) (g) (h)
(i) (j) (k) (l)

Fig. 7: Impacts of different components and losses on face super-resolution (16×16, $8\times$). (a) Interpolated unaligned NI-LR image. (b) Ground-truth UI-HR image (128×128 pixels). (c) Result without using the internal CPnet but a simple input convolutional layer instead. (d) Result without using the recursive external CPnet. (e) Result of Re-CPGAN without adopting L_{ic}. Note that specular appears in the left side of the forehead. (f) Result of Re-CPGAN trained by L_{mse}. (g) Result of Re-CPGAN trained by L_{mse} and L_{id}. (h) Result of CPUN (Re-CPGAN without employing L_{adv}). (i) Result without using an external guided face. Content features (F_C) replace guided features (F_G) in the copy block. (j) Result without data augmentation. (k) Result of Re-CPGAN with 8 Ex-CP units. (l) Result of Re-CPGAN. Note that, in this experiment 16 Ex-CP units are used for all the cases except (k).

3.2 Discriminative Network

Inspired by [6], [7], we employ a discriminative network to force the generated UI-HR faces to lie on the same manifold as real UI-HR ones. Our discriminative network consists of convolutional layers, max-pooling layers, dropout layers, and fully-connected layers. It is designed to determine whether an image is sampled from real face images or the hallucinated ones. As shown in Figs. 7(h) and (l), it can be clearly seen that the results of Re-CPGAN (see Fig. 7(l)) are more photo-realistic.

3.3 Training Procedure

We construct NI-LR/UI-HR face pairs $\{l_i, h_i\}$ for our training purpose, where h_i represents the aligned UI-HR face images (only eyes are aligned), and l_i represents the synthesized unaligned NI-LR face images. More details are provided in Sec. 4. Note that, our guided faces g_i are randomly selected from h_i and different from the ground-truth of l_i.

Our Re-CPGAN is trained in an end-to-end fashion. To compensate for the uneven illumination in output images, we develop

an illumination compensation loss L_{ic}. To minimize discrepancies between output images and their ground-truth counterparts, we employ two losses. The first one is an intensity similarity loss L_{mse} to maintain the pixel-wise intensity similarity and the second one is an identity similarity loss L_{id} to enforce the feature-wise similarity. To preserve the structural integrity of generated faces, we introduce the structure similarity loss L_h [7]. To enforce the output faces to resemble real ones, an adversarial loss L_{adv} [66] is also employed.

Illumination compensation loss: Inspired by the style loss in [11], we propose the illumination compensation loss L_{ic}. L_{ic} constrains the illumination characteristics of the reconstructed UI-HR face to be close to those of the guided UI-HR face in the latent subspace:

$$\begin{aligned} L_{ic} =& \mathbb{E}_{(\hat{h}_i,g_i)\sim p(\hat{h},g)}\{\sum_{j=1}^{L}\left\|\mu\left(\varphi_j(\hat{h}_i)\right)-\mu\left(\varphi_j(g_i)\right)\right\|_2 \\ &+\sum_{j=1}^{L}\left\|\sigma\left(\varphi_j(\hat{h}_i)\right)-\sigma\left(\varphi_j(g_i)\right)\right\|_2\}, \end{aligned} \tag{2}$$

where g_i represents the guided UI-HR image, $\hat{h}_i$ represents the generated UI-HR image, and $p(\hat{h},g)$ represents their joint distribution. $\varphi_j(\cdot)$ denote the outputs of relu1_1, relu2_1, relu3_1 and relu4_1 layers in a pre-trained VGG-19 model [67]. Here, μ and σ are the mean and variance of each feature channel. Fig. 7(e) shows that without employing L_{ic}, the upsampled face suffers from severe artifacts. This demonstrates that our L_{ic} significantly mitigates the illumination ambiguity in the upsampled results.

Intensity similarity loss: To enforce a generated UI-HR image $\hat{h}_i$ to be similar to its ground-truth image h_i in terms of intensities, an intensity similarity loss L_{mse} is employed:

$$\begin{aligned} L_{mse} &= \mathbb{E}_{(\hat{h}_i,h_i)\sim p(\hat{h},h)}\left\|\hat{h}_i-h_i\right\|_F^2 \\ &= \mathbb{E}_{(l_i,h_i)\sim p(l,h)}\left\|C_t\left(l_i\right)-h_i\right\|_F^2, \end{aligned} \tag{3}$$

where t and C are the parameters and the output of CPUN. l_i represents the input NI-LR face image. $p(\hat{h},h)$ represents the joint distribution of the generated UI-HR images $\hat{h}_i$ and the corresponding ground-truths h_i. Similarly, $p(l,h)$ represents the joint distribution of the input NI-LR images l_i and the corresponding ground-truths h_i.

As mention in [22], [62], only employing the intensity similarity loss L_{mse} in training often leads to overly smoothed results and the network may fail to generate high-frequency facial features (see Fig. 7(f)). Therefore, we incorporate an identity similarity loss to enhance our hallucinated results.

Identity similarity loss: Identity preservation is one of the most important goals in face hallucination [68]. Therefore, we adopt the identity similarity loss L_{id} to minimize the Euclidean distance between the high-level features of a hallucinated face and its ground-truth, thus endowing our Re-CPGAN with the identity preserving ability. The identity similarity loss L_{id} is expressed as:

$$\begin{aligned} L_{id} &= \mathbb{E}_{(\hat{h}_i,h_i)\sim p(\hat{h},h)}\left\|\Phi\left(\hat{h}_i\right)-\Phi\left(h_i\right)\right\|_F^2 \\ &= \mathbb{E}_{(l_i,h_i)\sim p(l,h)}\left\|\Phi\left(C_t\left(l_i\right)\right)-\Phi\left(h_i\right)\right\|_F^2, \end{aligned} \tag{4}$$

where $\Phi(\cdot)$ represents a feature representation of an input image extracted from the average pooling layer of the pre-trained ArcFace model [69]. As shown in Fig. 7(g), employing L_{id} indeed improves the generated results while producing more authentic facial details.

Structure similarity loss: To facilitate face alignment as well as constrain the structural consistency between the generated UI-HR image and the ground-truth one, the structure similarity loss L_h [7] is employed in training our heatmap estimation module (HEM), written as:

$$\begin{aligned} L_h &= \mathbb{E}_{(l_i,h_i)\sim p(l,h)}\frac{1}{P}\sum_{k=1}^{P}\left\|H^k\left(f_i\right)-H^k\left(h_i\right)\right\|_2^2 \\ &= \mathbb{E}_{(l_i,h_i)\sim p(l,h)}\frac{1}{P}\sum_{k=1}^{P}\left\|H^k\left(\tilde{C}_t\left(l_i\right)\right)-H^k\left(h_i\right)\right\|_2^2, \end{aligned} \tag{5}$$

where $H^k(f_i)$ represents the k-th predicted facial landmark heatmap estimated from the intermediate facial features f_i by a stacked hourglass module [64]. $H^k(h_i)$ denotes the k-th facial landmark heatmap generated by FAN [70] on the ground-truth image h_i. Here, we use 68 point facial landmarks to produce the ground-truth heatmaps.

Adversarial loss: Aiming at generating photo-realistic results, we infuse the discriminative information into our CPUN by adopting a discriminative network. Our goal is to make the discriminative network fail to distinguish hallucinated faces from ground-truth ones. The objective function L_D for the discriminative network is defined as follows:

$$L_D = -\mathbb{E}_{\left(\hat{h}_i,h_i\right)\sim p(\hat{h},h)}\left[\log D_d\left(h_i\right)+\log\left(1-D_d\left(\hat{h}_i\right)\right)\right], \tag{6}$$

where D and d represent the discriminative network and its parameters. During training, we update the parameters of the discriminative network by minimizing the loss L_D.

On the contrary, our CPUN is designed to produce realistic face images, which would be classified as real faces by the discriminative network. Thus, the corresponding adversarial loss L_{adv} is represented as:

$$\begin{aligned} L_{adv} &= -\mathbb{E}_{\hat{h}_i\sim p(\hat{h})}\log\left(D_d\left(\hat{h}_i\right)\right) \\ &= -\mathbb{E}_{l_i\sim p(l)}\log\left(D_d\left(C_t\left(l_i\right)\right)\right). \end{aligned} \tag{7}$$

To optimize the CPUN, we minimize the loss L_{adv}.

Total loss function: The objective function to hallucinate UI-HR face $\hat{h}$ is expressed as:

$$L_G = L_{mse}+\alpha L_{id}+\beta L_h+L_{ic}+\psi L_{adv}. \tag{8}$$

Since we intend to hallucinate UI-HR faces rather than generating random faces, we put lower weights on L_{id}, L_h and L_{adv}. Therefore, α,β and ψ in Eq. (8) are set to 0.01.

4 Data Augmentation

4.1 Illumination Rendering Engine: RaIN

Training a deep neural network often requires a large number of data to prevent over-fitting. However, existing public face datasets [9], [10] do not provide a sufficient number of NI/UI face pairs. To achieve enough training samples and improve the generalization ability of our network, we propose a tailor-made Random Adaptive Instance Normalization (RaIN) model as an illumination rendering engine for data augmentation.

