IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 January 2023, accepted 28 January 2023, date of publication 1 February 2023, date of current version 7 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3241606

RESEARCH ARTICLE

Joint Super-Resolution and Head Pose Estimation for Extreme Low-Resolution Faces

SAHAR RAHIMI MALAKSHAN, (Student Member, IEEE), MOHAMMAD SAEED EBRAHIMI SAADABADI, (Student Member, IEEE), MOKTARI MOSTOFA, (Student Member, IEEE), SOBHAN SOLEYMANI, (Member, IEEE), AND NASSER M. NASRABADI, (Fellow, IEEE)

Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, USA

Corresponding author: Sahar Rahimi Malakshan (sr00033@mix.wvu.edu)

ABSTRACT State-of-the-art deep learning-based Head Pose Estimation (HPE) techniques have reached spectacular performance on High-Resolution (HR) face images. However, they still fail to achieve expected performance on low-resolution images at large scales. This work presents an end-to-end HPE framework assisted by a Face Super-Resolution (FSR) algorithm. The proposed FSR model is specifically guided to enhance the HPE performance rather than considering FSR as an independent task. To this end, we utilized a Multi-Stage Generative Adversarial Network (MSGAN) which benefit from a pose-aware adversarial loss and head pose estimation feedback to generate super-resolved images that are properly aligned for HPE. Also, we propose a degradation strategy rather than simple down-sampling approach to mimic the diverse properties of real-world Low-Resolution (LR) images. We evaluate the performance of our proposed method on both synthetic and real-world LR datasets and show the superiority of our approach in both visual and HPE metrics on the AFLW2000, BIWI, and WiderFace Datasets.

INDEX TERMS Head pose estimation (HPE), face super-resolution (FSR), multi-stage generative adversarial networks (MSGAN), low-resolution (LR) face images.

I. INTRODUCTION

Single image Head Pose Estimation (HPE) plays a significant role in applications such as 3D face modeling, gaze direction detection, driver monitoring safety systems, surveillance face recognition, and face frontalization [1], [2], [3], [4], [5]. Existing HPE methods can be divided into 1) landmark-based and 2) landmark-free approaches. The landmark-based methods rely on accurate landmarks localization and use mean human head models to solve a 2D to 3D correspondence problem [6], [7]. Since accurate landmark localization is hard to achieve for unconstrained images, landmark-based approaches have severe limitations when used in real-world settings [8]. In the landmark-free approaches, the primary measure is to use a deep learning method to extract global representation from the input image [8], [9], [10]. These approaches leverage Convolutional Neural Networks (CNNs) to directly regress head Euler angles of a given face image [1], [9], [10], [11]. Despite the remarkable performance improvement made by landmark-free methods [9], [11], HPE in the real world (unconstrained) scenarios is far from optimal [9], [11].

One of the main reasons for the performance gap between constrained and unconstrained HPE is the lack of

FIGURE 1. Top: the samples from the 300WLP dataset (tarining dataset). Bottom: instances in the WiderFace dataset (evaluation dataset).

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

in-the-wild variations in the available training datasets, i.e., few low-quality images are present in the training datasets [12], [13]. Consequently, the network fails to explore these under-represented samples since they are less frequent, see Fig. 1 [14]. Generally, face image quality reflects the understandable nuisance factors in the face image such as pose-angle, illumination, distortion, and resolution [15], [16]. However, task-related image quality is the key to boosting the deep model performance [17]. For instance, head pose angle is regarded as an undesirable quality factor for a face recognition network [18], [19]. However, estimating head pose angle is the goal of a HPE algorithm and it is not considered as an undesirable factor of an input image.

Among the image characteristics that can affect image quality, the resolution is one of the primary challenges in HPE [9]. It is known that Low-Resolution (LR) synthetic data can mimic real-world LR images and increase the resolution diversity of the training dataset [9], [20], [21]. Therefore, training HPE module with synthetic LR data may narrow the performance gap between the high and low-resolution HPE [9], [11]. However, using synthesized LR samples, i.e., data augmentation during training of an HPE module, is a trade-off between positive gain from more diverse training instances and adverse effects of overfitting [12], [13]. The authors in [9] show that when the resolution variation increases, the performance on the original High-Resolution (HR) samples drops. Consequently, the model which is not trained with the augmentation achieves better performance on the HR samples by a considerable margin. Hence, the cost of improving performance on the LR data is the reduction of performance on HR samples, which is not desired.

Another possible approach to address the LR HPE is to utilize Super-Resolution (SR) techniques to retrieve the HR face images from their corresponding LR pairs [22], [23]. Among image generation models, the Generative Adversarial Networks (GANs) have shown incredible performance in producing images with high-frequency details, especially for tasks involving image-to-image translation such as SR [20], [24], [25], [26], [27]. However, the State-of-the-Art (SOTA) Face Super Resolution (FSR) methods did not consider face images with a wide range of pose variations, which is the most important aspect in designing an HPE module [22], [23], [26], [28]. Consequently, these methods failed to produce satisfactory results for the off-angle faces and added undesirable artifacts to the synthesized images, which severely lowered the HPE model performance [29]. Moreover, the majority of existing methods have investigated FSR up to $\times 4$ upscaling factor because extremely LR images, *e.g.*, 8×8 and 16×16 pixels, do not retain as much identity-related information as their corresponding HR images [30], [31], [32], [33], [34], [35].

In this work, we develop a guided FSR framework which is specifically designed to enhance the HPE module performance on extremely LR images, see Fig. 2. To this end, we seek to improve both visual and task-related quality of LR images. The former is achieved using reconstruction losses

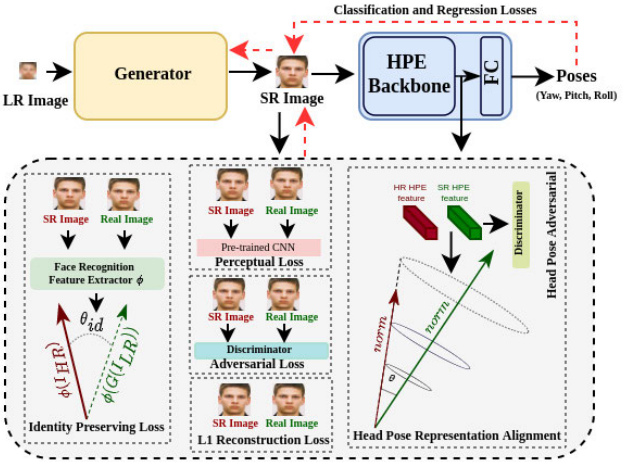

FIGURE 2. Overview of the proposed approach. Task-related (head pose adversarial, representation alignment, classification, and regression losses) and visual quality (reconstruction losses) feedbacks are used to guide the generator toward producing authentic SRd images.

and to satisfy the latter, we reduce the gap between the representation of Super-resolved (SRd) and HR images in terms of both angular and magnitude of HPE embedding. Furthermore, the pose-aware adversarial domain adaptation [36], [37] forces the generator to produces SRd output images that the HPE model cannot decipher from their real HR pairs. Moreover, to expilicitly supervise the HPE performance we use both classification and head pose regression losses.

We explore the effects of the input resolution on the representations obtained from the HPE module from the perspective of the magnitude, and the dimension spanned by the HPE embedding feature vectors. To the best of our knowledge, our proposed joint FSR and HPE is the first study which focuses on HPE for extreme LR images using a large upscaling factor. Our contributions in this paper can be summarized as follows:

1) We explore the effect of the image quality (resolution) on the features extracted from the penultimate layer of the HPE model and propose to use task-related quality measure to guide the FSR module to produce authentic and task-related high quality SRd images.
2) We introduce a novel pose-aware domain adaptation approach to ensure resolution-agnostic HPE.
3) We emperically show the dimension collaps in the embedding of current SOTA HPE models.
4) We introduce a practical degradation model which imitates the real-world LR images.

The rest of this paper is organized in the following manner. In Section II, we provide a literature review related to FSR, HPE and using SR to boost the downstream tasks. In Section III, we present our model and network architecture in detail. In Section IV, we describe the training details and datasets, and then present extensive experimental results, evaluations, and ablations studies to validate the effectiveness of the proposed approach. Finally, we conclude the paper in Section V.

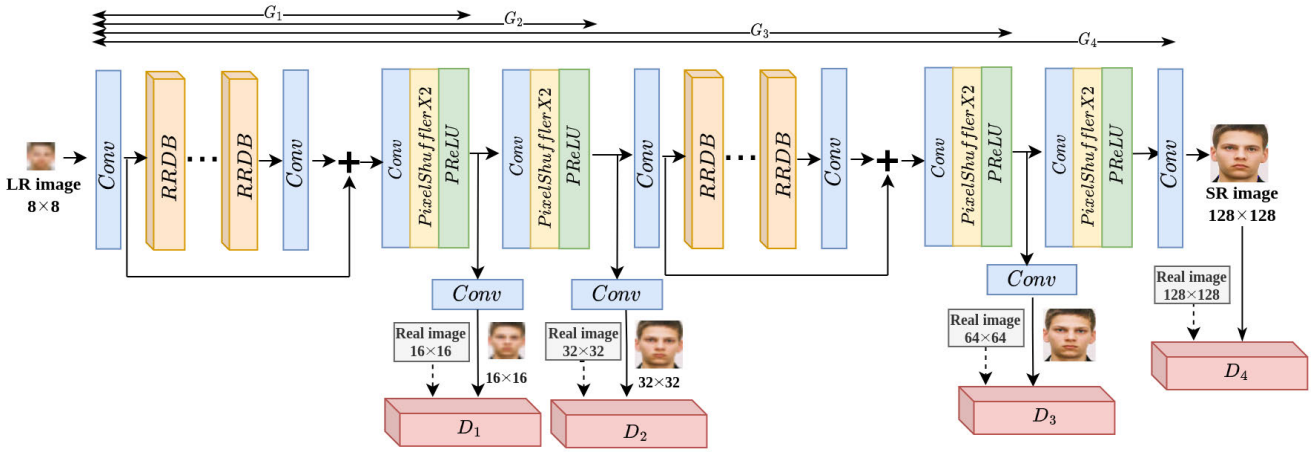

FIGURE 3. The architecture of the proposed Multi-Stage Generative Adversarial Network (MSGAN) for ×16 SR. We adopt Residual-in-Residual Dense Block (RRDB) from [20]. The first half of the network performs ×4 upsampling, and the other half performs the remaining ×4 upsampling.

