Face Hallucination with Finishing Touches

Yang Zhang, Ivor W. Tsang, Jun Li, Ping Liu, Xiaobo Lu, and Xin Yu

arXiv:2002.03308v1 [cs.CV] 9 Feb 2020

***Abstract*—Obtaining a high-quality frontal face image from a low-resolution (LR) non-frontal face image is primarily important for many facial analysis applications. However, mainstreams either focus on super-resolving near-frontal LR faces or frontalizing non-frontal high-resolution (HR) faces. It is desirable to perform both tasks seamlessly for daily-life unconstrained face images. In this paper, we present a novel Vivid Face Hallucination Generative Adversarial Network (VividGAN) devised for simultaneously super-resolving and frontalizing tiny non-frontal face images. VividGAN consists of a Vivid Face Hallucination Network (Vivid-FHnet) and two discriminators, *i.e.,* Coarse-D and Fine-D. The Vivid-FHnet first generates a coarse frontal HR face and then makes use of the structure prior, *i.e.,* fine-grained facial components, to achieve a fine frontal HR face image. Specifically, we propose a facial component-aware module, which adopts the facial geometry guidance as clues to accurately align and merge the coarse frontal HR face and prior information. Meanwhile, the two-level discriminators are designed to capture both the global outline of the face as well as detailed facial characteristics. The Coarse-D enforces the coarse hallucinated faces to be upright and complete; while the Fine-D focuses on the fine hallucinated ones for sharper details. Extensive experiments demonstrate that our VividGAN achieves photo-realistic frontal HR faces, reaching superior performance in downstream tasks, *i.e.,* face recognition and expression classification, compared with other state-of-the-art methods.**

***Index Terms*—Face hallucination, super-resolution, face frontalization, generative adversarial network.**

I. Introduction

As the requirement of public security increases, authentication plays an important role in daily life [1]. High-quality face images are one of the most discriminative biometric cues to provide identity verification. However, in most imaging conditions, spacing distance and relative location between public cameras and human faces are inevitable interference factors, leading to a vast collection of tiny non-frontal face images. Such low-quality face images not only impede human observation but also degrade the performance of downstream machine perception algorithms.

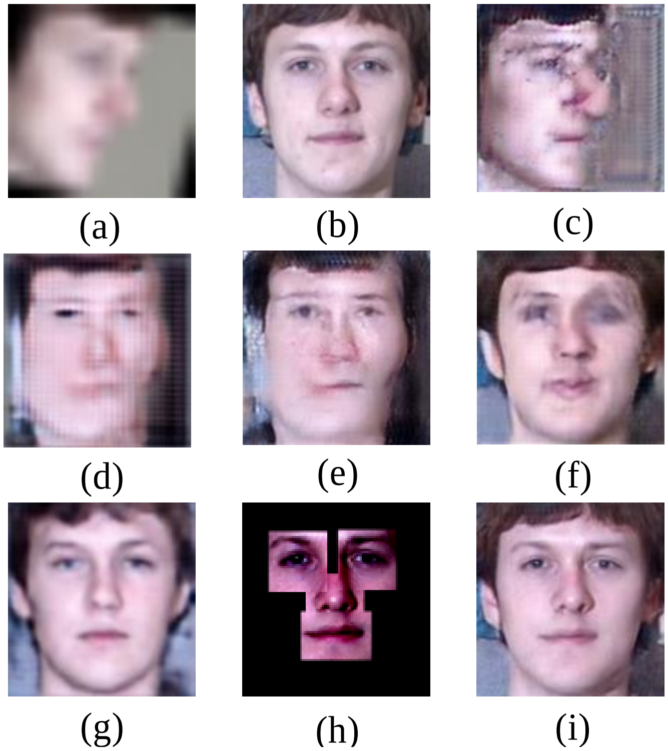


Fig. 1: Face super-resolution and frontalization results on a testing non-frontal LR face. (a) The input 16×16 non-frontal LR face. (b) The 128×128 ground truth frontal face (not available in training). (c) Result obtained by applying [2] on (a). (d) Face frontalization result of (a) by applying [3] after bicubic upsampling and alignment, a single STN network [4] is introduced to align (a) first. (e) Result obtained by applying [3] first and then [2]. (f) Result obtained by applying [2] first and then [3]. (g) Result of TANN [5]. (h) Fine-grained facial components. (i) Result of VividGAN.

Motivated by this challenge, many researchers resort to face super-resolution (SR) techniques to recover high-resolution (HR) face images from the tiny inputs. State-of-the-art face SR methods [2], [6]–[9] utilize deep convolutional neural networks to super-resolve near-frontal low-resolution (LR) faces. Meanwhile, some other works exploit face frontalization algorithms [10]–[13] to achieve frontal faces from their non-frontal counterparts. Conventional and emerging face frontalization techniques [14], [15] mainly warp 2D face images according to 3D face models based on real-time detected facial landmarks.