RaIN adopts an encoder-decoder architecture. The encoder is a pre-trained VGG-19 network [67] and the first few layers (up to relu4_1) in the encoder are fixed during training. Moreover, RaIN employs an Adaptive Instance Normalization (AdaIN) [11] layer and a Variational Auto-Encoder (VAE) [12] in the latent space. To be specific, AdaIN normalizes the features of a UI face image

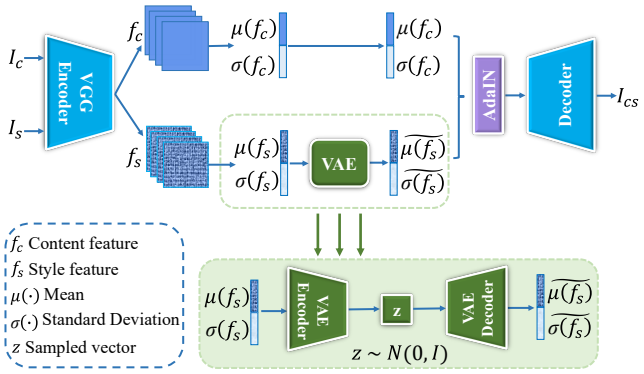


Fig. 8: The framework of our proposed RaIN model.

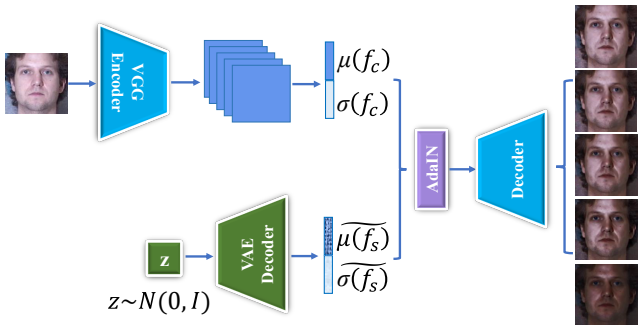


Fig. 9: The testing stage of our RaIN model. RAIN enables us to generate sufficient NI faces with random illumination conditions from some sampled vectors z.

and then enforces the features to share the same channel-wise mean and standard deviation as those of a selected NI face image features. The VAE, as shown in Fig. 8, is inserted before the AdaIN layer to produce a large number of plausible hypotheses for the feature statistics of the NI face image. Our RaIN model transfers illumination styles of face images in real-time and generates sufficient NI face samples from UI inputs (Fig. 8).

Fig. 8 illustrates that RaIN learns to render NI face images. First, given an input content image I_c (UI face) and a style image I_s (NI face), the VGG encoder encodes them into a common latent space, producing f_c and f_s. Then, the VAE first encodes $\mu(f_s) \oplus \sigma(f_s)$ ($\oplus$ denotes "concatenation") to a Gaussian distribution, and then decodes a sampled latent code z from the distribution to reconstruct of the style feature statistics. Afterwards, AdaIN adaptively normalizes f_c as follows:

$$t = \text{AdaIN}(f_c, f_s) = \widetilde{\sigma(f_s)} \left(\frac{f_c - \mu(f_c)}{\sigma(f_c)} \right) + \widetilde{\mu(f_s)}, \tag{9}$$

where $\mu(.)$ and $\sigma(.)$ denote the channel-wise mean and standard deviation. We simply scale the normalized content features f_c with $\widetilde{\sigma(f_s)}$, and shift it with $\widetilde{\mu(f_s)}$.

Finally, a randomly initialized decoder W is trained to map t back to the image space, generating a NI face image I_{cs}:

$$I_{cs} = W(t). \tag{10}$$

4.2 Training Settings

We first train our RaIN model using the MS-COCO[4] [71] and WikiArt[5] [72] databases as content and style images, respectively. Each database contains approximately 80,000 images. Then, we fine-tune the trained RaIN model on the Multi-PIE database [9]. Similar to [11], we employ a content loss and a style loss [11] to train the decoder D, and employ a MSE loss and a Kullback-Leibler divergence loss [12] to train the VAE.

4.3 Data Generation

We employ a pre-trained RaIN model as our "illumination rendering engine" and perform data augmentation. As shown in Fig. 9, we feed the UI face along with a random noise into the trained RaIN model. As a result, we can generate sufficient NI face images with random illumination conditions from an input UI one (see Fig. 10). We resize the synthesized NI face samples to 128×128 pixels and then apply 2D transforms, including rotations, translations, scaling and downsampling, to generate the NI-LR images. Meanwhile, we resize the corresponding UI ones to 128×128 pixels, and use them as our UI-HR images. Since illumination rendering would lead to hue shifts, the color tone of the generated samples is slightly different from the face images in Multi-PIE. However, when we downsample the generated NI faces to form the NI-LR ones, this color jittering can be largely reduced. As a result, we construct a large corpus of NI-LR/UI-HR face pairs. We will release the synthesized face pairs for academic and commercial applications.

In our work, the constructed NI-LR/UI-HR face pairs by RaIN are used to augment our training set. According to the comparison experiments, with data augmentation, our Re-CPGAN is able to hallucinate NI-LR faces even better (see Figs. 7(j) and (l)).

5 Experiments

5.1 Databases

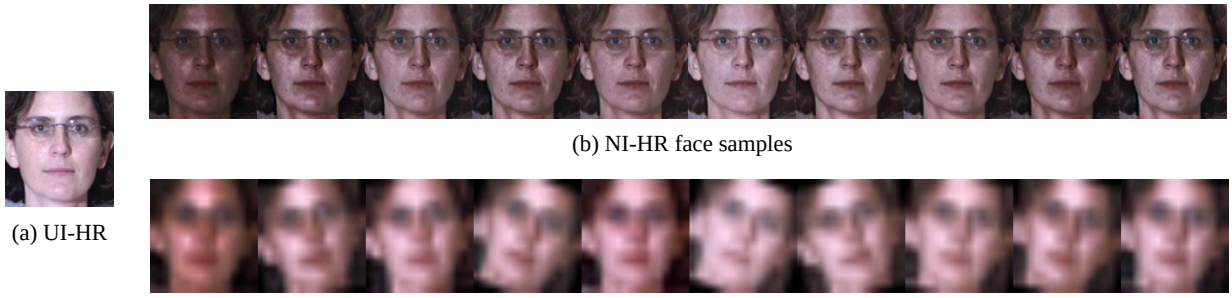
Re-CPGAN is trained and tested on the Multi-PIE database [9] (indoor) and the Celebrity Face Attribute (CelebA) database [10] (in-the-wild).

The **Multi-PIE** database [9] is a large face database with 750K+ images of 337 subjects under various poses, illumination conditions and expressions. We choose 16K NI/UI face pairs of all the subjects spanning across various illumination conditions, poses (0^o, $\pm 15°$, $\pm 30°$, $\pm 45°$, $\pm 60°$, $\pm 75°$, $\pm 90°$) and expressions ("smile", "disgust", "squint", "scream", "surprise", and "neutral"), for our experiments.

The **CelebA** database [10] only provides in-the-wild faces rather than NI/UI face pairs. Therefore, for the training purpose, we opt to synthesize NI faces for the UI ones. We first randomly select 18K cropped UI-HR faces from CelebA, resize them to 128×128 pixels, and use them as our ground-truth images. Then, similar to [73], the Adobe Photoshop Lightroom is adopted to render illumination on these UI-HR faces. Afterwards, we generate the unaligned NI-LR faces (8×8 / 16×16 pixels) by transforming and downsampling the rendered ones. As a result, we generate 18K NI-LR/UI-HR CelebA face pairs.

For each database, we choose 80 percent of the face pairs for training and 20 percent of the face pairs for testing, respectively. In this way, the training and testing sets do not overlap. Specifically,

4. http://cocodataset.org/home
5. https://www.wikiart.org/



(a) UI-HR

(b) NI-HR face samples

(c) Interpolated unaligned NI-LR face samples

Fig. 10: Illustration of the generated NI faces. (a) The UI face (original UI-HR face in Multi-PIE). (b) The generated NI face samples of (a). (c) The spatially transformed and downsampled versions of (b).