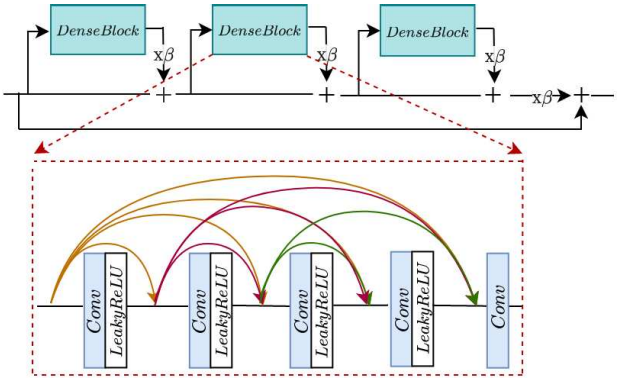

FIGURE 4. RRDB which combines multi-level residual network and dense connections.

II. RELATED WORK

A. HEAD POSE ESTIMATION

Approaches for HPE can be categorized into landmark-based methods, and landmark-free methods [6], [7], [8], [9], [10]. In landmark-based, first, key landmarks are localized, then pose estimation is performed by solving a 2D to 3D correspondences problem [6]. Despite the satisfactory results in constrained benchmarks, landmark-based approaches fail to maintain their performance in uncontrolled settings such as complete profile or LR face images [9]. The main reason for this lack of robustness is the strict reliance on the accurate landmark localization [8]. To address this issue, a large area of research has been dedicated to the landmark-free approaches [1], [9], [10]. Consequently, the landmark-free methods, mainly based on deep learning approaches, have achieved a significant performance improvement [1], [9], [10]. These approaches utilize CNNs to estimate head pose directly from image intensities [6], [7], [11], [38], [39].

Studies demonstrated that simultaneously learning related tasks yields better results than the training individual networks for each task (i.e., multi-tasking) [40], [41], [42], [43], [43]. Kummar et al. [41], intend to learn the global and local features simultaneously via using heatmap-CNN for face detection, and HPE [41]. Ranjan et al. use a CNN to learn HPE, gender classification, face detection, and landmark localization [43]. Recently, different studies use multiple HPE loss functions to train a HPE module, e.g., classification and regression [9]. Liu et al. [38] used a CNN to estimate the three angles of the head pose, but they used solely artificial datasets to train their models, which caused issues when the model was tested in real-world senario.

Patacchiola and Cangelosi [39] assessed the performance of several architectures in combination with an adaptive gradient learning. Ruiz et al. [9] improved the performance of their HPE model by employing both the Mean Squared Error (MSE) and cross-entropy losses. In FSA-Net [11], authors improved the HPE performance with an architecture based on regression and feature aggregation. Despite the significant improvement, mentioned methods fail to preserve their performance in the LR scenarios [9]. Among the SOTA methods, [9] has relative superior performance because of the down-sampling augmentation during the training. However, it obtains this robustness at the cost of losing performance on the original HR input images [9]. Moreover, according to the our investigation in Table 1, the performance of SOTA methods for LR face images (8×8 and 16×16) significantly drops.

B. FACE SUPER RESOLUTION (FSR)

FSR (also known as face hallucination) is a technique for creating a HR face image from a LR source [44]. FSR can be divided into classical and deep learning-based methods [45], [46], [47], [48], [49], [50]. In the classical method, LR images are SRd using global face statistical models [45], [46] or local patch-based representation methods [47], [48]. Although these methods can achieve impressive performances, most of them still suffer from the two drawbacks. First, they usually rely on the complex optimization methods to recover the HR images [51]. Second, they need manually tuned parameters to

obtain a good SR performance, making them inflexible [52]. Several papers provide a thorough review of traditional FSR techniques [22], [53].

Among the deep learning methods, GANs have received special attention [20], [24], [25]. The GAN framework is based on two competing networks. A generator network, G, and a discriminator network, D. In the GAN-based SR, the generator map LR input, I_{LR}, to its corresponding HR pair, I_{HR}. The trainable parameters of the generator are θ_g. On the other hand, the discriminator, $D(.;\theta_d)$, discriminates between the HR and SRd output of the generator using a binary classification. Consequently, G and D are playing a min-max game, and the loss function to train G and D can be formulated as follows:

$$l_{adv} = \min_{G} \max_{D} \{\mathbb{E}\left[\log D(I_{HR})\right] + \mathbb{E}\left[\log\left(1 - D(G(I_{LR}))\right)\right]\}. \tag{1}$$

Since many HR pairs exist for a LR face image, FSR is an inherently ill-posed inverse problem [54]. Therefore, prior information is the key to boosting the quality of the SRd image and GANs training stability [11], [32]. Various FSR studies have explored the benefits of prior facial information in FSR [11], [30], [32], [55], [56]. In ProgFSR [32], a new facial attention loss is developed via multiplication of the heatmap and the differences between SRd and HR images to focus on restoring face traits in more detail. Chen et al. [30] improved the reconstruction model performance by exploiting the prior geometric information extracted from the face images. SuperFAN [56] first SRd the LR images, then employed a prior estimation network to extract the heatmaps of SR and HR images, and constrain the heatmaps of the corresponding SR and HR images to be close. Despite the remarkable achievements, face images in the various poses, which are crucial for creating a HPE model, were not taken into the account by current FSR approaches. Most methods focus on producing visually acceptable output, not the output, which further improves the downstream task. Also, there are a limited number of studies deal with extreme LR input images, such as upscaling factors of $\times 8$ [30], [32] and $\times 16$ [33], [34].

C. SUPER-RESOLUTION TO BOOST DOWNSTREAM TASKS

SR has been shown to be a good preprocessing tool for LR input images in the various vision tasks. Shermeyer et al. [57] fine-tuned a pre-trained SR model to quantify its effect on object detection performance in the multi-resolution satellite pictures. Similarly, the proposed network in [58] jointly optimized the detection and SR modules to generate the SRd images. In [59], the effect of SR on the small-scale pedestrian detection has been extensively studied. Also, in SuperFan [56], joint training improved face SR by incorporating the facial landmark information in a GAN-based SR algorithm. Furthermore, Wu et al. [60] proposed joint face hallucination and recognition in which SR and a face recognition network are optimized iteratively. These works have motivated us to utilize SR framework to develop a resolution-agnostic HPE module. Our work focus on reducing the effects of resolution variation on the HPE for extremely LR and degraded images. To this end, we force the FSR to improve both visual and task-related quality of images

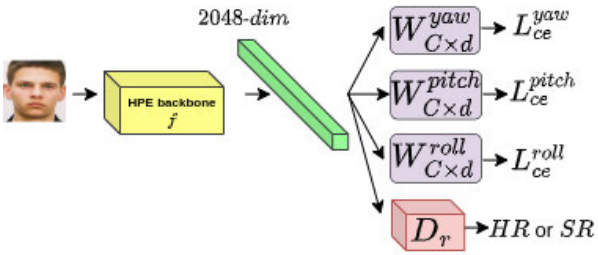

FIGURE 5. Showing the pose aware adversarial training. D_r is responsible to distinguish between the HPE representation of the HR and SRd output of the generator.

III. PROPOSED FRAMEWORK

In this section, we first begin by introducing our multi-stage generator. Then, we further explain the integration of our SR framework with the HPE module. The proposed generator produces the SRd version of an extremely LR input image in a way that is visually convincing and efficiently aligned to improve the HPE module performance. To this end, we integrate three HPE model feedbacks into our GAN framework: 1) Pose-Aware adversarial training, 2) head pose estimation criterion, and 3) aligning the SRd and HR representation of the HPE model in both angular and Euclidean space.

A. MULTI-STAGE GAN (MSGAN)

Our main goal is to map an extremely LR input face image to its corresponding HR counterpart, which will explicitly boosts the performance of HPE. To handle the information gap between the LR and HR data, one can stack more convolutional layers in the encoder-decoder architecture [20], [61]. However, more layers of continuous convolutions result in the loss of information which may be essential for synthesizing the SRd image [20], [62]. Therefore, we decompose the SR problem into multiple stages. Then the network can restore the HR information iteratively [63], [64], [65], [66]. The network learns the HR image distribution at different scales using a coarse-to-fine SR framework in our architecture. With the multi-stage framework, adversarial training is applied in the each stage:

$$l^i_{adv} = \min_{G_i} \max_{D_i} \left\{\mathbb{E}\left[\log D_i(I^i_{HR})\right] + \mathbb{E}\left[\log 1 - D_i(G(I^i_{LR}))\right]\right\}, \tag{2}$$

where $i \in \{1, 2, 3, 4\}$ refers to each stage and D_i/G_i refers to the discriminator/generator at each stage. We adopt the architecture of widely-used (SR network) ESRGAN [20] as our generator, i.e., a deep network with several residual-in-residual dense blocks (RRDB). We also double the original $\times 4$ ESRGAN architecture to have enough network capacity for performing SR with a scale factor of $\times 8$ and $\times 16$,

see Fig. 3. For our discriminator, we use a convolutional "PatchGAN" classifier [67]. Specifically, in conventional discriminators, the decision of being fake/real is made based on the whole input image (one output). However, here in PatchGan discriminator, the decision is made locally. Therefore, instead of one single output, the final output of the discriminator is a feature map indicating which part of the input image is real or fake. The discriminator of the proposed method consists of five consecutive modules of convolution-BatchNorm-ReLu [68]. For more details, please refer to IV-B.

B. HEAD POSE RELATED LOSSES

1) POSE-AWARE ADVERSARIAL TRAINING

Every softmax-based classification framework can be regarded as the stack of non-linear feature extractor layers (backbone) together with a softmax classification layer. The weight of the softmax layer are the centeroids of the classes [18]. Both the backbone and classifier will be trained end-to-end using a back-propagation algorithm. The goal is to increase the similarity between the output of the backbone (feature representation) and its corresponding softmax class centeroid and decrease the similarity with other class centeroids [12]. Given a fixed HPE module, the ideal scenario is to have identical SRd and HR feature representations from the HPE backbone (resolution-agnostic backbone).

To improve the similarity between the HR and SRd feature pair, we adopt the idea of adversarial domain adaptation, which aims at making the representations of two different versions of the same image as similar as possible [37]. To this end, we aim to fool a discriminator, which will be trained to distinguish between the SRd and HR representations obtained from the HPE backbone, as shown in Fig. 5. This mimics the adversarial policy used in the GANs as presented in section II-C [37]. However, the GAN discriminator distinguishes between the SRd and HR images, which is irrelevant to the downstream task. Here, to achieve the goal of resolution-agnostic HPE, we introduce an adversarial regularization strategy, Eq. 3, to guide the SR synthesis process so that the FSR module generates samples specified for the resolution-agnostic HPE:

$$l_{pa} = \min_{G} \max_{D_r} \{\mathbb{E}\left[\log D_r(f(I_{HR}))\right] + \mathbb{E}\left[\log\left(1 - D_r(f(G_4(I_{LR})))\right)\right]\}, \quad (3)$$

where f is the backbone of the HPE model and D_r is discriminating between $f(I_{HR})$ and $f(G_4(I_{LR}))$. At the same time, G_4 (SR generator) tries to generate SRd images suitable for accurate HPE. Applying adversarial training on the representation obtained from the HPE model aligns with our final goal of having resolution-agnostic HPE. In this context, if the features of the HR and SR images are not distinguishable, we can achieve results similar to the HR images from the SR counterparts.