In this study, we propose to hallucinate a frontal HR face from a non-frontal LR input. To achieve this goal, a naive idea is to *sequentially* combine existing face SR and frontalization models. However, as seen in Figs. 1(e) and (f), the results undergo obvious distortions and severe artifacts. This is mainly caused by two factors: (**i**) Both face SR and face frontalizaiton are ill-posed inverse problems, and the errors in

This work is done when the first author visits the University of Technology Sydney. Y. Zhang is with the School of Automatic, Southeast University, Nanjing 210096, China; the Centre for 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, J. Li and P. Liu are with the Centre for Artificial Intelligence, University of Technology Sydney, Ultimo, NSW 2007, Australia, e-mail: (ivor.tsang@uts.edu.au, Jun.Li@uts.edu.au and pino.pingliu@gmail.com).

X. B. Lu is with the School of Automatic, Southeast University, Nanjing 210096, China; Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing 210096, China, email: xblu2013@126.com.

X. Yu is with the Research School of Engineering, Australian National University, Canberra, Australia, e-mail: xin.yu@anu.edu.au.

either process cannot be eliminated by the other process but are exaggerated. (**ii**) Pose variation brings challenges for state-of-the-art face SR techniques (see the result in Fig. 1(c)) and low resolution causes difficulties for existing face frontalization algorithms (see the result in Fig. 1(d)). Consequently, we face a chicken-and-egg problem: face SR is better guided by face frontalization, while the latter requires a HR face. This challenging issue cannot be addressed by simply combining the two models.

More recently, [5] proposes a Transformative Adversarial Neural Network (TANN), which can jointly frontalize and super-resolve non-frontal LR face images. Compared with the straightforward combination of face frontalization and SR models, the joint mechanism, as performed in TANN, is able to avoid the artifacts. However, due to its single-stage mechanism, TANN does not have a "looking back" mechanism to preview and revise the faces. Therefore, TANN may fail to refine local details for the input faces under extreme poses or complex expressions. As shown in Fig. 1(g), unnatural artifacts still exist in the result of TANN.

Unlike previous works, we propose a novel Vivid Face Hallucination Generative Adversarial Network (VividGAN), targeting *jointly* super-resolve and frontalize tiny non-frontal face images without introducing unwanted blurs and artifacts. In this manner, the two tasks, face SR and frontalization can be alternatively facilitated with the guidance from each other in a unified framework. Specially, VividGAN is designed to *progressively* hallucinate the non-frontal LR face in a coarse-to-fine manner (see Fig. 2). Furthermore, we observe that inherent facial components, *i.e.,* eyes, noses and mouths, construct the facial structure. Therefore, we introduce the facial structure prior, *i.e.,* fine-grained facial components, as the semantic guidance to achieve realistic facial details as our "finishing touches".

Our VividGAN consists of a Vivid Face Hallucination Network (Vivid-FHnet) that comprises a coarse FH network and a fine FH network, as well as two discriminators, *i.e.,* Coarse-D and Fine-D. Their details are as follows: (**i**) The coarse FH network is designed to super-resolve and frontalize the input non-frontal LR face roughly. We do not assume the input face is aligned in advance. Instead, we interweave the spatial transformation networks (STNs) [4] with upsampling layers to compensate for the misalignment. In order to guarantee the content-integrity of hallucinated results, we introduce a mirror symmetry loss based on domain-specific knowledge of human faces. As a result, we can generate the coarse frontal HR face, which facilitates the next hallucination procedure. (**ii**) Subsequently, we propose the fine SR network to explicitly incorporate the structure prior, *i.e.,* fine-grained facial components, to achieve the fine frontal HR face image. First, we design a touching-up subnetwork, which integrates multi-scale upscale and downscale blocks along with skip connections, to effectively estimate the fine-grained facial components (see Fig. 1(h)). Then, we feed the structure prior and the coarse frontal HR face into a fine-integration subnetwork for further refinement. To do so, we put forward a facial component-aware module, which adopts the facial geometry guidance as clues, to align and merge the coarse face and prior information. (**iii**) Different from previous works [5], [16], [17] use a single discriminator, we choose to employ the two-level discriminators, which promote our coarse-to-fine Vivid-FHnet to produce photo-realistic results. In this manner, we can capture both the global outline of the face as well as detailed facial characteristics. Fig. 1(i) shows an example of our hallucinated frontal HR face, which is more visually appealing than the state-of-the-art methods.