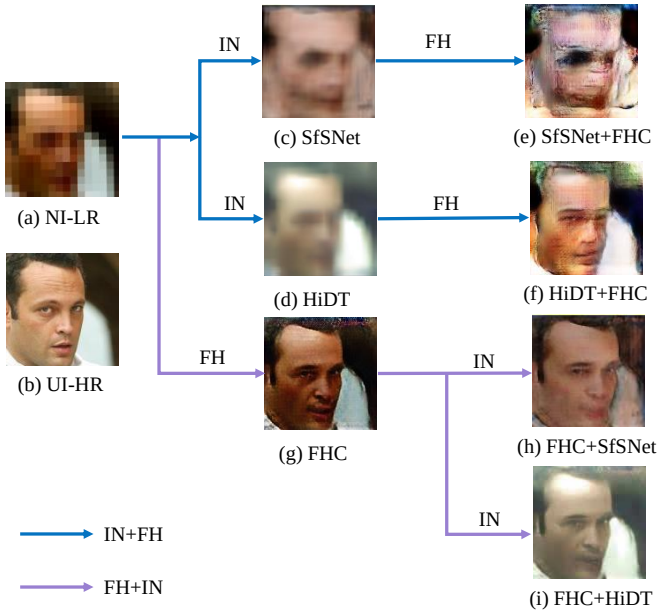


Fig. 11: Results of different combinations of face hallucination and illumination compensation methods. (a) Unaligned NI-LR inputs (16×16 pixels). (b) Ground-truth UI-HR images (128×128 pixels). (c) Bicubic interpolation + SfSNet [3]. (d) Bicubic interpolation + HiDT [43]. (e) SfSNet [3] + FHC [4]. (f) HiDT [43] + FHC [4]. (g) FHC [4]. (h) FHC [4] + SfSNet [3]. (i) FHC [4] + HiDT [43].

RaIN is adopted to perform data augmentation on the training set 10 times. During the training and testing processes, the external guided UI-HR images are *randomly* selected from the UI-HR ones. Our large-scale NI and UI face pair dataset, and the code will be available on https://github.com/SEU-yang.

5.2 Compared Methods

We conduct comparative experiments in the following four scenarios:

- FH: face hallucination methods (SRGAN [62], FSRnet [33], FHC [4]);
- IN+FH: illumination normalization techniques (HiDT [43]) followed by face hallucination methods (SRGAN [62], FSRnet [33] or FHC [4]) (we first upsample the NI-LR face images by bicubic interpolation, then apply [43], and downsample the normalized results for face hallucination);
- FH+IN: face hallucination methods (SRGAN [62], FSRnet [33] or FHC [4]) followed by illumination normalization techniques (SfSNet [3]);
- Joint FH+IN: CPGAN [5] and our Re-CPGAN with 16 Ex-CP units.

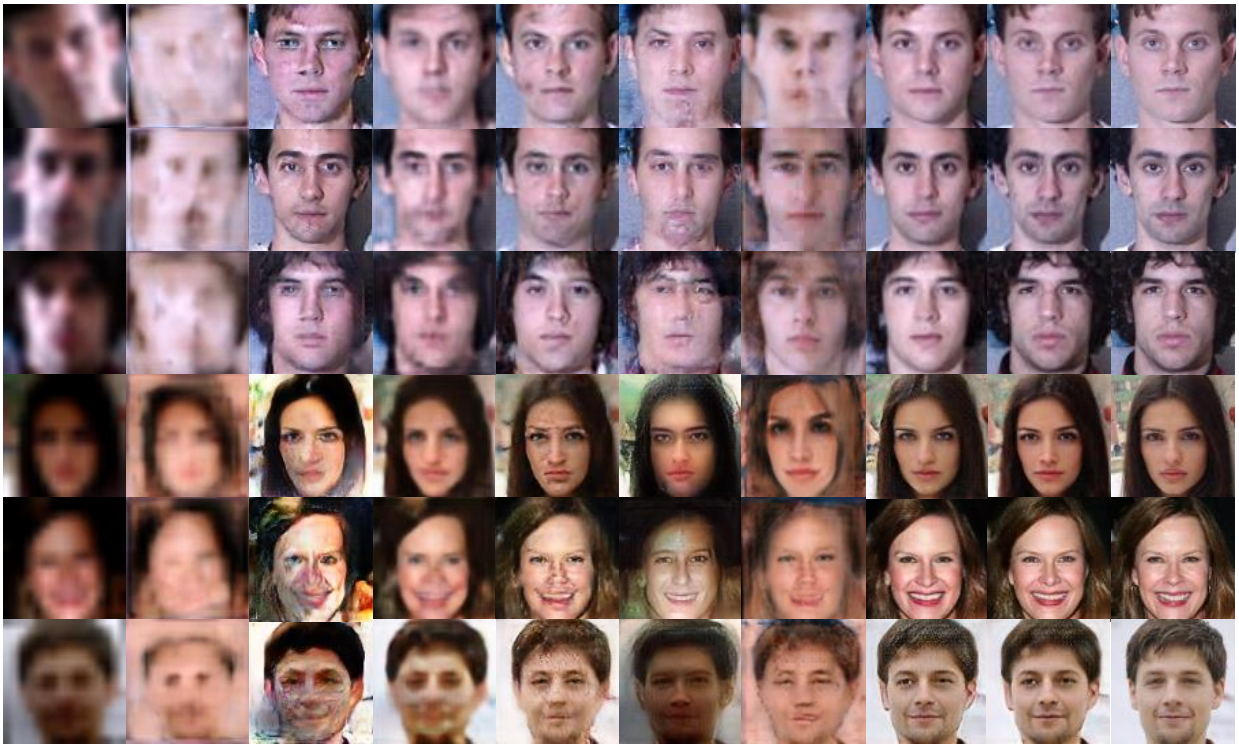
In the first fashion (FH), we hallucinate the NI-LR faces by state-of-the-art face hallucination methods directly. In the second fashion (IN+FH), we first normalize the NI-LR faces by popular illumination normalization techniques, and then hallucinate the normalized faces by state-of-the-art face hallucination methods. In the third fashion (FH+IN), we first hallucinate the NI-LR faces to achieve the NI-HR ones, and then normalize the hallucinated results. In the fourth fashion (Joint FH+IN), both CPGAN [5] and Re-CPGAN jointly tackle face hallucination and illumination normalization in a unified framework. As illustrated in Fig. 11, to achieve the best visual performance among various combinations, we employ HiDT [43] as the face illumination normalization technique in our IN+FH methods. Meanwhile, we employ SfSNet [3] as the face illumination normalization technique in our FH+IN methods.

For a fair comparison, we retrain these methods on our training sets. Since SRGAN [62] and FSRnet [33] cannot achieve face alignment during their upsampling procedure, we train an STN [60] to align the input unaligned LR faces to the upright position firstly. In contrast, FHC [4], CPGAN [5] and our Re-CPGAN super-resolve LR faces while aligning them.

5.3 Qualitative Comparisons with the SOTA

Fig. 12 illustrates the qualitative results of the compared methods. The hallucinated results obtained by Re-CPGAN are more photo-realistic and identity-preserving. As illustrated in Fig. 12(b), the combination of bicubic interpolation and face illumination compensation techniques [3] fails to generate photo-realistic facial details. Since bicubic upsampling only interpolates new pixels from neighboring pixels without generating new contents, the produced NI-HR images lack details. Consequently, the illumination compensation method fails to detect facial landmarks and thus outputs faces with severe artifacts and distorted contours.

SRGAN [62] is a generic super-resolution method and employs the framework of GAN [66] to improve the visual quality. Since non-uniform illumination induces more ambiguous mappings between NI-LR and UI-HR face images, SRGAN may not



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 12: Comparison with state-of-the-art methods (16×16, $8\times$). Columns: (a) Interpolated unaligned NI-LR inputs. (b) Bicubic interpolation + SfSNet [3]. (c) SRGAN [62]. (d) FSRnet [33]. (e) FHC [4]. (f) HiDT [43] + FHC [4]. (g) FHC [4] + SfSNet [3]. (h) CPGAN [5]. (i) Re-CPGAN. (j) Ground-truth UI-HR images (128×128 pixels). The first four columns: testing samples from **Multi-PIE**. The last four columns: testing samples from **CelebA**.

fully address those ambiguities. Hence, the super-resolved faces by SRGAN suffer blurriness and artifacts, as shown in Fig. 12(c).

FSRnet [33] incorporates facial geometry priors into the super-resolution of LR faces. FHC [4] exploits facial component information to encourage the upsampling stream to produce photo-realistic HR faces. These face hallucination methods hallucinate high-frequency facial details with the help of facial structure priors. However, uneven illumination of input faces would degenerate the performance of facial prior estimation, and inaccurate facial priors may provide misleading information in face hallucination. As shown in Figs. 12(d) and (e), those methods produce blurry face structures and deteriorated facial details.

As aforementioned, simply combining existing face hallucination and illumination normalization methods cannot address this challenging issue. This is verified by the results of the strategy IN+FH (see Fig. 12(f)), where upsampled face regions suffer severe distortions and ghosting artifacts. Similarly, the strategy FH+IN also fails to recover authentic facial details, as visible in Fig. 12(g).

CPGAN [5] is the first attempt to jointly address face hallucination and illumination compensation in a whole framework. In this manner, two tasks facilitate each other mutually. Therefore, CPGAN generates satisfying results, as shown in Fig. 12(h). However, due to the limited network capacity of CPGAN, its performance is restricted. Motivated by this, we improve the network capacity by introducing recursive learning into CPGAN without increasing network parameters, known as Re-CPGAN. Our Re-CPGAN generates visually more appealing UI-HR faces from very LR inputs with uneven illumination. As visible in Fig. 12(i), the facial parts covered by shading artifacts in the third and fourth rows, such as the mouth and jaw, are better reconstructed compared to CPGAN.