Considering Eq. 2 and Eq. 3, we are explicitly guiding our SR model to produce visually appealing SRd images which can improve the performance of the HPE model on the LR images. Eq. 3 aims to make the representations of SRd and HR images obtained from the HPE model indistinguishable. However, it does not explicitly impose the HPE accuracy error loss. To further impose the low HPE error, we leverage the HPE loss, as it was used in [9], and the representation learning criterion, which we elaborate on in the next two sections.

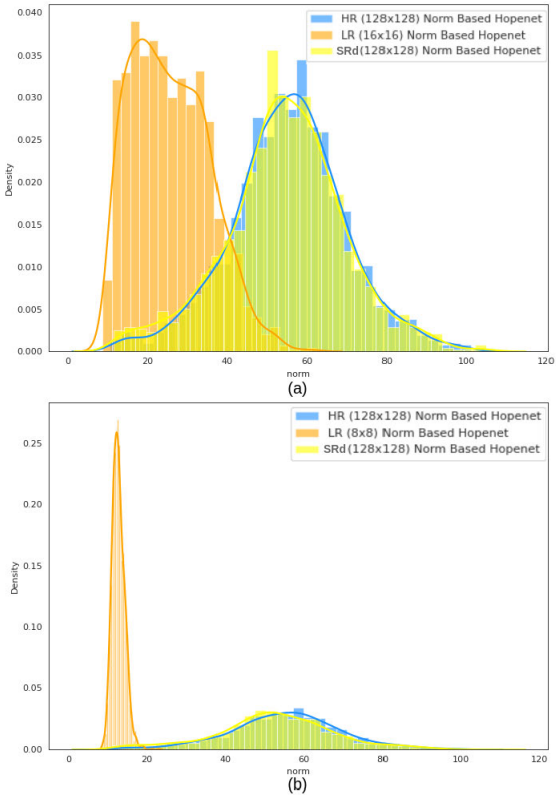

FIGURE 6. Distribution of the l_2 norm of features obtained from the penultimate layer of the HPE module for (a) LR (16 × 16), (b) LR (8 × 8) with their corresponding super-resolved and ground truth HR l_2 norms for the AFLW2000 dataset.

2) HPE LOSS

Assume that the Euler angle's range is $[a, b]$ and is divided into n bins. Then, the Ground Truth (GT) Euler angels, θ_{GT}, fall into one of the bins. Therefore, we can define a one-hot label for each GT and calculate the cross entropy loss L_{ce}. Also, from the softmax probabilities (output), the corresponding predicted head pose angle can be calculated. Eq. 4 shows how to compute the predicted Euler angle (yaw, pitch, and roll) given the softmax output, θ.

$$\hat{\theta} = a + \frac{b-a}{n} \sum_{i=1,\ldots,n} (i-1)\theta_i, \quad (4)$$

where θ_i denotes the output of the softmax layer for i^{th} bin and $\hat{\theta}$ is the predicted angle. From Eq. 4, a regression loss ($L_{mse} = (\hat{\theta} - \theta_{GT})$) is added to derive the fine-grained prediction. The following (estimation loss) is the total pose estimation loss

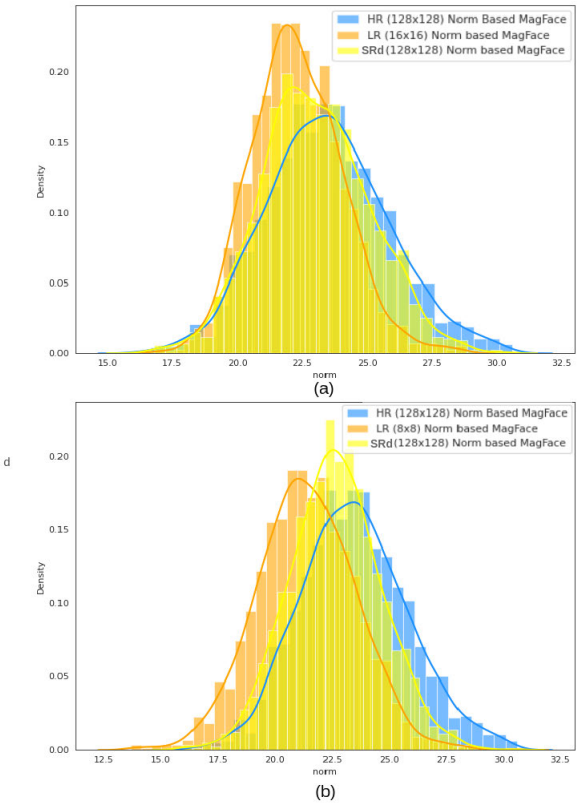

FIGURE 7. Distribution of the l_2 norm of features obtained from MagFace [15] for (a) LR (16 × 16), (b) LR (8 × 8) with their corresponding super-resolved and ground truth HR l_2 norms for the AFLW2000 dataset.

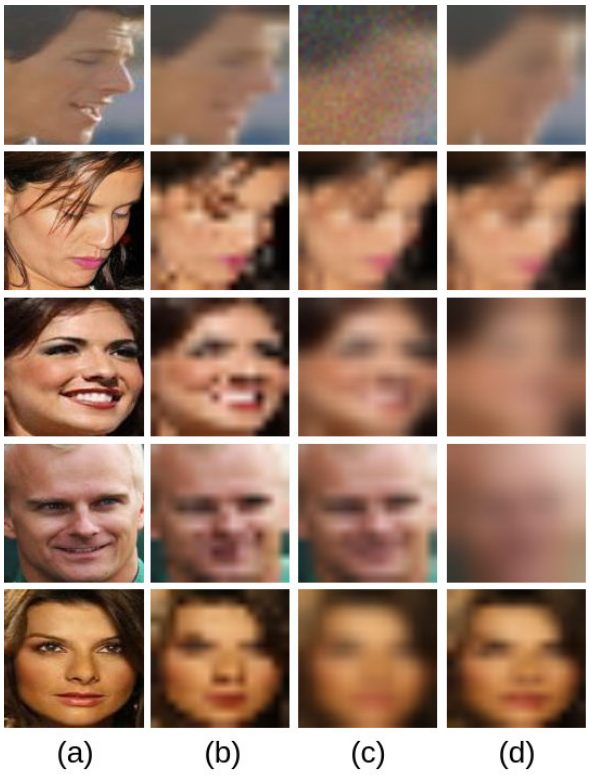

FIGURE 8. Visualization of some degraded samples based on proposed degradation method. Column (a) shows the HR images, (b), (c), and (d) are the different degraded images correspond to each HR sample.

for each Euler angle:

$$L_{est} = L_{ce} + \beta_{mse} L_{mse}, \tag{5}$$

where β_{mse} is the regression coefficient. Note that three final losses for three Euler angles need to be calculated.

3) REPRESENTATION ALIGNMENT

It is widely accepted that representation learned via Softmax has an angular distribution [18]. Also, face recognition literature empirically shows that the l_2 norm of the feature reflects the quality of the input image [12], [13], [15]. Fig. 6 shows the effect of input resolution on the l_2 norm of features obtained from the HPE backbone. Fig. 6 shows that L_2 norm of features of the HPE backbone is severely affected by input resolution. Therefore, considering the angular distribution of the Softmax representations, it is essential to consider the alignment of HR and SRd features in both 1) the angular space and 2) the l_2 norm space. The inner product between the normalized representations from the HR and SRd images can provide the angular similarity. In this regard, we define similarity loss, L_{sim}, in the angular space as:

$$L_{sim} = 1 - \frac{f(I_{HR})}{||f(I_{HR})||} \cdot \frac{f(G_4(I_{LR}))}{||f(G_4(I_{LR}))||}, \tag{6}$$

where $f(.)$ represent the feature vector obtained from the penultimate layer of HPE model, and $||.||$ is the L_2 norm operator.

For adding the norm regularization, we first normalized the feature norm by the batch statistics, μ and σ:

$$\widehat{||f(.)||} = \frac{||f(.)|| - (\mu)}{\sigma}. \tag{7}$$

To relax μ and σ from the batch size, we calculate them in an exponential moving average over the training iterations:

$$\sigma = \alpha\sigma_t + (1-\alpha)\sigma_{t-1}, \tag{8}$$

$$\mu = \alpha\mu_t + (1-\alpha)\mu_{t-1}. \tag{9}$$

Then we define the square root of the difference between $||\widehat{f(I_{HR})}||$ and $||\widehat{f(G(I_{LR})}||$ as the norm regularization loss:

$$L_m = |||\widehat{f(G_4(I_{LR}))}|| - ||\widehat{f(I_{HR})}|||^2. \tag{10}$$

With the mapping of the norm of the feature to the area between [−1,1], Eq. 7, the distribution of the feature norm, $\widehat{||f(.)||}$, is made roughly unit Gaussian. It is known that around 68% of the unit Gaussian distribution lies between −1 and 1. We can scale the σ to map most values to fall between −1 and 1. To this end, we replace the σ with $0.33 * \sigma$ [12]. This is done to make the value of angular and magnitude loss comparable.

Finally, the loss function for guiding the SR network toward increasing the performance of HPE model is as follow:

$$L_p = \lambda_{pa} L_{pa} + \lambda_{est} L_{est} + \lambda_{sim} L_{sim} + \lambda_m L_m. \quad (11)$$

In a softmax-based classifier, representations' angular alignment is crucial for correctly classifying. Consequently, we separately optimize the angle and magnitude of representations. Aligning the magnitude of HPE embedding forces the generator to increase the task-related quality of its output. Without this constraint, the generator only focuses on improving the visual quality of the output.

C. RECONSTRUCTION LOSSES

Due to the importance of identity preservation, the global face shape and local attributes in FSR need to be handled cautiously. To this end, three loss functions opt-in training: 1) the L_1 reconstruction loss, 2) Learned Perceptual Image Patch Similarity (LPIPS) [69], and 3) identity preservation loss.

1) L_1 LOSS

The L_1 loss function is chosen for the reconstruction rather than L_2 since L_1 encourages less blurring [67]:

$$L_{l1}(G^i(I_{LR}), I_{HR}) = \frac{1}{whc} \sum_{i,j,k} |(G^i(I_{LR}))_{i,j,k} - (I_{HR})_{i,j,k}|, \quad (12)$$

where the height, width, and channel number are represented by h, w, and c, respectively.