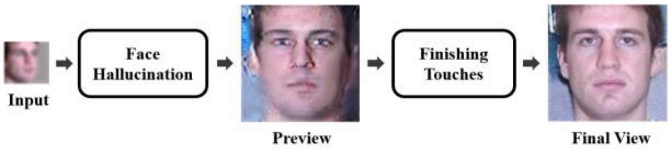


Fig. 2: The pipeline of our VividGAN.

In summary, our contributions are threefold:

- We propose a novel framework, dubbed Vivid Face Hallucination Generative Adversarial Network (VividGAN), to jointly tackle face SR and face frontalization in a whole framework. Instead of directly hallucinating the non-frontal LR input, we first upsample and frontalize the non-frontal LR face in a coarse granularity and then make use of the structure prior, *i.e.,* fine-grained facial components, to achieve the fine frontal HR face.
- We propose a facial component-aware module to facilitate effective enhancement and alignment, based on the facial geometric guidance. In this way, we merge the fine-grained facial components to the coarse frontal HR face seamlessly.
- Our experiments demonstrate that VividGAN is able to frontalize and super-resolve (by an upscaling factor of $8\times$) very LR face images (*e.g.,* 16×16) accompanied with large poses (*e.g.,* 90^o) and complex expressions (*e.g.,* "disgust", "fear"). Moreover, the hallucinated face images of VividGAN achieve superior results in downstream tasks, *i.e.,* face recognition and expression classification, compared with other state-of-the-art methods.

II. Related Work

A. Face Super-resolution

Face Super-resolution (SR) aims at establishing the intensity relationships between input LR and output HR face images. The previous 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.* [18] formulate a linear mapping between LR and HR images to achieve face SR based on an Eigen-transformation of LR faces. [19] incorporates a bilateral filtering to mitigate the ghosting artifacts. [20] morphs HR faces from aligned LR ones based on optimal transport and subspace learning. However, they require precisely aligned input LR and reference HR faces with 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. [21] proposes to super-resolve local LR patches based on a weighted sum of exemplar facial patches in reference HR database. Liu *et al.* [22] put forward a locality-constrained bi-layer network to jointly super-resolve LR faces as well as eliminate noise and outliers. Moreover, SIFT flow [23] and facial landmarks [24] are introduced to locate facial components for further super-resolution. Thus, these techniques need to localize facial components in LR inputs preciously. In this manner, they cannot process very LR faces.

Recently, deep learning based face SR methods are actively explored and achieve superior performance than traditional methods. Yu *et al.* [6] put forward a GAN-derived model to hallucinate very LR face images. Huang *et al.* [25] incorporate the wavelet coefficients into deep convolutional networks to super-resolve LR inputs with multiple upscaling factors. Cao *et al.* [8] design an attention-aware mechanism and a local enhancement network to alternately enhance facial regions for further super-resolution. Xu *et al.* [26] achieve joint super-resolution and deblurring for face and text images based on the GAN framework with a multi-class adversarial loss. Dahl *et al.* [27] present an autoregressive Pixel-RNN [28] to hallucinate pre-aligned LR faces. Yu *et al.* [16] present a multiscale transformative discriminative network to hallucinate unaligned input LR face images with different resolutions. However, these methods focus on super-resolving near-frontal LR faces. Thus, they are restricted to the inputs under small pose variations.

Afterwards, some face SR techniques are put forward to super-resolve the LR faces under large pose variations by introducing facial prior information [2], [9], [29]. Chen *et al.* [9] incorporate the facial geometry priors into their SR model to super-resolve LR faces. Yu *et al.* [2] exploit the facial component information based on the intermediate upsampled features to encourage the upsampling stream to produce photo-realistic HR faces. However, these techniques only super-resolve non-frontal LR faces without frontalizing them for better visual perception.

B. Face Frontalization

Frontal view synthesis, termed as face frontalization, is a challenging research for its ill-posed nature, such as self-occlusions and pose variations. Conventional and emerging researches on face frontalization can be grouped into three classes: 2D/3D local feature warping, statistic modeling as well as deep learning methods.

The first category researches date back to the 3D Morphable Model (3DMM) [14], which extracts the shape and texture bases of a face in PCA subspace. Driven by 3DMM, [30] proposes to formulate new expressions of input faces by estimating the 3D surface from face appearance. Meanwhile, some approaches [15], [31], [32], [32]–[35] put forward to generate frontal faces by mapping each non-frontal face in the gallery set onto a 3D reference surface mesh. However, detecting facial landmarks at acquisition time [36] is the fundamental prerequisite for these approaches to determine the transformation between the gallery image and the query template. Thus, they fail to process the face images with extreme yaw angles or in low resolutions.