5.4 Quantitative Comparisons with the SOTA

To evaluate the super-resolution performance quantitatively, we report the average Peak Single-to-Noise Ratio (PSNR), Structural SIMilarity (SSIM) values as well as average Facial Landmark Localization Error (FLLE) on the entire testing set in Tab. 1. FLLE measures the Euclidean distance between the estimated facial landmarks and the ground-truth ones. We employ a state-of-the-art face alignment method, *i.e.*, FAN [70], to detect 68 point facial landmarks.

As shown in Tab. 1, our Re-CPGAN achieves remarkably better quantitative results than other state-of-the-art methods on both indoor and in-the-wild databases. Specifically, on the Multi-PIE testing set, Re-CPGAN outperforms the second best method CPGAN with a large margin of 0.57 dB in PSNR. This is mainly because our external CPnet progressively improves super-resolution results via a recursive fashion. As a result, the hallucinated faces by Re-CPGAN are more similar to the ground-truths. Furthermore, Tab. 1 also demonstrates that both IN+FH and FH+IN methods fail to achieve satisfying quantitative performance. This implies that jointly addressing face hallucination and illumination compensation is more suitable and effective for this challenging task.

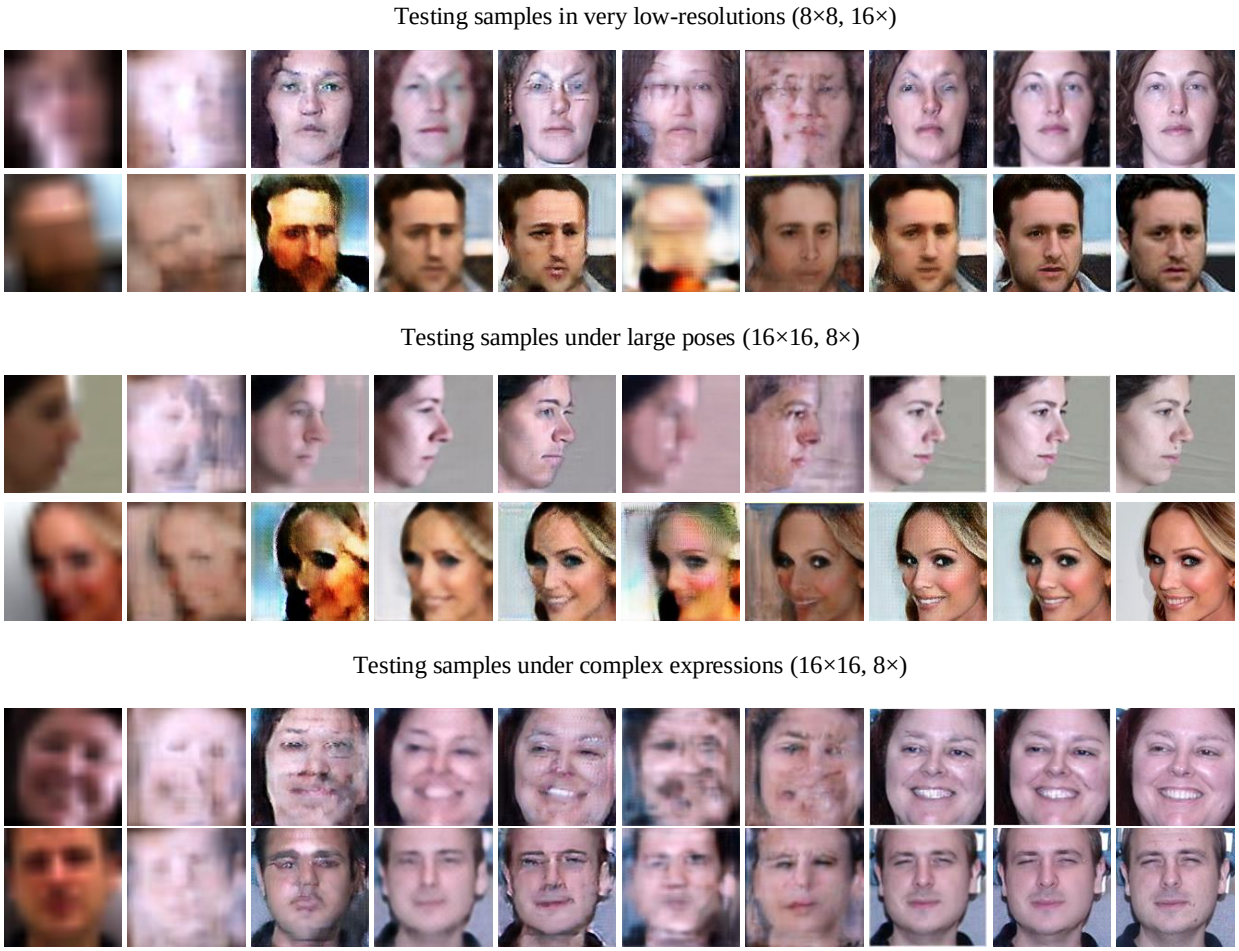


Fig. 13: Comparison with state-of-the-art methods on very challenging situations. Columns: (a) Interpolated unaligned NI-LR inputs. (b) Bicubic interpolation + SfSNet [3]. (c) SRGAN [62]. (d) FSRnet [33]. (e) FHC [4]. (f) HiDT [43] + FHC [4]. (g) FHC [4] + SfSNet [3]. (h) CPGAN [5]. (i) Re-CPGAN. (j) Ground-truth UI-HR images (128×128 pixels).

5.5 Comparisons of Upsampling Very LR Faces

We evaluate our method qualitatively on very low-resolution face images (8×8 pixels) with a $16\times$ magnification factor, compared with other state-of-the-art methods. This is a very challenging case for face hallucination because 16×16 pixels will be recovered from a single pixel and an input NI-LR image only has 8×8 pixels.

As shown in Fig. 13, our Re-CPGAN achieves pleasant hallucination performance on such challenging images (see the first two lines in Fig. 13(i)). However, the results of the state-of-the-art methods deviate from the ground-truth appearance severely. On the contrary, our Re-CPGAN recovers authentic global structures and local details. Moreover, the joint face hallucination and illumination compensation mechanism performed in Re-CPGAN significantly reduces artifacts.

5.6 Robustness Towards Poses and Expressions

We also evaluate our method qualitatively on the NI-LR faces under large poses and complex expressions. Since these input faces are accompanied by not only non-uniform illumination but also self-occlusions, it is challenging to normalize and super-resolve them.

Fig. 13 shows that the state-of-the-art methods fail to reconstruct plausible UI-HR faces, where the edges are blurry and the structures are distorted. In contrast, our Re-CPGAN achieves superior performance when the input LR faces undergo large poses (*e.g.,* 90^o) and complex facial expressions (*e.g.,* "smile", "disgust"). This demonstrates the robustness of our method towards pose and facial expression variations. Since recursive learning is applied to the Ex-CP unit, our Re-CPGAN is able to recover diverse characteristics of NI-LR inputs progressively.

5.7 Ablation Study

5.7.1 Impact of Increasing Recursion Depths

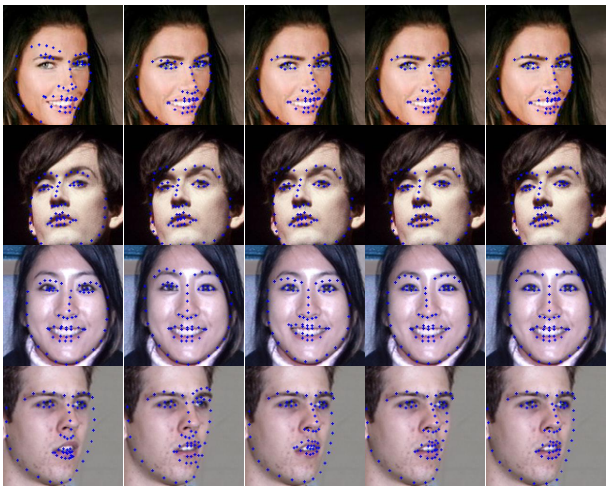
To study the effect of recursion depths, Re-CPGAN variants with different numbers of recursions (*e.g.,* 1, 4, 8, 12, 16) are compared.

Facial landmarks estimation: Here, we study the impacts of recursion depths on our heatmap estimation module. Fig. 14 shows that the estimated facial landmarks are more precise as the recursion depth increases. We attribute such better performance to the deeper model structure introduced by more recursion operations.

Face hallucination: We also investigate the effect of recursion depths on face hallucination performance qualitatively and quantitatively. As shown in Fig. 6(b), hallucinated faces become more

TABLE 1: Average PSNR [dB], SSIM and FLLE results of compared methods on the testing sets (16×16, $8\times$).