2) PERCEPTUAL LOSS

Most of the widely used perceptual losses are based on VGG. However, the VGG-based perceptual loss [70] does not produce precise results for large upscaling factor SR [69]. Also, the VGG network was designed to classify images, and it might not be the ideal solution for the SR quality criteria objective. LPIPS, which measures the distance between two images in a deep feature space, is more in line with human judgment. Hence, the LPIPS is used as the perceptual loss [71]:

$$L_{lpips}(G_4(I_{LR}), I_{HR}) = \sum_k \tau^k(\phi^k(G_4(I_{LR})) - \phi^k(I_{HR})), \quad (13)$$

where τ transforms the deep embedding to the LPIPS score and ϕ is the feature extractor. The score is computed and averaged for k layers.

3) IDENTITY PRESERVATION

Adversarial learning of GANs, Eq. 2, encourages the generator to characterize the attribute in the HR data. However, it does not ensure that the identity information is preserved on the generated SRd output [17]. Moreover, aligning the representation of the HPE backbone by Eq. 10 will encourage the generator to mix the identity-related information between different subjects. Hence, the cosine similarity between the feature vector of the SRd and HR pair obtained from a pre-trained face recognition module is used to enforce identity preservation. This identity preservation loss, L_{ip}, is defined as:

$$L_{ip}(G_4(I_{LR}), I_{HR}) = 1 - \frac{\phi_2(G_4(I_{LR}))}{||\phi_2(G_4(I_{LR}))||_2} \cdot \frac{\phi_2(I_{HR})}{||\phi_2(I_{HR})||_2}, \quad (14)$$

where ϕ_2 is the face recognition model.

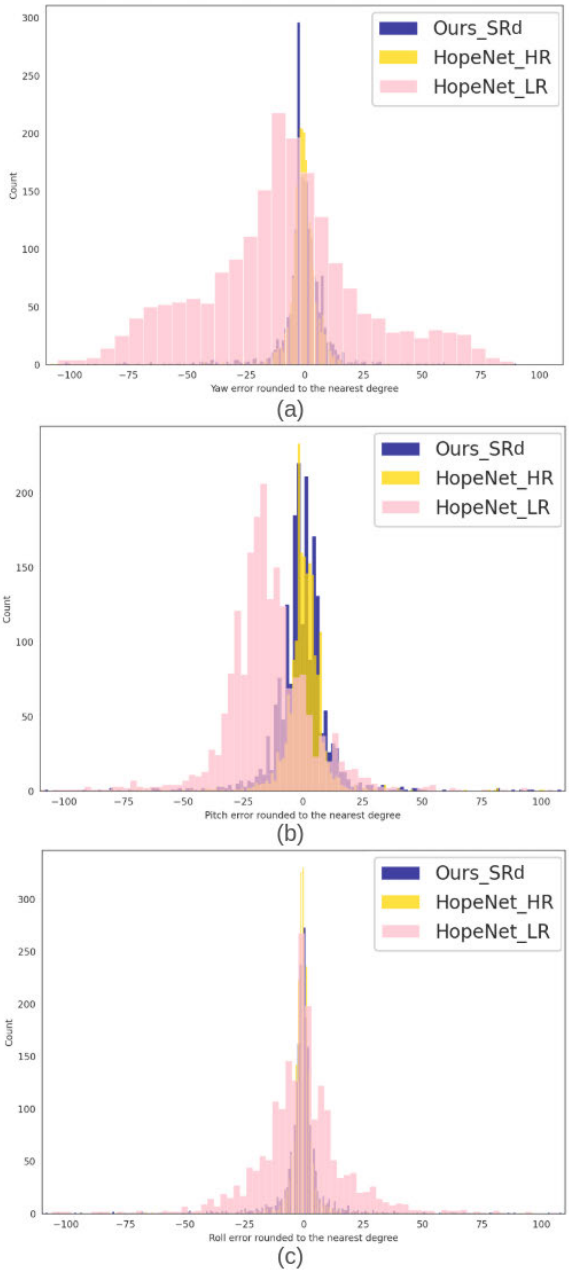

FIGURE 9. a) Yaw, b) pitch, and c) roll prediction error distribution on the AFLW2000 dataset. Comparison between our method and HopeNet when using the LR (8×8) and HR (128×128) images as inputs.

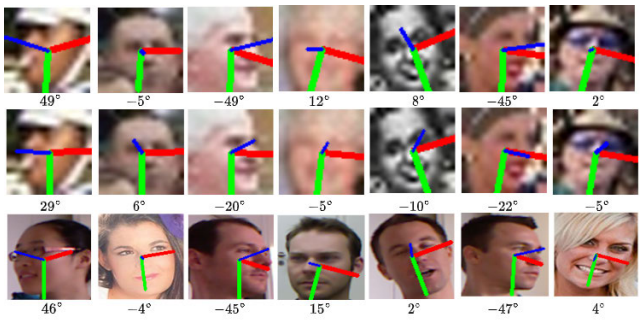

FIGURE 10. Comparison of pose estimation on the the WiderFace dataset between our method (first row) and HopeNet (second row). The blue axis points toward the front of the face, the green pointing downward, and the red pointing to the side. Since there is no head pose label for the WiderFace dataset, we demonstrate the labeled samples in the third row, showing that our approach's prediction is more authentic than the original Hopenet.

By considering the input size of 8×8 and generating images in 4 stages, the overall reconstruction loss in the each stage $i = 1, \ldots, 4$ is a combination of the aforementioned losses:

$$L_{rec(i)} = \lambda_{adv} L^i_{adv} + \lambda_{l1} L^i_{l1} + \lambda_{lpips} L^i_{lpips} + \lambda_{ip} L^i_{ip}, \quad (15)$$

where $\lambda_{(.)}$ coefficients are the regularization parameters. Therefore, the final training loss for all the stages can be defined as:

$$L_{compound} = \sum_{i=1,\ldots,4} \left[\alpha_i L_{rec(i)}\right] + \gamma_p * L_p, \quad (16)$$

where α_i are the regularization parameters for reconstruction loss at each stage and γ_p indicates the weight associated with the pose estimation loss compared to the reconstruction loss. Note that L_p is being applied to the final SRd output.

IV. EXPERIMENTS

A. DATASETS

We employ 300W-LP [6] and CelebA [72] as our training datasets, which contain 60,000 and 500,000 samples, respectively. For evaluation, we use the AFLW2000 [73], BIWI [74], and WiderFace [75] datasets. The AFLW2000 contains 2,000 in-the-wild samples with large variations in pose, illumination, expression, and occlusions. The BIWI [74] consists of videos of participants in indoor environment (15,000 images from 20 subjects). The WiderFace [75] is a challenging face detection dataset containing face images with high degree of variability in pose, occlusion, and degradation.

B. ARCHITECTURE

The Residual-in-Residual Dense Block (RRDB), used in the generator architecture (see Figures 3 and 4) connects all layers within a residual block to enhance the network's capacity [78]. A residual scaling in Fig. 4 is the process of multiplying a parameter β from the range [0, 1] to the residual which prevents instability [20]. As shown in the Fig. 3, in the case of ×16 upscaling factor, the first half of the network performs ×4 upsampling (using two successive PixelShuffle(×2)) and the other half performs the remaining ×4 upsampling. Likewise, in the case of ×8, the first half of the network performs ×2 upsampling, and the other half performs the remaining ×4 upsampling.

Our multi-stage generator provides outputs from the several intermediatory layers of the network rather than just the final layer [79]. In the case of ×16 SR, we enforce constraints on our network at four different image resolutions 16×16, 32×32, 64×64 and 128×128 (see Fig. 3). $D_i \in \{1, 2, 3, 4\}$ are patch-based discriminators that progressively contribute to the generator learning finer details in every scale. We utilize stage-specific patch-based discriminators at multiple SR stages to further push the texture of generated images toward being indistinguishable from I_{HR} [67]. Other important issues about the training of GANs are exploding and vanishing gradients [80]. The former causes the training instability [81], [82]. The latter leads either to the bad local minima or stalled training before convergence [81]. Hence, we employ spectral normalization (SN) in each layer of discriminators to stabilize our adversarial training [82]. Furthermore, D_r is a multilayer perceptron with two hidden layers of size 256, followed by a batch normalization and leaky Relu activation function. At the top of these hidden layers is a single neuron with a Sigmoid activation function.

C. IMAGE DEGRADATION

While the bicubic degradation is rarely suitable for mimicking the real-world LR images, we employ a practical degradation model to synthesize training pairs. Primarily, HR images are degraded by applying downsampling, Gaussian blur, and noise [83]. To further reduce the gap between real/synthesized LR, we adopt the motion blur kernels generated by applying sub-pixel interpolation to the random trajectory vectors [84]. The order of the mentioned steps is shuffled, leading to expanding the degradation space to imitate diverse real-world LR images.

Generally, we can state that blur, downsampling, and noise are the three key factors contributing to the degradation of real images. The proposed degradation framework can be mathematically modeled by:

$$I_{LR} = (I_{HR} \otimes K) \downarrow_s + n, \quad (17)$$

where I_{LR} is obtained by convolving the I_{HR} with a point spread function K (Gaussian or motion blur kernel), followed by a downsampling operation $\downarrow_s$ with scale factor s and addition of white Gaussian noise n with standard deviation σ. Toward a more practical model, blurriness is achieved by two convolutions with Gaussian kernel and motion blur kernel. Downsampling includes bilinear, bicubic, and nearest neighbor interpolations operators. The noise is modeled by additive white Gaussian noise (AWGN). Moreover, instead of using the blur/downsampling/noise-addition pipeline, we randomly shuffle the order of applying

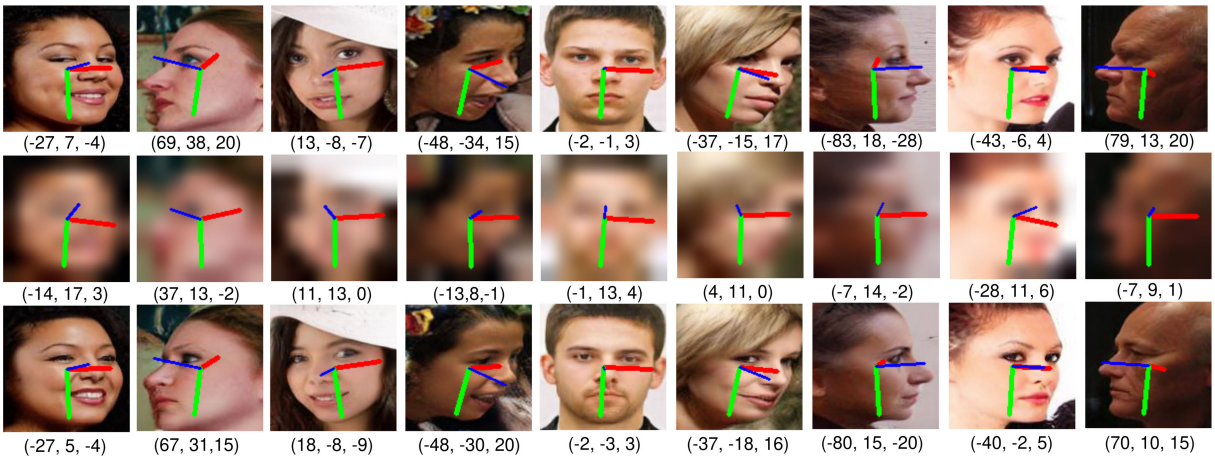

FIGURE 11. Comparison of pose estimation for GT, LR, and SRd images. The pose predictions are plotted over the images of AFLW2000. The blue axis points toward the front of the face, the green pointing downward, and the red pointing to the side. Below each image, we stated (yaw, pitch, roll) values. The first row is the ground truth pose; the second row corresponds to the pose estimated by HopeNet on LR (8×8) images and the third row shows the pose estimated for LR (8×8) images using our proposed method.