Due to the fact that the frontal face has the minimum rank of all various poses, the second category researches propose to infer frontal views based on statistical models by solving a constrained low-rank minimization problem. Sagonas *et al.* [11] achieve joint face frontalization and facial landmark detection in a whole framework. However, the appearance of their results cannot keep consistent with the frontal views.

More recently, deep learning based methods [3], [10], [13], [37]–[41] have been leveraged for face frontalization research. Kan *et al.* [42] put forward the stacked progressive auto-encoders to progressive frontalize non-frontal HR faces based on the learned pose-robust features. Cole *et al.* [39] present a face recognition network to decompose a input face image into a sparse set of facial landmarks as well as aligned texture maps. Then, a differentiable image warping operation is conducted on the extracted features to produce the frontal view. [43] proposes a CNN-based model to learn 3DMM shapes as well as texture parameters. However, they render frontal faces without taking the image intensity similarity into consideration, resulting in distorted results.

Later, the community has witnessed rapid development of image generative models based on the adversarial learning strategy. Under this trend, the GAN framework [44] has dominated face frontalization research. Tran *et al.* [3] propose a GAN-based model to learn the disentangled representation of input faces, achieving label-assisted face frontalization and pose-invariant face recognition. Jie *et al.* [45] propose the high fidelity pose in-variant model to synthesize the frontal face based on estimated dense correspondence fields and recovered facial texture maps.

Above all, the state-of-the-art techniques treat face frontalization as a HR image-to-image translation problem without taking face SR into consideration simultaneously. On the contrary, our goal is to frontalize profile faces in low resolution as well as super-resolve facial details at the same time, resulting in more challenging synthesis task.

III. Proposed Method: VividGAN

Our VividGAN consists of a Vivid-FHnet that comprises a coarse FH network and a fine FH network, and two discriminators, *i.e.,* Coarse-D and Fine-D. We design the coarse FH network to provide a preview for the fine FH network to further revise. Meanwhile, the tailor-made two-level discriminators are introduced to force the hallucinated frontal HR faces to be more realistic.

A. Coarse FH network

First, we propose the coarse FH network to recover the coarse frontal HR face image. The architecture of our coarse FH network is illustrated in Fig.3. First, an input LR face is encoded to latent features by a convolutional layer.

Then, we hallucinate the latent features by a cascade of Feature Transform module (FT module), which is composed of a spatial transform network (STN) [4], a residual block [46], and a deconvolutional layer. Inspired by [5], [16], in each FT module, the spatial transform network (STN) [4] layer is used

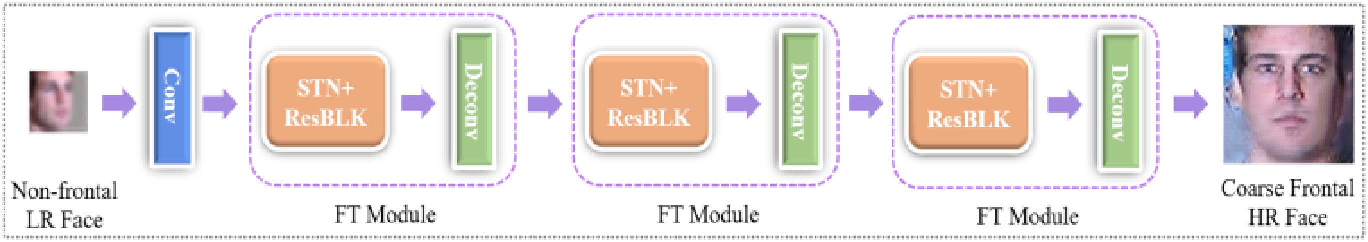


Fig. 3: The architecture of the coarse FH network.

to line up the intermediate features. Afterwards, the deconvolutional layer is adopted to upsample the aligned features. Meanwhile, the residual block [46] is introduced to promote the recovery of high-frequency details as well as increase the network capacity. Since the STN and upsampling layers are interwoven together, our coarse FH network can effectively learn to eliminate the undesired influence of misalignment.

During training, we introduce not only the pixel-wise intensity similarity, known as pixel-wise l_2 loss, but also the feature-wise identity similarity, known as perceptual loss [47]. They are able to enforce the distribution of the latent features to resemble their corresponding frontal counterparts. Specially, to fully utilize the domain-specific knowledge on human faces, we introduce a mirror symmetry loss to guarantee the content-integrity of the generated frontal face. Similar to the works [16], [17], the discriminative loss is employed to achieve upright frontal HR faces. In this way, as shown in Fig. 6(b), our coarse FH network can effectively perform face SR and frontalization on the unaligned non-frontal LR face.