Method	Multi-PIE			CelebA		
	FH					
	PSNR	SSIM	FLLE	PSNR	SSIM	FLLE
Bicubic	12.838	0.385	21.312	12.794	0.377	22.156
SRGAN	16.769	0.546	10.361	17.951	0.506	10.083
FSRnet	19.342	0.611	7.972	19.854	0.587	8.926
FHC	20.680	0.634	6.709	21.130	0.652	7.841
Method	Multi-PIE			CelebA		
	IN [43]+FH					
	PSNR	SSIM	FLLE	PSNR	SSIM	FLLE
Bicubic	13.315	0.399	20.891	13.018	0.404	21.074
SRGAN	15.044	0.486	13.102	16.512	0.491	11.277
FSRnet	15.810	0.497	13.336	18.124	0.518	9.902
FHC	18.263	0.595	8.409	19.508	0.579	8.947
Method	Multi-PIE			CelebA		
	FH+IN [3]					
	PSNR	SSIM	FLLE	PSNR	SSIM	FLLE
Bicubic	12.960	0.387	19.405	12.945	0.392	21.898
SRGAN	14.252	0.448	14.126	15.237	0.448	12.055
FSRnet	15.449	0.491	13.228	15.825	0.453	12.953
FHC	17.554	0.582	8.977	16.459	0.490	11.280
Method	Multi-PIE			CelebA		
	Joint FH+IN					
	PSNR	SSIM	FLLE	PSNR	SSIM	FLLE
CPGAN	24.639	0.778	3.654	23.972	0.723	3.991
Re-CPGAN	**25.211**	**0.794**	**2.762**	**24.348**	**0.759**	**2.934**



(a) (b) (c) (d) (e)

Fig. 14: Facial landmark estimation by Re-CPGAN variants on CelebA and Multi-PIE faces. (a) Re-CPGAN-r1. (b) Re-CPGAN-r4. (c) Re-CPGAN-r8. (d) Re-CPGAN-r12. (e) Re-CPGAN-r16. Here, Re-CPGAN-rN means the Re-CPGAN variant with N EX-CP units. Please zoom in to see the improvements.

photo-realistic when we apply more recursions to Re-CPGAN. It indicates that increasing recursion depths not only leads to a deep yet compact model but also boosts performance. Moreover, Fig. 15 demonstrates that the quantitative results improve as more recursions are performed. It is noteworthy that the performance gains become negligible when continuing to increase the recursion depth more than 10.

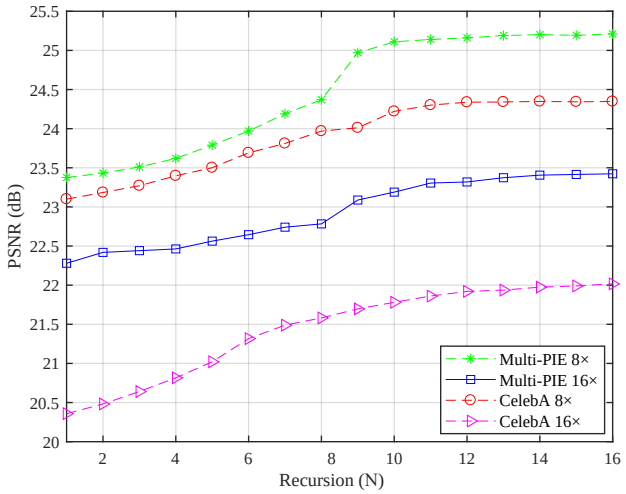


Fig. 15: Impacts of recursion depths on CelebA and Multi-PIE testing sets.

TABLE 2: Ablation study of different sub-networks on CelebA and Multi-PIE databases (16×16, $8\times$).

w/o CPnet	Multi-PIE		CelebA	
	PSNR	SSIM	PSNR	SSIM
w/o IN	23.215	0.752	22.679	0.715
w/o EX	22.890	0.733	22.151	0.709
w/o D	23.287	0.756	22.908	0.722
Re-CPGAN	**25.211**	**0.794**	**24.348**	**0.759**

5.7.2 *Impacts of Internal CPnet*

As indicated by the quantitative results of the IN+FH and FH+IN combination methods in Tab. 1, simply combining existing face hallucination and illumination compensation methods leads to sub-optimal hallucination performance. Our Re-CPGAN embeds an internal CPnet to initially offset non-uniform illumination by exploiting the internal guidance, and it is able to reduce the ambiguous mapping caused by shading artifacts in CPUN, thus facilitating the latter upsampling operations.

To evaluate the impact of the internal CPnet, we replace it with a convolutional layer and use this convolutional layer to encode NI-LR faces. As shown in Tab. 2, the performance of only using the input convolutional layer, marked by w/o IN, degrades almost 2 dB in terms of PSNR on Multi-PIE. Moreover, as shown in Fig. 7(c), the Re-CPGAN variant without the internal CPnet produces flawed results with obvious distortions and blurred artifacts. It implies that the internal CPnet recovers high-frequency facial details from non-uniform illumination and thus improves face hallucination performance.

5.7.3 *Impacts of Recursive External CPnet*

Compared to our previous work [5], Re-CPGAN does not require more network parameters or wider architectures for better performance. As illustrated in Fig. 3, we design the recursive external CPNet to normalize illumination and enhance facial details at the feature level. Benefiting from the recursive learning, we improve the face hallucination performance significantly in Re-CPGAN compared to CPGAN and do not introduce new parameters for additional layers.

TABLE 3: Ablation study of different losses (16×16, $8\times$)

		Multi-PIE		CelebA	
		PSNR	SSIM	PSNR	SSIM
w/o L_{ic}	L_G^-	22.013	0.704	21.174	0.652
	$L_G^\dagger$	22.732	0.729	21.755	0.698
	$L_G^\ddagger$	23.148	0.743	22.108	0.707
	L_G	23.759	0.780	22.855	0.720
w/ L_{ic}	L_G^-	22.640	0.717	22.043	0.705
	$L_G^\dagger$	23.287	0.756	22.908	0.722
	$L_G^\ddagger$	23.944	0.782	23.276	0.738
	L_G	**25.211**	**0.794**	**24.348**	**0.759**

We remove the recursive external CPNet and then feed the output of the internal CPnet to the face reconstruction net directly, and this variant is marked as w/o EX, in Tab. 2. As demonstrated in Tab. 2, the performance of w/o EX degrades 2.32 dB compared to our Re-CPGAN on Multi-PIE. It implies that the recursive external CPnet recovers facial details authentically. Furthermore, without using the recursive external CPnet, the reconstructed results suffer ghosting artifacts, such as blurry edges, as seen Fig. 7(d). It also demonstrates that the recursive external CPnet plays a crucial role in our model.

5.7.4 Impacts of Different Losses

We report the performance of Re-CPGAN variants that are trained with different loss combinations on Multi-PIE and CelebA (see Tab. 3 and Fig. 7). We denote the compared loss combinations as follows: (i) L_G^-: L_{mse} and L_h; (ii) $L_G^\dagger$: L_{mse}, L_{id} and L_h; (iii) $L_G^\ddagger$: L_{mse}, L_{id} and L_{adv}; (iv) L_G: L_{mse}, L_{id}, L_h and L_{adv}. Note that L_h is a prerequisite objective in training our heatmap estimation module.

As demonstrated in Tab. 3, using our illumination compensation loss (L_{ic}) leads to better quantitative results. Because L_{ic} constrains the illumination characteristics of the reconstructed UI-HR faces to be close to the guided UI-HR ones in the latent subspace, the hallucinated faces will achieve uniform illumination, thus being similar to their UI-HR ground-truths. Only employing the intensity similarity loss L_{mse} leads to unpleasant results (L_G^- in Tab. 3). The identity similarity loss (L_{id}) not only improves the visual quality as seen in Fig. 7(g)), but also increases the quantitative performance ($L_G^\dagger$ as indicated in Tab. 3). This experiment demonstrates that L_{id} forces the high-level features of the hallucinated faces, *i.e.,* identity information, to be similar to their ground-truth counterparts and thus improves hallucination performance. In addition, we also verify the effectiveness of the structure similarity loss (L_h). As indicated in Tab. 3 ($L_G^\ddagger$), removing L_h leads to degraded quantitative performance since L_h enforces the structure of hallucinated faces to resemble their ground-truths. The adversarial loss is used to enforce upsampled faces not only to be realistic but also to exhibit in a well-illuminated condition. Hence, with the help of the adversarial loss (L_{adv}), Re-CPGAN achieves photo-realistic face images (see Fig. 7(l)) and the highest quantitative results (*i.e.*, L_G in Tab. 3).

5.7.5 Impacts of Guided Faces

Our method employs an external guided UI-HR face for illumination compensation. We design the external CPnet to learn illumination features from a guided face, and propose an illumination compensation loss to enforce the illumination characteristics of the reconstructed UI-HR face to be similar to those of the guided face in the latent subspace. In this fashion, we offset non-uniform illuminations of input NI-LR faces under the guidance of external lighting information rather than only depending on internal information. As shown in Fig. 7(i), the face hallucination performance of Re-CPGAN degrades without using the guided face. Note that we remove the guided face branch and retrain the entire network.