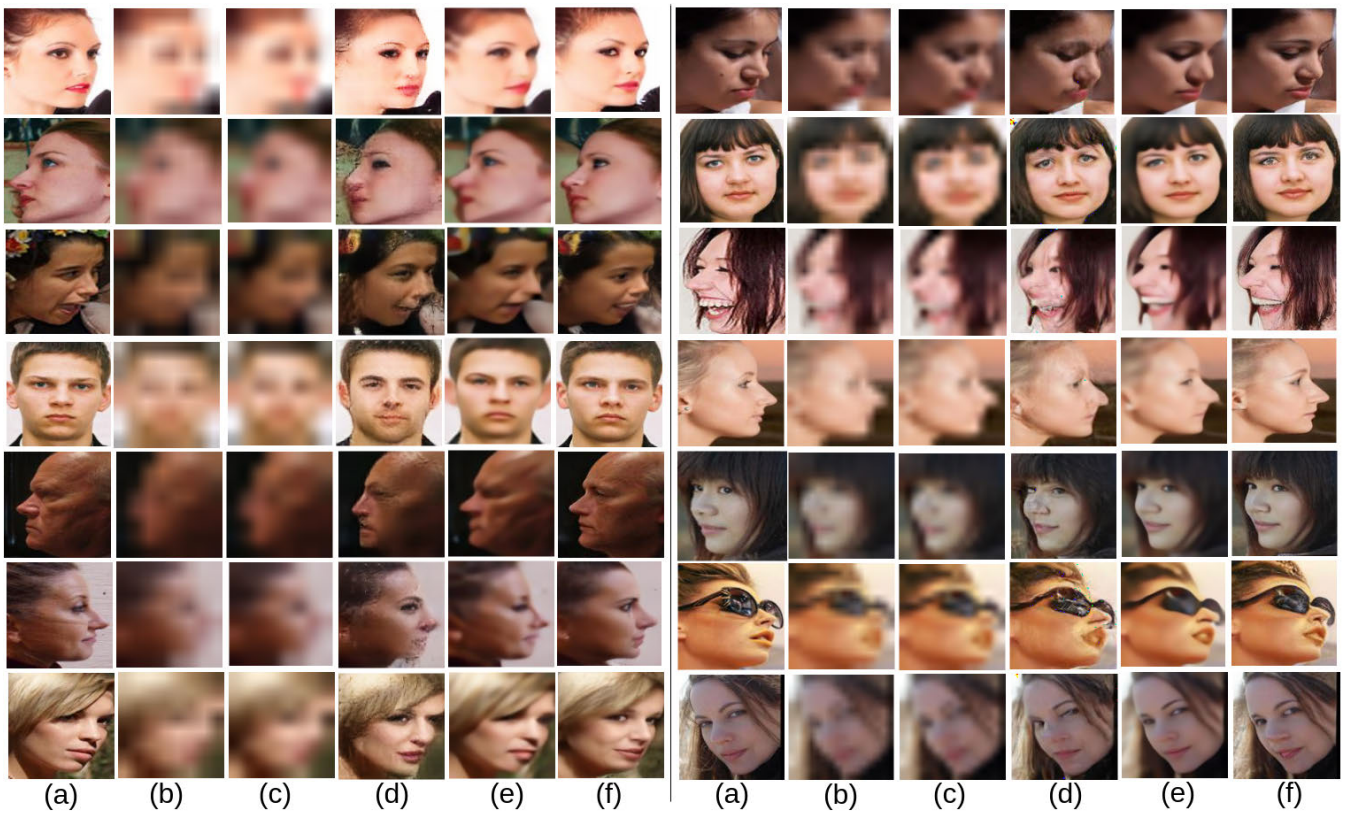

FIGURE 12. Qualitative SR comparison on the AFLW2000 dataset. Left part: (a) HR (128×128), b) LR (8×8), c) bicubic interpolation, d) ExtremeSR [69], e) MGBPv2 [76], and f) ours for the upsampling rate of ×16. Right part: a) HR (128×128), b) LR (16×16), c) bicubic interpolation, d) ProgFSR [32], e) SPARNet [77] and f) ours for the upsampling rate of ×8. Please zoom in to have a better detailed view of the images.

degradation components [85]. Consequently, we can generate various LR images corresponding to each HR image with a wide range of degradations. Fig. 8 shows three different degraded version of each HR sample which are generated during the training.

D. IMPLEMENTATION DETAILS

In degradation process, we randomly apply different downsampling techniques on the I_{HR}, including bilinear, nearest neighbor, and bicubic interpolations. Regarding blur setting, as we mentioned in section IV-C we use

Gaussian and motion blur functions with arbitrary kernel parameters to expand degradation space. The size of Gaussian blur kernel is uniformly sampled from $\{(7 \times 7), (9 \times 9), \ldots, (21 \times 21)\}$. Also, the Gaussian kernel width is uniformly sampled from $[0.1, 2.8]$. We used the method proposed in [84] to generate the motion blur effect. The standard deviation of Gaussian noise is uniformly sampled from $\{1/255, 2/255, \ldots, 25/255\}$. The probability of applying blur and noise degradation steps on an HR image is 0.5.

The training process is divided into two stages; training the stand-alone generator using L_1 reconstruction loss and then training the multi-stage GAN model using the loss function given by Eq. 16. We utilize the pre-trained HopeNet model [9] as the HPE sub-network. We employ the MTCNN [86] to detect faces instead of the Dockerface detector recommended by Hopenet [87], which improves the predicted pose angles compared to the original paper [9]. We fine-tune the pre-trained FaceNet model [88] on our training dataset and use it as an identity feature extractor for identity-preserving loss. blProposed framework benefits from the relative influence of different loss functions, which is guided by the regularization parameters: $\lambda_{pa} = 0.1$, $\lambda_{est} = 1.0$, $\lambda_{sim} = 0.2$, $\lambda_m = 0.01$, $\beta_{mse} = 0.01$, $\lambda_{adv} = 0.01$, $\lambda_{l1} = 1.0$, $\lambda_{lpips} = 0.001$, $\lambda_{ip} = 0.1$, $\gamma_p = 1.0$, $\alpha_1 = 0.001$, $\alpha_2 = 0.005$, $\alpha_3 = 0.01$, and $\alpha_4 = 1$, which are obtained empirically. For our qualitative comparisons, we report both umsampling factors of $\times 8$ and $\times 16$, from 16×16 and 8×8 pixels, to 128×128 pixels.

E. EVALUATION METRICS

1) PEAK SIGNAL-TO-NOISE RATIO (PSNR)

PSNR is a quality metric based on the Mean Squared Error (MSE) of pixels for each channel between the GT and generated SRd image [89]:

$$MSE = \frac{1}{whc} \sum_{i,j,k} ((G(I_{LR}))_{i,j,k} - (I_{HR})_{i,j,k})^2, \tag{18a}$$

$$PSNR = 10 \log_{10}(\frac{M^2}{MSE}), \tag{18b}$$

where the height, width, channel number and maximum possible pixel value are represented by h, w, c, and M, respectively.

2) STRUCTURAL SIMILARITY

Structural Similarity Index Measure (SSIM) examines the homogeneity and phase coherence of the gradient magnitude on the original and reconstructed images to quantifies image quality degradation. The structure, brightness, and contrast of the photos are used to determine how similar they are [89]:

$$SSIM(SR, HR) = \frac{(2\mu_{SR}\mu_{HR} + C_1) + (2\sigma_{SR,HR} + C_2)}{(\mu_{SR}^2 + \mu_{HR}^2 + C_1)(\sigma_{SR}^2 + \sigma_{HR}^2 + C_2)}, \tag{19}$$

where (μ_{SR}, μ_{HR}) and $(\sigma_{SR}, \sigma_{HR})$, and $\sigma_{SR,HR}$ denote the average, standard deviation, and correlation of intensity value of the generated and original image, respectively. C_1 and C_2 are constants for avoiding instability.

TABLE 1. The comparison between the prediction error of the SOTA HPE methods on LR (8×8 and 16×16) and HR input images (128×128): $Error_{LR} - Error_{HR}$. The results are reported for the AFLW2000 dataset.

Method		MAE(°)			
		Yaw	Pitch	Roll	Avg.
RealHePoNet [90]	8×8	54.13	45.63	-	-
FSA [11]		25.14	11.06	10.11	15.44
img2pos [91]		24.83	18.67	12.33	18.6
HopeNet [25]		22.69	20.57	9.52	15.01
Ours		**2.41**	**2.07**	**2.32**	**2.16**
RealHePoNet [90]	16×16	32.19	25.37	-	-
FSA [11]		9.9	4.74	5.37	6.67
img2pos [91]		7.88	5.52	7.11	6.83
HopeNet [25]		5.4	4.6	4.16	4.62
Ours		**0.47**	**0.42**	**0.81**	**0.48**

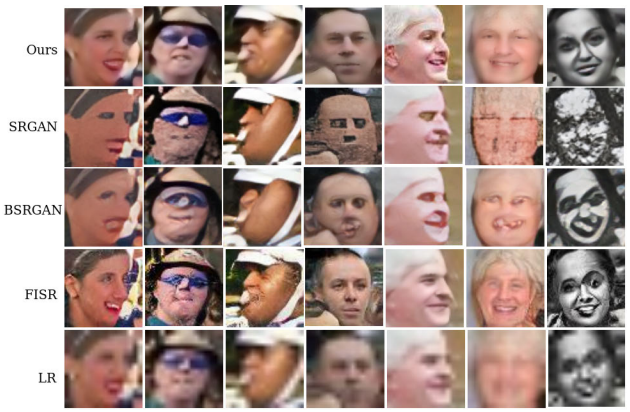

FIGURE 13. Results produced by our model, FISR [92], BSRGAN [83] and SRGAN [25] on real-world low-resolution faces images (16×16) from the WiderFace dataset [75]. Our proposed method generates superior reconstruction over the existing methods for different degraded images.