B. Fine FH Network

Then, in the following fine FH network, we design the touching-up subnetwork to estimate the structure prior, *i.e.,* fine-grained facial components. Afterwards, the fine-integration subnetwork is proposed to merge the coarse frontal HR face and prior information seamlessly. In this manner, the fine frontal HR face is generated.

1) Touching-up subnetwork for prior estimation: After obtaining the coarse frontal HR face, the inherent facial components, *i.e.,* eyes, noses, mouths, can be easily localized based on facial landmarks. Then, our touching-up subnetwork takes each coarse facial component as input and estimates the fine-grained one (see Fig. 4).

Inspired by the U-net architecture [48], our touching-up subnetwork integrates successive convolutional and deconvolutional layers along with skip connections to achieve feature fusion. In this way, multi-scale features of facial components can be captured to achieve photo-realistic enhancement. During training, we utilize the intensity similarity loss to constrain the fine-grained facial components to be similar to the corresponding ground-truths.

To effectively align and integrate the coarse frontal HR face with prior information, the fine-grained facial components are stitched onto a masked template to produce a fused view (see Fig. 6(c)). Meanwhile, the max-out fusing strategy is adopted to reduce the stitching artifacts on the overlapping areas in this procedure.

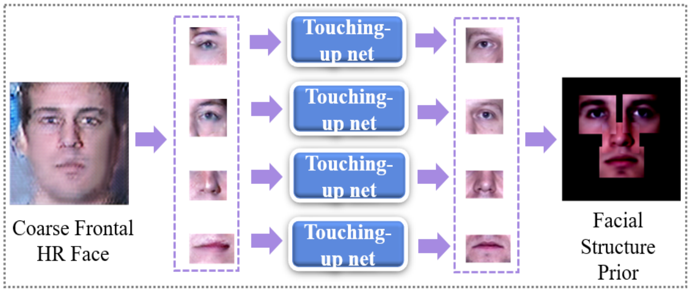


Fig. 4: The architecture of the touching-up subnetwork.

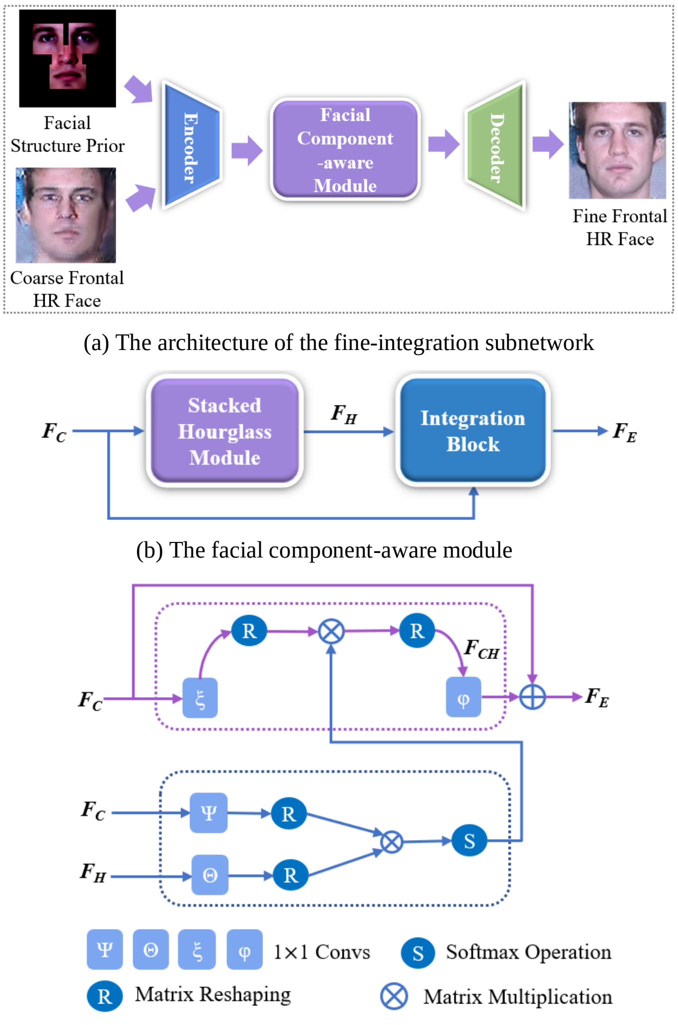
Specially, we have established the Inherent Facial Component Dataset (IFC-set), consisting of left eyes, right eyes, noses and mouths. The IFC-set is constructed based on public and self-built databases, including the Multi-PIE database [49], the MMI Facial Expression database [50], the Celebrity Face Attributes database [51] as well as our proposed PSD-HIGHROAD database [52]. Part of our IFC-set will be released on our website.