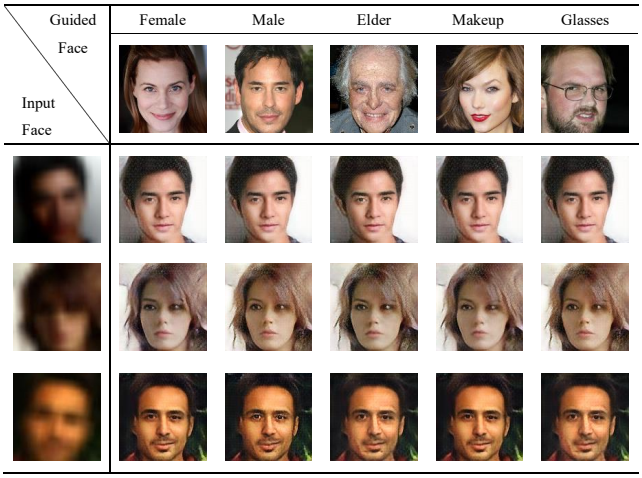


Fig. 16: The hallucinated results of our Re-CPGAN with different guided faces.

Then, we explore the impact of different guided faces on our method. As seen in Fig. 16, by using guided faces with different facial attributes (*e.g.,* "gender", "age", "makeup", and "glasses"), we still obtain photo-realistic results and the hallucinated faces are not affected by different facial attributes of guided faces. The is mainly because we employ guided faces including different variations in training Re-CPGAN and it learns to focus on the illumination style rather than other attributes.

Furthermore, Fig. 16 also implies that our method can generate identity-preserving results regardless of the identities of guided faces. The reasons are as follows: (1) Since our copy block is designed to explicitly learn the illumination pattern from an external guided UI-HR face under the supervision of the illumination compensation loss, it only adopts illumination features from the guided images rather than the facial contents. (2) We apply the identity similarity loss to hallucinated faces, thus enabling our Re-CPGAN to preserve identity information.

6 Discussion

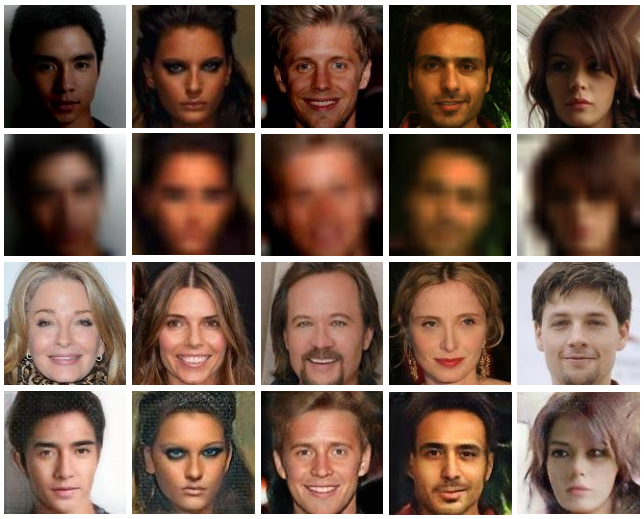
6.1 Comparisons with SOTA on UI Faces

As shown in Fig. 2(d), the illumination processing method [3] is not suitable to remove illuminations of NI-LR faces and lead to artifacts in normalized results. Then, the produced artifacts would affect the performance of face hallucination methods (see Fig. 2(e)). In contrast, we jointly remove illuminations and hallucinate faces instead of treating these two task separately. Therefore, we significantly alleviate side-effects caused by either of these two processes.

Furthermore, we conduct experiments to evaluate Re-CPGAN and the state-of-the-art face hallucination methods on the LR faces with normal illumination. In this case, the illumination normalization methods are not necessary and thus not employed.

(a) (b) (c) (d) (e) (f) (g) (h) (i)

Fig. 17: Comparison with state-of-the-art methods. Columns: (a) UI-LR inputs (16×16 pixels). (b) Bicubic interpolation. (c) FSRnet [33]. (d) SRGAN [62]. (e) WaveletSRnet [26]. (f) FHC [4]. (g) CPGAN [5]. (h) Re-CPGAN. (i) Ground-truth UI-HR images (128×128 pixels).



(a) (b) (c) (d) (e)

Fig. 18: Results of super-resolving NI-LR face images on CelebA (16×16, $8\times$). First row: real NI-HR images. Second row: unaligned NI-LR images. Third row: guided UI-HR images. Fourth row: our normalized and hallucinated results.

As shown in Fig. 17, Re-CPGAN still outperforms the state-of-the-art methods. Note that, our previous method CPGAN [5] intends to increase the network depth to achieve better hallucination performance but is limited by the GPU memory. In this work, we employ a recursive external CPnet in Re-CPGAN and thus achieve a deep yet compact model. This also demonstrates that the effectiveness of our recursive learning for the external CPnet.

6.2 Performance on Real NI-LR Faces

Although our model is trained on the synthesized NI/UI CelebA face pairs, our method can effectively hallucinate the faces under real illumination conditions. To demonstrate this, we randomly select face images with real non-uniform illumination from CelebA excluding the faces used for generating our training dataset. Then, we obtain the NI-LR face samples (16×16 pixels) by transforming and downsampling these images. These NI-LR faces do not share illumination styles with the examples in the training dataset, and thus these samples are much more challenging. As shown in Fig. 18, our Re-CPGAN achieves superior normalization and hallucination performance on such randomly chosen images. Therefore, our Re-CPGAN is not restricted to certain illumination styles.

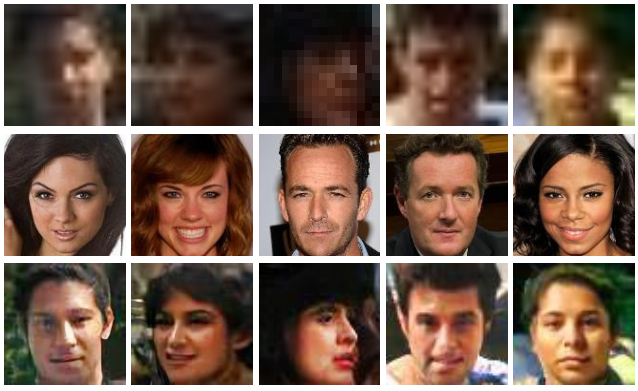
Moreover, we also evaluate our model on real-world NI-LR faces. To do so, we randomly choose face images from the Widerface database [74] for testing. Widerface contains in-the-wild faces which are affected by various degradation, illumination and noise types. Here, our Re-CPGAN model is trained on the CelebA training set. As seen in Fig. 19, our Re-CPGAN not only hallucinates the visually appealing HR faces but also reduces the shading artifacts.

6.3 Comparisons with the SOTA on Face Recognition

We demonstrate that our Re-CPGAN boosts the performance of low-resolution face recognition. We adopt the "recognition via hallucination" framework to conduct face recognition experiments on the **Multi-PIE** [9] database. Concretely, aggressively downsampled faces are first hallucinated by face hallucination methods and then used for recognition.

Experimental Settings: First, we partition the **Multi-PIE database** [9] into subject disjoint training and testing sets. Then, we train the compared face hallucination methods on the training set and then conduct face recognition experiments on the testing set. The testing set includes the NI-LR/UI-HR Multi-PIE face pairs of 50 testing individuals under 10 illumination conditions. Here, for all the testing images, we crop the aligned face regions, resize them to 128×128 pixels, and thus generate our HR face images. We generate the NI-LR faces (16×16 pixels) by transforming and downsampling the NI-HR ones. Then, the NI-LR and UI-HR face images construct the probe and gallery sets respectively. Afterwards, we employ a state-of-the-art pre-trained face recognition model (SphereFaceNet [75]) to conduct face recognition experiments on LR faces and hallucinated HR faces from LR ones by different methods. Finally, we compute the cosine distance of the extracted deep features for face recognition. In particular, we train a CycleGAN [41] to alleviate the domain gap between gallery faces and hallucinated ones.

Evaluation: The performance comparisons of Re-CPGAN and other face hallucination methods are shown in Tab. 4. Here, we compare the performance of face hallucination methods as well as the combinations of face hallucination and illumination compensation. As indicated by Tab. 4, the face recognition rates of our hallucinated UI-HR faces are superior to those of the



(a) (b) (c) (d) (e)

Fig. 19: Results of upsampling NI-LR face images on Widerface (16×16, $8\times$). First row: real NI-LR images. Second row: guided UI-HR images. Third row: our normalized and hallucinated results.

TABLE 4: Face recognition performance comparison on the Multi-PIE database.

FH method	Accuracy		
	FH	IN+FH	FH+IN
Bicubic	59.60%	49.84%	48.12%
SRGAN [62]	62.04%	52.31%	51.79%
FSRnet [33]	63.15%	53.62%	52.83%
FHC [4]	65.29%	54.47%	53.06%
CPGAN [5]	84.36%		
NI-LR	61.21%		
UI-HR	98.13%		
Re-CPGAN	87.45%		

NI-LR faces and other methods' results. This demonstrates that our Re-CPGAN achieves remarkable identity preservation ability, which substantially satisfies the need of the downstream face recognition task. Moreover, we can see that direct combinations of face hallucination and illumination compensation achieve inferior performance compared to face hallucination methods. This also implies that it is more reasonable to take these two tasks in a unified framework.