3) PERCEPTUALLY-LEARNED METRIC

A weighted Euclidian distance between deep features of HR and SRd images provides the LPIPS similarity score [71]. It is shown that the LIPIPS effectively reflect the human perceptual similarity [71]. We chose the LPIPS configuration based on the AlexNet network with the weights learned from the BAPPS dataset [71]. The lower LPIPS score indicates that the two images are more similar.

4) MEAN ABSOLUTE ERROR (MAE)

To measure the head position estimation error, we employ the MAE metric. MAE is obtained as the average error values for the yaw, roll, and pitch and is expressed in degrees (°).

F. RESULTS

In Table 1, we show the differences between the HR (128×128), and LR (8×8, 16×16) HPE error for SOTA methods. According to our investigation in Table 1, the estimation error for LR face images is significantly high. The

results show that the proposed method can reduce the performance gap between the HR and LR HPE errors.

Table 2 shows the proposed method performance compared to the SOTA algorithms. This table presents a performance comparison from both aspects of the HPE error and the qualitative measure of generated SRd images. To have fair comparison, we fine-tuned Super-FAN [56], ProgFSR [32], FSRNet [30], SRGAN [25], SPARNet [77], MGBPv2 [76] and ExtremeSR [76] on our training datasets. Then, the best results before or after fine-tune are reported. Considering the HPE performance, we report the MAE on all three head angles, i. e., yaw, pitch, and roll. From the HPE aspect, the proposed method outperforms the other competitors by a considerable margin for both $\times 16$ and $\times 8$ upscaling factors on the AFLW2000 and BIWI datasets. Moreover, Fig. 9 shows the prediction error distributions on the AFLW2000 dataset. This diagram shows that our framework effectively reduces the error resulting from the LR input images.

Fig. 11 visually compares the performance of the proposed approach, and HopeNet [9] on the AFLW2000 dataset. The GT pose of corresponding HR input images is plotted for better comparison. Our framework fills the gap between high and low-resolution HPE by boosting the performance of the HPE model on LR images. To further investigate the efficiency of the proposed algorithm on the real LR images, we compare the performance of our method with HopeNet [9] on the WiderFace dataset in Fig. 10. Since there is no GT head pose label for the WiderFace dataset, we visually compare the results on the yaw angle with samples of the training dataset, which are visually similar to the presented sample of WiderFace. From Fig. 10, we can observe that the proposed approach estimates the head pose more accurately than the HopeNet [9]. Moreover, considering the arrows on the images that show the front, downward, and side of the faces, our proposed method clearly produces results that are more reliable by human inference.

Fig. 12 investigates the visual quality of our SRd images and compares them with the output of other SOTA algorithms. The first observation from Fig. 12 is that the proposed method effectively generates images that preserve the texture and identity despite the low-resolution inputs. In comparison with other methods, fewer undesired artifacts are produced, and at the same time, realistic details are added to the SRd image. In the case of frontal faces, ProgFSR [32] shows comparable performance to ours. However, in the off-angle scenarios, failure of ProgFSR is apparent; take the fifth row as an example. Our method generates superior reconstruction over other methods, specifically in off-angle face images. Moreover, considering the quantitative measures in Table 2, in $\times 16$ upsampling SR, our method outperforms other studies, and for $\times 8$ upsampling SR achieves the best LPIPS scores. In order to assess the performance of our SR method on the real world images with the resolution of 16 $\times$16, we provide some visual result (Fig. 13) on the extremely degraded LR face images from the Widerface dataset [75] and compare them with ISR [92], BSRGAN [83] and SRGAN [25]. It should be noted that there are no corresponding ground-truth HR images for WiderFace testset.

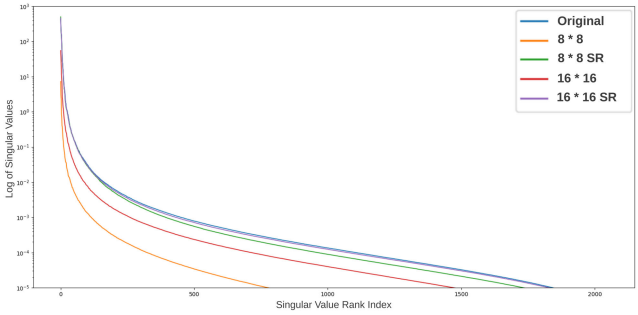

FIGURE 14. Singular value spectrum of embedding spaces. The representations were obtained from pre-trained HopeNet on 4,000 randomly selected samples of the BIWI dataset. Each feature vector has 2,048 dimensions. The singular value spectrum represents each dimension's relative contribution to the module's final output.

G. ABLATION STUDY

1) REPRESENTATION ANALYSIS

Our goal is to establish a resolution-agnostic HPE framework. In our method, the parameters of HPE module are fixed, so we are sure that the performance of the original HR data is maintained. Therefore, we attempted to make the HR and SRd features as similar as possible using the SR sub-network. So far, we have shown the effectiveness of our framework from the quantitative perspective of HPE error and SR metrics. In this section, we show that the proposed method effectively aligns the HR and SRd representations obtained by the HPE embeddings.

Figures 6 and 7 show the distribution of the magnitude of the feature representations obtained from the HPE backbone [9] and MagFace [15] for HR, LR and our SRd face images. The magnitudes of the feature embeddings of the HPE network in Fig. 6 reflects the resolution or quality of the input sample to some extent ((a) 16 $\times$ 16 and (b) 8 $\times$ 8) [13], [15]. Also, Fig. 6 shows the efficacy of our method in reducing the gap between the HR and SRd representations and increasing the task-related quality of images. Moreover, Magface is widely used to measure image quality [15]. In Fig. 7, we demonstrate the quality measure obtained from the MagFace (magnitude of its feature). Reducing the gap between the LR and HR feature magnitude of MagFace via applying our proposed method shows that our approach improves the HPE performance and, at the same time, produces visually acceptable outputs.

To further show the effect of LR input on the HPE module, we investigate the dimension collapse in the HPE module. We evaluate the dimensionality of the HPE backbone by collecting the representation vector of 4,000 randomly selected samples from the BIWI dataset, $Z \in R^{2048\times 4000}$. HPE module embedding has 2,048 dimensions [9]. Then, we compute the covariance matrix, $C \in R^{2048\times 2048}$ of the embedding:

$$C = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(z_i - \bar{z})^T, \tag{20}$$

TABLE 2. HPE and visual quality comparison of the proposed approach with the SOTA methods on the AFLW2000 and BIWI datasets (original HR image is 128×128). For qualitative results, see Figure 4.

		AFLW2000 Dataset							BIWI Dataset						
		MAE (°)				Quality Metrics			MAE (°)				Quality Metrics		
Method	Factor	Yaw↓	Pitch↓	Roll↓	Avg.↓	PSNR↑	SSIM↑	LPIPS↓	Yaw↓	Pitch↓	Roll↓	Avg.↓	PSNR↑	SSIM↑	LPIPS↓
Original		4.63	7.18	5.88	5.89				3.61	4.67	2.87	3.71			
Bicubic		10.03	11.78	9.74	10.51	24.72	0.703	0.204	6.36	9.99	5.48	7.27	27.80	0.794	0.118
SRGAN [25]		8.13	10.12	8.52	8.92	25.13	0.713	0.184	5.10	8.15	4.11	5.78	27.91	0.799	0.098
ProgFSR [32]		6.67	8.88	7.27	7.16	24.63	0.719	0.094	4.10	6.05	3.96	4.70	27.10	0.801	0.096
FSRNet [30]	×8	5.36	7.96	6.84	6.72	26.66	0.771	0.110	4.20	5.91	3.65	4.58	28.14	0.805	0.096
Super-FAN [56]		5.10	7.80	6.40	6.43	26.91	0.788	0.068	4.01	5.59	3.18	4.26	29.05	0.812	0.102
SPARNet [77]		5.12	7.82	6.44	6.52	**28.45**	**0.838**	0.088	3.98	5.69	3.29	4.32	**30.64**	**0.883**	0.081
Ours		**5.03**	**7.54**	**6.23**	**6.26**	27.04	0.792	**0.050**	**3.91**	**5.36**	**3.07**	**4.11**	30.28	0.855	**0.054**
Bicubic		27.32	20.57	15.10	20.99	21.12	0.558	0.239	20.80	18.45	9.11	16.12	23.37	0.656	0.16
SRGAN [25]		10.12	12.35	11.52	11.33	19.90	0.598	0.190	8.86	8.15	6.11	7.70	22.91	0.699	0.106
ExtremeSR [69]	×16	8.74	9.71	8.99	9.14	20.05	0.416	0.154	6.79	6.98	4.43	6.06	23.61	0.708	0.098
MGBPv2 [76]		9.52	10.63	8.58	9.57	20.56	0.525	0.169	6.90	6.63	5.30	6.27	22.89	0.682	0.115
Ours		**6.95**	**9.20**	**7.81**	**7.98**	**22.93**	**0.652**	**0.085**	**4.75**	**5.95**	**3.71**	**4.80**	**25.57**	**0.739**	**0.063**

TABLE 3. Comparison for using the prior information about the image resolution, or Quality Estimator (QE) to determine whether FSR module should be used or not. The result of HPE module on LR images are shown in row ∈ [2 : 6]. The results are showing the performance for AFLW2000 dataset.

Input	MS		FSR		MAE			
	GT	QS	No	Yes	Yaw	Pitch	Roll	Avg.
Original			✓		4.63	7.18	5.88	5.89
Original		✓		✓	4.76	7.34	5.96	6.02
16 × 16			✓		10.03	11.78	9.74	10.51
16 × 16	✓			✓	5.03	7.54	6.23	6.26
16 × 16		✓		✓	5.10	7.69	6.34	6.37
8 × 8			✓		27.32	20.57	15.10	20.99
8 × 8	✓			✓	6.95	9.20	7.81	7.98
8 × 8		✓		✓	7.04	9.25	7.9	8.05

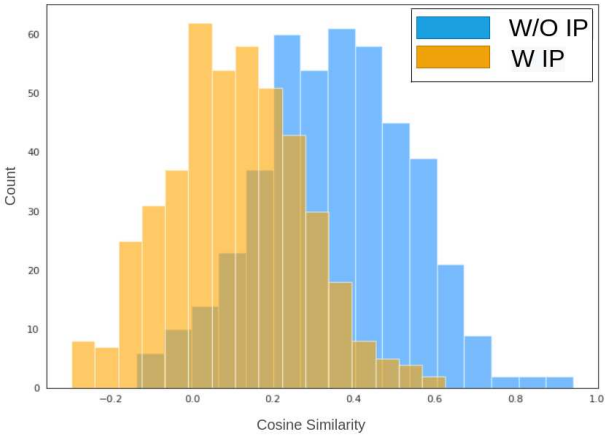

FIGURE 15. Cosine similarity between 500 randomly selected pairs of AFLW2000 dataset. The representation obtained from [88].

where $N = 4,000$ is the number of samples. Fig. 14 shows the singular value decomposition of C in logarithmic scale and sorted order. Extremely small numbers shows no contribution of those dimensions in the final objective [93]. Fig. 14 shows that more than half of the dimensions collapsed when the input is in size of 8×8 and 16×16. This phenomenon can be investigated in future works, such as employing regularization methods (e.g., dropout) during training of HPE network to force the model to use all of the dimensions efficiently. Also, the final classifier's centroids distribution can be further analyzed to alleviate the phenomenon. Another observation is that our SR framework efficiently reduces the gap between singular values of feature vectors from LR and HR input. The severe drop in the 8×8 is in line with Fig. 6, in which the norm of features drastically changes for 8×8 input.