2) Fine-integration subnetwork: After inferring the coarse frontal HR face and prior information, *i.e.,* fine-grained facial components, performing accurate alignment and integration becomes the most challenging task. To do so, we design the fine-integration subnetwork, as shown in Fig. 5(a), to generate the fine frontal HR face in three steps: (*i*) the *encoder* concatenates the coarse frontal HR face and prior information to produce the fused features; (*ii*) the *facial component-aware module* achieves facial components alignment and enhancement in feature level; (*iii*) the *decoder* reconstructs the fine frontal HR face.

Facial component-aware module: After obtaining the fused features from the encoder, we apply the facial component-aware module to supervise accurate alignment and enhancement. Fig. 5(b) demonstrates the composition of our facial component-aware module. First, the staked hourglass module [53] is adopted to predict the facial landmark heatmaps from the fused features. Then, we fed the facial landmark heatmaps and the fused features to the integration block for recalibrating.

Inspired by the non-local block [54], the operation procedure of our integration block is as follows.

First, the landmark heatmaps F_H and the fused features F_C are normalized and transformed into two feature spaces θ and ψ to calculate their similarity. Then, the attention features F_{CH} can be formulated as a weighted sum of the landmark heatmaps F_H that are similar to the corresponding positions



(a) The architecture of the fine-integration subnetwork

(b) The facial component-aware module

(c) The processing procedure of the integration block

Fig. 5: The detail of our fine-integration subnetwork.

on the fused features F_C. For the i th output response:

$$F_{CH}^{i} = \frac{1}{M(F)} \sum_{\forall j} \left(\exp\left(W_\theta^T \left(F_C^i\right)^T F_H^j W_\psi \right) F_C^j W_\zeta \right), \quad (1)$$

where $M(F) = \sum_{\forall j} \exp\left(W_\theta^T \left(F_C^i\right)^T F_H^j W_\psi \right)$ is the sum of all output responses over all positions. In Eq. 1, the embedding transformations W_θ, W_ψ and W_ζ are learnt during the training process.

Subsequently, the attention features F_{CH} are transformed into the feature space φ. Finally, we integrate the fused features F_C and the attention features F_{CH}, forming the final refined features F_E, showing in Eq. 4.

$$F_E = \sigma \left(F_{CH} W_\varphi \right) + F_C, \quad (2)$$

where W_θ is also learnt during the training process. σ is a trade-off parameter, which is set to 1 in our experiment. Because the facial landmark heatmaps provide facial geometric guidance, our integration block provides spatial configuration of the inherent facial components.

To verify the effectiveness our facial component-aware module, we conduct comparison experiments. As shown in Fig. 6(f), the VividGAN variant without the facial component-aware module produces flawed results. As expected, VividGAN generates photo-realistic faces (see Fig. 6(j)).

C. Two-level Discriminators

Inspired by [16], [17], we employ the discriminative network to force the generated faces to lie on the same manifold as real frontal faces.

Our Vivid-FHnet hallucinates the non-frontal LR face in a coarse-to-fine manner and produces coarse and fine faces. In this way, the single discriminator, which is commonly used in previous works [5], [16], [17], [55], is unable solve them both. Thus, we propose the two-level discriminators, *i.e.,* Coarse-D and Fine-D, to address these two kinds of inconsistencies. The Coarse-D enforces the coarse hallucinated face to be a upright and complete preview; while the Fine-D focuses on the fine hallucinated one for more sharper details. In this manner, dual discriminative supervision is formulated in our hallucination procedure to avoid the poor convergence problem of GAN.

To analyse the effect of our two-level discriminators, we perform the results of different VividGAN variants, as shown in Figs. 6 (g),(h),(i) and (j). It can be obviously seen that the results of VividGAN (Fig. 6 (j)) capture both the global outline of the face as well as detailed facial characteristics.

D. Loss Terms

In our work, we incorporate five individual losses, including the mirror symmetry loss (L_{sys}) with pixel-wise intensity similarity loss (L_{mse}), feature-wise identity similarity loss (L_{id}), structure-wise similarity loss (L_h) and class-wise discriminative loss (L_{adv}). Here, we detail their basic forms.

1) Mirror symmetry loss: Human faces, like many biological forms, show high levels of mirror symmetry. Thus, we introduce the mirror symmetry loss (L_{sys}) to guarantee the content integrity for the hallucinated faces.

The mirror symmetry loss L_{sys} of a generated image $\hat{h}_i$ takes the formulation:

$$L_{sym} = \mathbb{E}_{(\hat{h}_i)\sim p(\hat{h})} \left\| \vec{h}_i - \hat{h}_i \right\|_F^2, \quad (3)$$

where $\vec{h}_i$ represents the horizontal flipped image for the generated one $\hat{h}_i$.