6.4 Comparisons with the SOTA on Face Expression Classification

Furthermore, we manifest that our Re-CPGAN also benefits low-resolution face expression classification tasks.

Experimental Settings: We perform a standard 10-fold subject-independent cross-validation on the **Multi-PIE expression dataset** [9]. First, the NI-LR/UI-HR Multi-PIE face pairs with complex expressions, *i.e.,* "smile", "disgust", "squint", "scream", "surprise", and "neutral", are split into 10 subsets according to the identity information and the individuals in any two subsets are mutually exclusive. In each experiment, 9 subsets are used for training and the remaining one for testing. We train all the compared hallucination models on the same training database and employ a state-of-the-art expression classification model, VGG-VD-16 [67], to identify the facial expressions of UI-HR faces hallucinated from NI-LR ones. Here, state-of-the-art face hallucination methods are used to upsample the testing faces, while the classification results of the NI-LR faces upsampled by bicubic interpolation and the ground-truth UI-HR faces are also provided as baselines. At last, the expression classification performance for each method is obtained by averaging the results of the 10 folds, as indicated in Tab. 5.

TABLE 5: Face expression classification results for different methods on the Multi-PIE expression dataset.

FH method	Accuracy		
	FH	IN+FH	FH+IN
Bicubic	60.94%	51.26%	50.35%
SRGAN [62]	64.32%	55.14%	54.49%
FSRnet [33]	66.55%	57.60%	56.72%
FHC [4]	70.61%	60.08%	58.17%
CPGAN [5]	88.13%		
NI-LR	62.64%		
UI-HR	97.92%		
Re-CPGAN	93.57%		

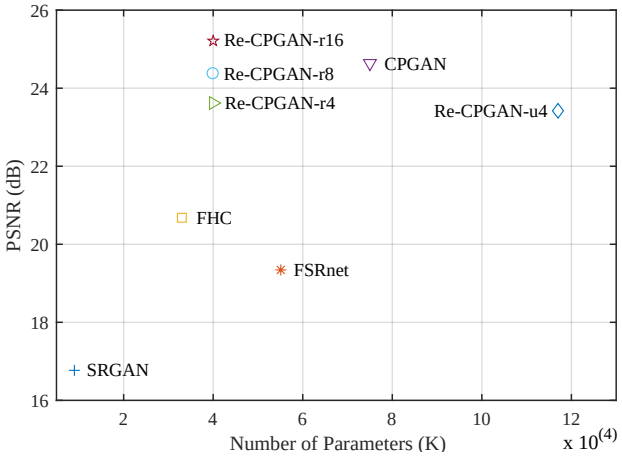


Fig. 20: PSNR values and parameter numbers of compared models on Multi-PIE (16×16, $8\times$). Here, Re-CPGAN-rN represents the Re-CPGAN variant with N Ex-CP units, and Re-CPGAN-uN represents the Re-CPGAN variant which directly stacks N Ex-CP units without using recursive learning.

Evaluation: As indicated in Tab. 5, the upsampled faces of our method achieves superior face expression classification rates. This also demonstrates that the hallucinated face images of our Re-CPGAN are more authentic to the ground-truth UI-HR faces in comparison to the state-of-the-art. Particularly, the face expression classification rate of our hallucinated faces exceeds that of the NI-LR ones by a large margin of 30.93%. Hence, Re-CPGAN indeed facilitates the low-resolution face expression classification task.

6.5 Model Size Analyses

As indicated in Fig. 20, the parameters of Re-CPGAN are much smaller than CPGAN [5]. Meanwhile, we achieve improvements on the quantitative performance, as indicated by Fig. 20. Since recursive learning is employed to construct a very deep network, we effectively reuse network parameters and boost the model performance. Note that, increasing the recursion depths, our model parameters do not increase. In contrast, simply stacking the CPnet

will increase parameters dramatically but does not necessarily lead to superior performance.

6.6 Limitations

Since our work uses the bilinear downsampling to generate NI-LR faces artificially, we do not contain real-world degradation types (*e.g.,* motion blur, compression artefacts, sensor noise) in the training dataset. Therefore, we produce relative blurry results when applied to real-world NI-LR images (see Fig. 19). This deterioration is mainly caused by the significant domain shift between LR face data. We will address the domain shift caused by different image degradation factors in face hallucination as our future work.

7 Conclusion

In this paper, we presented a recursive copy and paste generative adversarial network (Re-CPGAN) to jointly hallucinate the NI-LR face images and compensate for the non-uniform illumination. With the internal and recursive external CPNets, our method progressively upsample and refine facial features based on the spatial distribution of facial structure up to a magnification factor of 16×. In particular, Re-CPGAN offsets non-uniform illumination and upsamples facial details alternatingly. Meanwhile, the RaIN model is presented to generate sufficient face pairs under diverse illumination conditions. Our RaIN model not only significantly enriches our training dataset but also improves the generalization of our Re-CPGAN. Extensive results demonstrate that our Re-CPGAN produces HR identity-preserving face images and substantially boosts the performance of downstream tasks, *i.e.,* face recognition and expression classification. This makes our Re-CPGAN more desirable in practice.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No.61871123, No.61976017) and Key Research and Development Program in Jiangsu Province (No.BE2016739). This work was supported in part by Australian Research Council under Grant DP180100106 and DP200101328.

References

[1] X. Chen, M. Chen, J. Xin, and Q. Zhao, "Face illumination transfer through edge-preserving filters," in *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, 2011.

[2] H. Zhou, J. Sun, Y. Yacoob, and D. W. Jacobs, "Label denoising adversarial network (ldan) for inverse lighting of faces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[3] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs, "Sfsnet: Learning shape, reflectance and illuminance of faces 'in the wild'," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, "Face super-resolution guided by facial component heatmaps," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2018, pp. 217–233.

[5] Y. Zhang, I. W. Tsang, Y. Luo, C.-H. Hu, X. Lu, and X. Yu, "Copy and paste gan: Face hallucination from shaded thumbnails," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[6] X. Yu, F. Porikli, B. Fernando, and R. Hartley, "Hallucinating unaligned face images by multiscale transformative discriminative networks," *International Journal of Computer Vision*, pp. 1–27, 2019.

[7] X. Yu, B. Fernando, R. Hartley, and F. Porikli, "Super-resolving very low-resolution face images with supplementary attributes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 908–917.

[8] X. Yu, F. Shiri, B. Ghanem, and F. Porikli, "Can we see more? joint frontalization and hallucination of unaligned tiny faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2148–2164, 2019.

[9] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[10] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015, p. 3730–3738.

[11] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2017, pp. 1501–1510.

[12] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2013, pp. 1–14.

[13] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 35, no. 3, pp. 425–434, 2005.

[14] C. Liu, H.-Y. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *International Journal of Computer Vision*, vol. 75, no. 1, pp. 115–134, 2007.

[15] S. Kolouri and G. K. Rohde, "Transport-based single frame super resolution of very low resolution face images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4876–4884.

[16] X. Ma, J. Zhang, and C. Qi, "Hallucinating face by position-patch," *Pattern Recognition*, vol. 43, no. 6, pp. 2224–2236, 2010.

[17] R. A. Farrugia and C. Guillemot, "Face hallucination using linear models of coupled sparse support," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4562–4577, 2017.

[18] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Face super-resolution via multi-layer locality-constrained iterative neighbor embedding and intermediate dictionary learning," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4220–4231, 2014.

[19] L. Liu, C. P. Chen, S. Li, Y. Y. Tang, and L. Chen, "Robust face hallucination via locality-constrained bi-layer representation," *IEEE Transactions on Cybernetics*, vol. 48, no. 4, pp. 1189–1201, 2017.

[20] M. F. Tappen and C. Liu, "A bayesian approach to alignment-based image hallucination," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2012, pp. 236–249.

[21] C.-Y. Yang, S. Liu, and M.-H. Yang, "Hallucinating compressed face images," *International Journal of Computer Vision*, vol. 126, no. 6, pp. 597–614, 2018.

[22] X. Yu and F. Porikli, "Imagining the unimaginable faces by deconvolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2747–2761, 2018.

[23] X. Yu, F. Xu, S. Zhang, and L. Zhang, "Efficient patch-wise non-uniform deblurring for a single image," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1510–1524, 2014.

[24] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2016, pp. 318–333.

[25] ——, "Face hallucination with tiny unaligned images by transformative discriminative neural networks," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017, p. 4327–4333.

[26] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet domain generative adversarial network for multi-scale face hallucination," *International Journal of Computer Vision*, vol. 127, no. 6-7, pp. 763–784, 2019.

[27] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li, "Attention-aware face hallucination via deep reinforcement learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 690–698.

[28] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, "Learning to super-resolve blurry face and text images," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2017, pp. 251–260.

[29] R. Dahl, M. Norouzi, and J. Shlens, "Pixel recursive super resolution," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2017, pp. 5439–5448.

[30] A. V. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proceedings of Machine Learning Research*, vol. 48, 2016, pp. 1747–1756.