2) MODEL SELECTION

So far, we have shown our proposed framework's capacity to boost the HPE performance using an FSR subnetwork. One major challenge in the real-world application would be deciding whether the FSR is necessary. The soft attention mechanism is not applicable; therefore, we used a hard-selection method [94]. To this end, we use the HopeNet feature to determine whether the input should be fed to FSR subnetwork first or not. We used the HopeNet and trained a binary classifier to decide whether the input image should be SRd. We trained the classifier on augmented 300W-LP. The augmented dataset has both original and down-sampled images. Table 3 compare the results in different scenarios.

In Table 3, we show the performance with and without using the proposed Model Selection (MS) strategy.Also, we report the results of HPE without using the FSR module to present better the performance gained by the whole framework. The table shows the result in three different settings with LR input. We show the HPE performance in the first setting without using the FSR module. In the second setting, we use the prior information (GT) about the data resolution to decide whether to use the FSR module or not. In the other setting, we use the quality classifier to determine whether the image needs to go through the SR network. The automatic quality classification results in negligible performance degeneration from Table 3. However, comparing the results obtained from our framework with the results of the HPE model on the LR input images, we still have better

TABLE 4. Ablation analysis: MAE (of yaw, pitch and roll), PSNR, SSIM and LPIPS for ×8 and ×16 upscaling factors across different models on the AFLW2000 dataset. MSGAN^{lpa} is MSGAN when it is utilized with l_{pa}.

	×8 upscaling factor							×16 upscaling factor						
	MAE (°)				Quality Metrics			MAE (°)				Quality Metrics		
Method	Yaw↓	Pitch↓	Roll↓	Avg.↓	PSNR↑	SSIM↑	LPIPS↓	Yaw↓	Pitch↓	Roll↓	Avg.↓	PSNR↑	SSIM↑	LPIPS↓
Original	4.63	7.18	5.88	5.89										
Bicubic	10.03	11.78	9.47	10.51	24.72	0.703	0.204	27.32	20.57	15.10	20.99	21.12	0.558	0.239
Generator	6.26	8.57	7.46	7.43	28.20	0.931	0.099	8.34	10.92	9.13	9.46	23.34	0.572	0.134
MSGAN	5.13	7.61	6.31	6.35	26.89	0.790	0.060	7.22	9.48	7.96	8.22	22.40	0.619	0.087
MSGAN$^{l_{pa}}$	5.08	7.59	6.31	6.32	26.81	0.788	0.055	7.15	9.42	7.91	8.16	22.37	0.610	0.087
Joint MSGAN	5.03	7.54	6.23	6.26	27.04	0.792	0.050	6.95	9.20	7.81	7.98	22.93	0.652	0.085

TABLE 5. Verification results on the LFW dataset using FaceNet.

Upscaling factor	×8			×16		
Input	Original	LR	SRd	Original	LR	SRd
AUC	0.9683	0.8951	0.9139	0.9683	0.6478	0.8948

performance by a considerable margin. In this manner, we can boost the performance on the LR images and maintain the low prediction error on the HR input.

3) PREVENTING IDENTITY MIXING

In the task of HPE, identity-related information is irrelevant, and the network tries to ignore any identity information. Therefore, reducing the HPE error and increasing the similarity between HR and LR representations obtained from the HPE backbone leads the generator toward mixing the identities. To alleviate this, we utilize an identity preserving loss described in section III-C3. To validate, we experiment on the WiderFace dataset. To this end, 500 negative, i. e., presenting different identities, pairs (1000 images) are randomly selected. We apply two versions of our method, with and without l_{ip}, on these images and then calculate the cosine similarity score for each pair. Fig. 15 clearly shows that without using the identity preservation loss function, the generator tends to mix the identities of different subjects.

Furthermore, we provide the face verification results in Table 5 on the LFW dataset [95] using FaceNet [88] in terms of the Area Under the Curve (AUC). Training the FSR network to boost HPE performance, can result in identity mixing; however, the proposed method could maintain the identity characteristics and increase the face verification performance for both ×8 and ×16 upscaling factor.

4) CONTRIBUTION OF LOSSES

On the AFLW2000 dataset, we provide ablation experiments to study the followings: stand-alone generator just using L_{l1}, Multi Stage GAN (MSGAN) model after adding L_{l1} at other stages and adding other reconstruction losses to the GAN model ($L^i_{adv} + L^i_{l1} + L_{lpips} + L_{ip}$), $MSGAN^{L_{pa}}$ which adds (L_{pa}) to the previous step, and finally joint MSGAN model that integrates head pose related losses ($L_{est} + L_{sim} + L_m$). GAN-based methods underperform in terms of MSE due to the introduction of adversarial losses, which tend to allow the models to achieve the perceptually better SR results but result in the more reconstruction errors. Detailed loss configurations are shown in the Table 4. Based on these results, our proposed models provide decent results which are almost comparable to the original HR images.

V. CONCLUSION

To address the challenge of HPE in the LR face images, we proposed a method that jointly optimizes SR and HPE. We demonstrated that our network significantly increase the pose estimation accuracy for the LR face images. We showed remarkable improvements in SR for extremely LR face images by using head pose related losses. Our proposed network has two important components. First, we employ the reconstruction loss (including: GAN adversarial losses, L_1 loss, perceptual loss, and identity preserving loss) in our MSGAN to add structural and identity details to the SRd images. Secondly, the total loss integrates the head pose related losses with reconstruction loss to guide the SR sub-network to generate the SRd images conducive to recognizing poses. To validate the effectiveness of the proposed approach, we presented extensive experimental results, evaluations, and ablation studies.

REFERENCES

[1] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009.
[2] Z. Wang, J. Liu, K. He, Q. Xu, X. Xu, and H. Liu, "Screening early children with autism spectrum disorder via response-to-name protocol," *IEEE Trans. Ind. Informat.*, vol. 17, no. 1, pp. 587–595, Jan. 2021.
[3] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12026–12035.
[4] D. C. Luvizon, D. Picard, and H. Tabia, "Multi-task deep learning for real-time 3D human pose estimation and action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2752–2764, Aug. 2020.
[5] Z. Song, Z. Yin, Z. Yuan, C. Zhang, W. Chi, Y. Ling, and S. Zhang, "Attention-oriented action recognition for real-time human–robot interaction," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 7087–7094.
[6] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 146–155.
[7] R. Valle, J. M. Buenaposada, A. Valdés, and L. Baumela, "Face alignment using a 3D deeply-initialized ensemble of regression trees," *Comput. Vis. Image Understand.*, vol. 189, Dec. 2019, Art. no. 102846.
[8] Z. Cao, Z. Chu, D. Liu, and Y. Chen, "A vector-based representation to enhance head pose estimation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1188–1197.

[9] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2074–2083.

[10] H.-W. Hsu, T.-Y. Wu, S. Wan, W. H. Wong, and C.-Y. Lee, "QuatNet: Quaternion-based head pose estimation with multiregression loss," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1035–1046, Apr. 2019.

[11] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang, "FSA-Net: Learning fine-grained structure aggregation for head pose estimation from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1087–1096.

[12] M. Kim, A. K. Jain, and X. Liu, "AdaFace: Quality adaptive margin for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18750–18759.

[13] M. S. E. Saadabadi, S. R. Malakshan, A. Zafari, M. Mostofa, and N. M. Nasrabadi, "A quality aware sample-to-sample comparison for face recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 6129–6138.

[14] W. Wang, H. Zhang, Z. Yuan, and C. Wang, "Unsupervised real-world super-resolution: A domain adaptation perspective," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4318–4327.

[15] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "MagFace: A universal representation for face recognition and quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14225–14234.

[16] N. Najafzadeh, H. Kashiani, M. S. E. Saadabadi, N. A. Talemi, S. R. Malakshan, and N. M. Nasrabadi, "Face image quality vector assessment for biometrics applications," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV) Workshops*, Jan. 2023, pp. 511–520.

[17] F. Liu, M. Kim, A. Jain, and X. Liu, "Controllable and guided face synthesis for unconstrained face recognition," 2022, *arXiv:2207.10180*.

[18] M. S. E. Saadabadi, S. R. Malakshan, S. Soleymani, M. Mostofa, and N. M. Nasrabadi, "Information maximization for extreme pose face recognition," 2022, *arXiv:2209.03456*.

[19] M. Mostofa, M. S. E. Saadabadi, S. R. Malakshan, and N. M. Nasrabadi, "Pose attention-guided profile-to-frontal face recognition," 2022, *arXiv:2209.07001*.

[20] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 1–16.

[21] H. Kashiani, S. M. Sami, S. Soleymani, and N. M. Nasrabadi, "Robust ensemble morph detection with domain generalization," 2022, *arXiv:2209.08130*.

[22] J. Jiang, C. Wang, X. Liu, and J. Ma, "Deep learning-based face super-resolution: A survey," *ACM Comput. Surv.*, vol. 55, no. 1, pp. 1–36, Jan. 2023.

[23] M. Zhang and Q. Ling, "Supervised pixel-wise GAN for face super-resolution," *IEEE Trans. Multimedia*, vol. 23, pp. 1938–1950, 2021.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[25] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE CVPR*, Jul. 2017, pp. 4681–4690.

[26] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1415–1424.

[27] Y.-J. Ju, G.-H. Lee, J.-H. Hong, and S.-W. Lee, "Complete face recovery GAN: Unsupervised joint face rotation and de-occlusion from a single-view image," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 3711–3721.

[28] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H.-P. Seidel, P. Pérez, M. Zollhofer, and C. Theobalt, "StyleRig: Rigging StyleGAN for 3D control over portrait images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6142–6151.

[29] X. Hu, W. Ren, J. LaMaster, X. Cao, X. Li, Z. Li, B. Menze, and W. Liu, "Face super-resolution guided by 3D facial priors," in *Proc. 16th Eur. Conf.* Glasgow, U.K.: Springer, 2020, pp. 763–780.