2) Pixel-wise intensity similarity loss: To enforce the hallucinated face to approximate to the ground truth face in intensity level, the intensity similarity loss L_{mse} is introduced:

$$L_{mse} = \mathbb{E}_{(\hat{h}_i, h_i)\sim p(\hat{h}, h)} \left\| \hat{h}_i - h_i \right\|_F^2, \quad (4)$$

where $p(\hat{h}, h)$ represents the joint distribution of the generated results $\hat{h}_i$ and the corresponding ground truths h_i.

As mention in [56], L_{mse} leads to high peak signal-to-noise ratio (PSNR) values. However, only employing L_{mse} in feed-forward optimization is insufficient to capture the high-frequency features, resulting in over-smoothed results (see Fig. 6(d)).

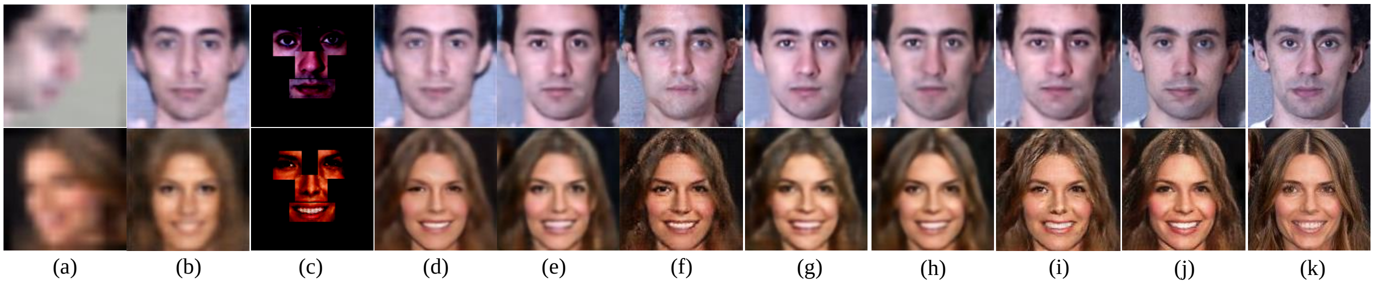


Fig. 6: Ablation study on the effect of different sub-nets and losses on Multi-PIE and CelebA faces. (a) The input 16×16 LR images. (b) The coarse frontal HR faces. (c) The fine-grained facial components. (d) The results only using L_{mse} and L_h. (e) The results using L_{mse}, L_{id} and L_h. (f) The results of VividGAN w/o facial component-aware module (w/o L_h). (g) The results of Vivid-FHnet. (h) The results of Vivid-FHnet and Coarse-D. (i) The results of Vivid-FHnet and Fine-D. (j) The results of VividGAN. (k) The ground truth 128×128 frontal HR images.

3) Feature-wise identity similarity loss: Identity preservation is one of the most important goals in face hallucination [57]. Thus, we adopt the identity similarity loss L_{id} to calculate the Euclidean distance between the high-level features of the hallucinated face and the ground truth one, enabling the identity preserving ability of VividGAN.

The identity similarity loss L_{id} is defined as:

$$L_{id} = \mathbb{E}_{\left(\hat{h}_i, h_i\right) \sim p\left(\hat{h}, h\right)} \left\| \Phi\left(\hat{h}_i\right) - \Phi\left(h_i\right) \right\|_F^2, \tag{5}$$

where $\Phi(\cdot)$ represents the extracted feature vector from the AveragePooling layer of the Resnet50 model [46] for the input images. As Fig. 6(e) shows, L_{id} promotes the generated results retaining more details.

4) Structure-wise similarity loss: To facilitate face alignment as well as provide spatial configuration of facial components, the structure-wise similarity loss [17] L_h is introduced as supervision for our facial component-aware module, which is defined as:

$$L_h = \mathbb{E}_{\left(l_i, h_i\right) \sim p(l, h)} \frac{1}{P} \sum_{k=1}^{P} \left\| H^k\left(f_i\right) - H^k\left(h_i\right) \right\|_2^2, \tag{6}$$

where $H^k\left(f_i\right)$ represents the k-th predicted facial landmark heatmap, which is estimated by the stacked hourglass module [53] on the intermediate facial features f_i. $H^k\left(h_i\right)$ denotes the k-th ground truth facial landmark heatmap obtained by running the Face Alignment Network (FAN) [58] on the ground truth image h_i. Here, we use 68 point facial landmarks to produce the ground-truth heatmaps. Note that L_h is the prerequisite constrain for our facial component-aware module.