[31] S. Menon, A. Damian, M. Hu, N. Ravi, and C. Rudin, "Pulse: Self-supervised photo upsampling via latent space exploration of generative

models," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[32] Y. Zhang, I. W. Tsang, Y. Luo, C.-H. Hu, X. Lu, and X. Yu, "Copy and paste gan: Face hallucination from shaded thumbnails," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7355–7364.

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

[34] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a gan to learn how to do image degradation first," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2018, pp. 185–200.

[35] A. Shashua and T. Riklin-Raviv, "The quotient image: Class-based re-rendering and recognition with varying illuminations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 129–139, 2001.

[36] Z. Liu, Y. Shan, and Z. Zhang, "Expressive expression mapping with ratio images," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001, pp. 271–276.

[37] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, "Face swapping: automatically replacing faces in photographs," in *ACM SIGGRAPH 2008 papers*, 2008, pp. 1–8.

[38] B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *Amfg*, vol. 19, no. 6, pp. 1635–1650, 2010.

[39] L. Q. Tran, X. Yin, and X. Liu, "Representation learning by rotating your faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[40] Y. Zhang, C. Hu, and X. Lu, "Il-gan: Illumination-invariant representation learning for single sample face recognition." *J. Vis. Commun. Image Represent.*, vol. 59, pp. 501–513, 2019.

[41] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2223–2232.

[42] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz, "Few-shot unsupervised image-to-image translation," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2019, pp. 10 551–10 560.

[43] I. Anokhin, P. Solovev, D. Korzhenkov, A. Kharlamov, T. Khakhulin, A. Silvestrov, S. Nikolenko, V. Lempitsky, and G. Sterkin, "High-resolution daytime translation without domain labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7488–7497.

[44] P. N. Belhumeur and D. J. Kriegman, "What is the set of images of an object under all possible illumination conditions?" *International Journal of Computer Vision*, vol. 28, no. 3, pp. 245–260, 1998.

[45] V. Blanz, T. Vetter *et al.*, "A morphable model for the synthesis of 3d faces." in *Siggraph*, vol. 99, 1999, pp. 187–194.

[46] Z. Wang, X. Yu, M. Lu, Q. Wang, C. Qian, and F. Xu, "Single image portrait relighting via explicit multiple reflectance channel modeling," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–13, 2020.

[47] V. Blanz, T. Vetter, and A. Rockwood, "A morphable model for the synthesis of 3d faces," *Acm Siggraph*, pp. 187–194, 2002.

[48] Y. Wang, L. Zhang, Z. Liu, G. Hua, Z. Wen, Z. Zhang, and D. Samaras, "Face relighting from a single image under arbitrary unknown lighting conditions." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1968–1984, 2009.

[49] J. T. Barron and J. Malik, "Shape, illumination, and reflectance from shading." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1670–1687, 2015.

[50] Z. Shu, S. Hadap, E. Shechtman, K. Sunkavalli, S. Paris, and D. Samaras, "Portrait lighting transfer using a mass transport approach," *ACM Transactions on Graphics*, vol. 37, no. 1, p. 2, 2018.

[51] S. Saito, L. Wei, L. Hu, K. Nagano, and H. Li, "Photorealistic facial texture inference using deep neural networks." in *CVPR*. IEEE Computer Society, 2017, pp. 2326–2335.

[52] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras, "Neural face editing with intrinsic image disentangling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5541–5550.

[53] P. H. O. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling." in *ICML*, ser. JMLR Workshop and Conference Proceedings, vol. 32. JMLR.org, 2014, pp. 82–90.

[54] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3367–3375.

[55] T. Ying, Y. Jian, and X. Liu, "Image super-resolution via deep recursive residual network," in *IEEE Computer Vision & Pattern Recognition*, 2017.

[56] Y. Zheng, X. Yu, M. Liu, and S. Zhang, "Single image deraining via recurrent residual multiscale networks," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2020.

[57] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3d object classification," in *Advances in neural information processing systems*, 2012, pp. 656–664.

[58] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution." in *CVPR*. IEEE Computer Society, 2016, pp. 1637–1645.

[59] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2018, pp. 286–301.

[60] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in Neural Information Processing Systems(NIPS)*, 2015, pp. 2017–2025.

[61] X. Yu, F. Porikli, B. Fernando, and R. Hartley, "Hallucinating unaligned face images by multiscale transformative discriminative networks," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 500–526, 2020.

[62] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4681–4690.

[63] A. Bulat and G. Tzimiropoulos, "Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 109–117.

[64] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2016, pp. 483–499.

[65] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803.

[66] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 2672–2680.

[67] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2015, pp. 1–14.

[68] F. Shiri, X. Yu, F. Porikli, R. Hartley, and P. Koniusz, "Identity-preserving face recovery from stylized portraits," *International Journal of Computer Vision*, vol. 127, no. 6-7, pp. 863–883, 2019.

[69] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[70] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2017, pp. 1021–1030.

[71] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 740–755.

[72] F. Phillips and B. Mackintosh, "Wiki art gallery, inc.: A case for critical thinking," *Issues in Accounting Education*, vol. 26, no. 3, pp. 593–608, 2011.

[73] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style aggregated network for facial landmark detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 379–388.

[74] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5525–5533.

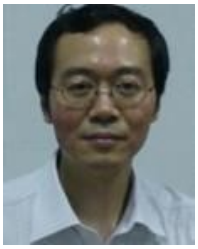
[75] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Yang Zhang received her B.S. degree in Chongqing University of Posts and Telecommunications in 2013. She is currently working toward the Ph.D. degree with the School of Automation, Southeast University, Nanjing, China. She was a visiting student with the Australian Institute of Artificial Intelligence, University of Technology Sydney, Sydney, Australia, under the supervision of Prof. Ivor W.Tsang. She is the recipient of the China National Scholarship in 2019 and the Nanjing AI Project Scholarship in 2018. Her current research interests include computer vision and image processing.

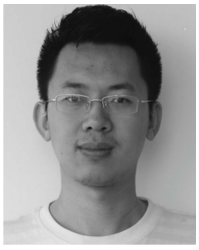
Ivor W.Tsang is an ARC Future Fellow and Professor of Artificial Intelligence with the University of Technology Sydney, Australia. He is also the Research Director of the Australian Artificial Intelligence Institute. In 2013, Prof Tsang received his prestigious ARC Future Fellowship for his research regarding Machine Learning on Big Data. In 2019, his JMLR paper titled "Towards ultrahigh dimensional feature selection for big data" received the International Consortium of Chinese Mathematicians Best Paper Award. In 2020, Prof Tsang was recognized as the AI 2000 AAAI/IJCAI Most Influential Scholar in Australia for his outstanding contributions to the field of AAAI/IJCAI between 2009 and 2019. His research on transfer learning granted him the Best Student Paper Award at CVPR 2010 and the 2014 IEEE TMM Prize Paper Award. In addition, he received the IEEE TNN Outstanding 2004 Paper Award in 2007. He serves as a Senior Area Chair/Area Chair for NeurIPS, ICML, AISTATS, AAAI and IJCAI, and the Editorial Board for JMLR, MLJ, JAIR and IEEE TPAMI.

Yawei Luo received his B.S. degree and Ph.D degree in Huazhong University of Science and Technology in 2013 and 2020, respectively. He was a visiting student at ReLER Lab in University of Technology Sydney for two years. He is the recipient of the China National Scholarship in 2016. He is currently a Postdoc in College of Computer Science and Technology, Zhejiang University. His work focuses on computer vision, domain adaptation, 3D reconstruction and semantic segmentation.

Changhui Hu received the B.S. degree in automation from Huazhong University of Science and Technology Wuchang Branch, Wuhan, China, in 2005, the M.S. degree in power electronic and the Ph.D. degree with the School of Automation, Southeast University, Nanjing, China, in 2017. He is currently a lecturer with the College of Automation and the College of Artificial Intelligent, Nanjing University of Posts and Telecommunications, Nanjing, China. His current research interests include image processing, face recognition, and pattern recognition.



Xiaobo Lu received the B.S. degree from Shanghai Jiao Tong University, Shanghai, China, the M.S. degree from Southeast University, Nanjing, China, and the Ph.D. degree from Nanjing University of Aeronautics and Astronautics. He did his postdoctoral research with Chien-Shiung Wu Laboratory, Southeast University, from 1998 to 2000. He is currently a Professor with the School of Automation and deputy Director of the Detection Technology and Automation Research Institute, Southeast University. He is a coauthor of the book An Introduction to the Intelligent Transportation Systems (Beijing, China Communications, 2008). His research interests include image processing, signal processing, pattern recognition, and computer vision. Dr. Lu has received many research awards, such as the First Prize in Natural Science Award from the Ministry of Education of China and the prize in the Science and Technology Award of Jiangsu province.



Xin Yu received his B.S. degree in Electronic Engineering from University of Electronic Science and Technology of China, Chengdu, China, in 2009, and received his Ph.D. degree in the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2015. He also received a Ph.D. degree in the College of Engineering and Computer Science, Australian National University, Canberra, Australia, in 2019. He is currently a Lecturer at University of Technology Sydney. His interests include computer vision and image processing.