[30] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "FSRNet: End-to-end learning face super-resolution with facial priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2492–2501.

[31] X. Li, M. Liu, Y. Ye, W. Zuo, L. Lin, and R. Yang, "Learning warped guidance for blind face restoration," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 272–289.

[32] D. Kim, M. Kim, G. Kwon, and D.-S. Kim, "Progressive face super-resolution via attention to facial landmark," 2019, *arXiv:1908.08239*.

[33] T. Shang, Q. Dai, S. Zhu, T. Yang, and Y. Guo, "Perceptual extreme super resolution network with receptive field block," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 440–441.

[34] S. Gu et al., "AIM 2019 challenge on image extreme super-resolution: Methods and results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3556–3564.

[35] T. Xie, X. Yang, Y. Jia, C. Zhu, and X. Li, "Adaptive densely connected single image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3432–3440.

[36] M. Ghifary, D. Balduzzi, W. B. Kleijn, and M. Zhang, "Scatter component analysis: A unified framework for domain adaptation and domain generalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1414–1430, Jul. 2017.

[37] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7167–7176.

[38] X. Liu, W. Liang, Y. Wang, S. Li, and M. Pei, "3D head pose estimation with convolutional neural network trained on synthetic images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1289–1293.

[39] M. Patacchiola and A. Cangelosi, "Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods," *Pattern Recognit.*, vol. 71, pp. 132–143, Nov. 2017.

[40] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *Proc. 13th Eur. Conf.* Zurich, Switzerland: Springer, 2014, pp. 109–122.

[41] A. Kumar, A. Alavi, and R. Chellappa, "KEPLER: Keypoint and pose estimation of unconstrained faces by learning efficient H-CNN regressors," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 258–265.

[42] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2879–2886.

[43] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

[44] Y. Liang, J.-H. Lai, W.-S. Zheng, and Z. Cai, "A survey of face hallucination," in *Proc. 7th Chin. Conf. (CCBR)*. Guangzhou, China: Springer, 2012, pp. 83–93.

[45] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 35, no. 3, pp. 425–434, Aug. 2005.

[46] J.-S. Park and S.-W. Lee, "An example-based face hallucination method for single-frame, low-resolution facial images," *IEEE Trans. Image Process.*, vol. 17, no. 10, pp. 1806–1816, Oct. 2008.

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

[48] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun./Jul. 2004, p. 1.

[49] R. Zhang, L. Xu, Z. Yu, Y. Shi, C. Mu, and M. Xu, "Deep-IRTarget: An automatic target detector in infrared imagery using dual-domain feature extraction and allocation," *IEEE Trans. Multimedia*, vol. 24, pp. 1735–1749, 2022.

[50] R. Zhang, S. Yang, Q. Zhang, L. Xu, Y. He, and F. Zhang, "Graph-based few-shot learning with transformed feature propagation and optimal class allocation," *Neurocomputing*, vol. 470, pp. 247–256, Jan. 2022.

[51] T. Yang, P. Ren, X. Xie, and L. Zhang, "GAN prior embedded network for blind face restoration in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 672–681.

[52] C. Tian, Y. Xu, W. Zuo, B. Zhang, L. Fei, and C.-W. Lin, "Coarse-to-fine CNN for image super-resolution," *IEEE Trans. Multimedia*, vol. 23, pp. 1489–1502, 2021.

[53] H. Liu, X. Zheng, J. Han, Y. Chu, and T. Tao, "Survey on GAN-based face hallucination with its model development," *IET Image Process.*, vol. 13, no. 14, pp. 2662–2672, Dec. 2019.

[54] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2439–2448.

[55] Z.-S. Liu, W.-C. Siu, and Y.-L. Chan, "Features guided face super-resolution via hybrid model of deep learning and random forests," *IEEE Trans. Image Process.*, vol. 30, pp. 4157–4170, 2021.

[56] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 109–117.

[57] J. Shermeyer and A. Van Etten, "The effects of super-resolution on object detection performance in satellite imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.

[58] M. Mostofa, S. N. Ferdous, B. S. Riggan, and N. M. Nasrabadi, "Joint-SRVDNet: Joint super resolution and vehicle detection network," *IEEE Access*, vol. 8, pp. 82306–82319, 2020.

[59] Y. Pang, J. Cao, J. Wang, and J. Han, "JCS-Net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 12, pp. 3322–3331, Dec. 2019.

[60] J. Wu, S. Ding, W. Xu, and H. Chao, "Deep joint face hallucination and recognition," 2016, *arXiv:1611.08091*.

[61] L. Jiang, N. Wang, Q. Dang, R. Liu, and B. Lai, "PP-MSVSR: Multi-stage video super-resolution," 2021, *arXiv:2112.02828*.

[62] Y. Li, B. Jiang, Y. Lu, and L. Shen, "Fine-grained adversarial image inpainting with super resolution," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[63] X. Li, J. Wu, Z. Lin, H. Liu, and H. Zha, "Recurrent squeeze-and-excitation context aggregation net for single image deraining," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 254–269.

[64] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3883–3891.

[65] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14821–14831.

[66] G. Cheng, A. Matsune, Q. Li, L. Zhu, H. Zang, and S. Zhan, "Encoder–decoder residual network for real super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.

[67] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE CVPR*, Jul. 2017, pp. 1125–1134.

[68] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[69] Y. Jo, S. Yang, and S. J. Kim, "Investigating loss functions for extreme super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 424–425.

[70] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. 14th Eur. Conf.* Amsterdam, The Netherlands: Springer, 2016, pp. 694–711.

[71] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[72] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[73] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 787–796.

[74] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3D face analysis," *Int. J. Comput. Vis.*, vol. 101, no. 3, pp. 437–458, 2013.

[75] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3676–3684.

[76] P. N. Michelini, W. Chen, H. Liu, and D. Zhu, "MGBPv2: Scaling up multi-grid back-projection networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3399–3407.

[77] C. Chen, D. Gong, H. Wang, Z. Li, and K.-Y.-K. Wong, "Learning spatial attention for face super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 1219–1231, 2021.

[78] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[79] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1947–1962, Aug. 2019.

[80] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.

[81] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[82] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*.

[83] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4791–4800.

[84] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "DeblurGAN: Blind motion deblurring using conditional adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8183–8192.

[85] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 702–703.

[86] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[87] N. Ruiz and J. M. Rehg, "Dockerface: An easy to install and use faster R-CNN face detector in a Docker container," 2017, *arXiv:1708.04370*.

[88] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[89] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[90] R. Berral-Soler, F. J. Madrid-Cuevas, R. Muñoz-Salinas, and M. J. Marín-Jiménez, "RealHePoNet: A robust single-stage ConvNet for head pose estimation in the wild," *Neural Comput. Appl.*, vol. 33, no. 13, pp. 7673–7689, Jul. 2021.

[91] V. Albiero, X. Chen, X. Yin, G. Pang, and T. Hassner, "img2pose: Face alignment and detection via 6DoF, face pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7617–7627.

[92] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a GAN to learn how to do image degradation first," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 185–200.

[93] L. Jing, P. Vincent, Y. LeCun, and Y. Tian, "Understanding dimensional collapse in contrastive self-supervised learning," 2021, *arXiv:2110.09348*.

[94] M. Kolahdouzi, A. Sepas-Moghaddam, and A. Etemad, "Face trees for expression recognition," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Dec. 2021, pp. 1–5.

[95] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Dept. Comput. Sci., Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.

SAHAR RAHIMI MALAKSHAN (Student Member, IEEE) received the B.Sc. degree in electrical engineering and the M.Sc. degree in biomedical engineering from the K. N. Toosi University of Technology, Tehran, Iran. She is currently pursuing the Ph.D. degree in electrical engineering with West Virginia University, WV, USA. Her area of research interests include image processing, deep learning, and their applications in computer vision and biometrics.

MOHAMMAD SAEED EBRAHIMI SAADABADI (Student Member, IEEE) received the B.Sc. degree in electrical engineering and the M.Sc. degree in biomedical engineering from the K. N. Toosi University of Technology, Tehran, Iran, in 2017 and 2021, respectively. He is currently pursuing the Ph.D. degree with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, USA. His research interests include deep learning, computer vision, and machine learning.

MOKTARI MOSTOFA (Student Member, IEEE) was born in Dhaka, Bangladesh, in 1993. She received the B.Sc. degree in electrical and electronic engineering and the M.Sc. degree in communication and signal processing from the University of Dhaka, Bangladesh, in 2016 and 2018, respectively. She is currently pursuing the Ph.D. degree in electrical engineering with West Virginia University, WV, USA. Since 2018, she has been a Research Assistant with the Deep Learning Laboratory and doing research with Prof. Nasser M. Nasrabadi. Her research interests include the applications of deep learning, machine learning, and image processing.

SOBHAN SOLEYMANI (Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Tehran, the M.Sc. degree in electrical engineering from the École Polytechnique Fédérale de Lausanne (EPFL), and the Ph.D. degree in electrical engineering from the Lane Department of Computer Science and Electrical Engineering, West Virginia University, in 2021. His areas of research interests include computer vision, machine learning, signal and image processing, and their application in biometrics. He has served as a Reviewer for IEEE Transactions on Biometrics, Behavior, and Identity Science (TBIOM), IEEE Transactions on Neural Networks and Learning Systems (TNNLS), *Neural Information Processing Systems (NeurIPS)*, International Conference on Learning Representations (ICLR), International Joint Conference on Biometrics (IJCB), and Winter Conference on Applications of Computer Vision (WACV).

NASSER M. NASRABADI (Fellow, IEEE) received the B.Sc. (Eng.) and Ph.D. degrees in electrical engineering from the Imperial College of Science and Technology, University of London, London, U.K., in 1980 and 1984, respectively. In 1984, he was at IBM, U.K., as a Senior Programmer. From 1985 to 1986, he was at the Philips Research Laboratory, New York, NY, USA, as a member of the Technical Staff. From 1986 to 1991, he was an Assistant Professor at the Department of Electrical Engineering, Worcester Polytechnic Institute, Worcester, MA, USA. From 1991 to 1996, he was an Associate Professor at the Department of Electrical and Computer Engineering, State University of New York at Buffalo, Buffalo, NY, USA. From 1996 to 2015, he was a Senior Research Scientist at the U.S. Army Research Laboratory. Since 2015, he has been a Professor with the Lane Department of Computer Science and Electrical Engineering. His current research interests include image processing, computer vision, biometrics, statistical machine learning theory, sparsity, robotics, neural networks, and image processing. He is a fellow of the International Society for Optics and photonics, ARL, and SPIE. He has served as an Associate Editor for the IEEE Transactions on Image Processing, the IEEE Transactions on Circuits and Systems for Video Technology, and the IEEE Transactions on Neural Networks and Learning Systems.