5) Class-wise discriminative loss: Aiming at generating photo-realistic results, we infuse the class discriminative information into our Vivid-FHnet by adopting the two-level discriminators. Our goal is to make the two-level discriminators fail to distinguish hallucinated faces from ground truth ones.

The objective function L_D for the discriminator is defined as below:

$$L_D = -\mathbb{E}_{\left(\hat{h}_i, h_i\right) \sim p\left(\hat{h}, h\right)} \left[\log D_d\left(h_i\right) + \log\left(1 - D_d\left(\hat{h}_i\right)\right)\right], \tag{7}$$

where D and d represent the discriminator and its parameters. During training, we minimize the loss L_D and update the parameters for the discriminator.

However, our Vivid-FHnet is designed to produce realistic face images, which should be classified as real faces by the discriminators. Thus, the discriminative loss L_{adv} can be represented uniformly as:

$$L_{adv} = -\mathbb{E}_{\hat{h}_i \sim p\left(\hat{h}\right)} \log\left(D_d\left(\hat{h}_i\right)\right). \tag{8}$$

Aiming at optimizing the Vivid-FHnet, we minimize the loss L_{adv} to update the parameters.

E. Training details

We train the coarse FH network, the touching-up subnetwork and the fine-integration subnetwork to generate coarse frontal HR faces, fine-grained facial components as well as fine frontal HR faces in our unified model. As mentioned in Sec. III. B and C, their training losses are distinct.

The objective function for the coarse FH network, L_C, is expressed as:

$$L_C = L_{mse}^c + L_{sys}^c + \alpha_1 L_{id}^c + \psi_1 L_{adv}^c. \tag{9}$$

The objective function for the touching-up subnetwork, L_T, is expressed as:

$$L_T = L_{mse}^t. \tag{10}$$

The objective function for the fine-integration subnetwork, L_F, is expressed as:

$$L_F = L_{mse}^f + \alpha_2 L_{id}^f + \gamma_2 L_h^f + \psi_2 L_{adv}^f. \tag{11}$$

Above all, the total objective function for Vivid-FHnet, L_G, is expressed as:

$$L_G = L_C + L_T + L_F. \tag{12}$$

Since we intend to hallucinate frontal HR faces rather than generate random faces, we set lower weights on L_{id}, L_h as well as L_{adv}. Thus, $\alpha_1, \psi_1, \alpha_2, \gamma_2$ and ψ_2 in Eqs. 9, 10 and 11 are set to 0.01. Not that, the superscripts for symbols in all the equations always refer to the corresponding networks.

Specially, the training process of our VividGAN model is in three steps: (i) Pre-train the coarse FH network with loss L_C (Eq. 9); (ii) Pre-train the touching-up subnetwork with loss L_T (Eq. 10); (iii) Train the whole VividGAN model with three separate losses: L_G (Eq. 12) is for the VividFHnet, loss L_D^c (Eq. 7) is for the Coarse-D and loss L_D^f (Eq. 7) is for the Fine-D. In the final step, the learning rate of our fine-integration subnetwork is set to be 10 times as the former two subnetworks.

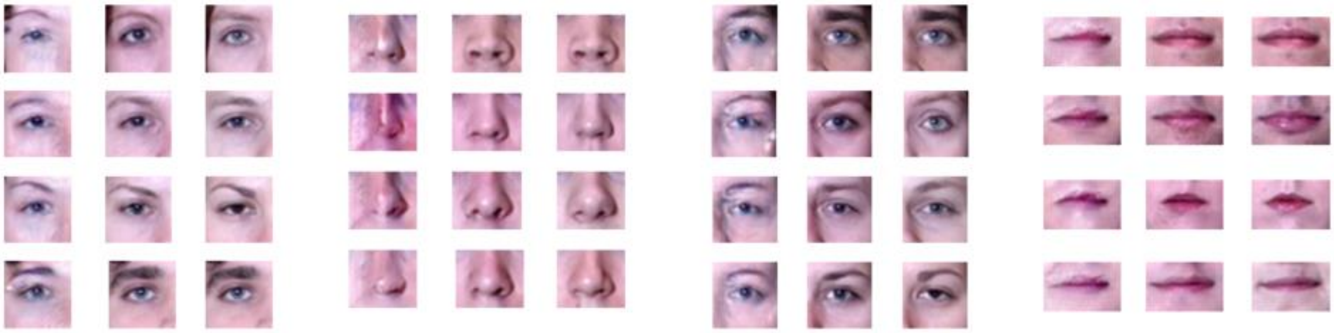


Fig. 7: The inherent facial components. For each part, the first column is the coarse facial components, the second column is the fine-grained facial components, the third column is the ground truth facial components.

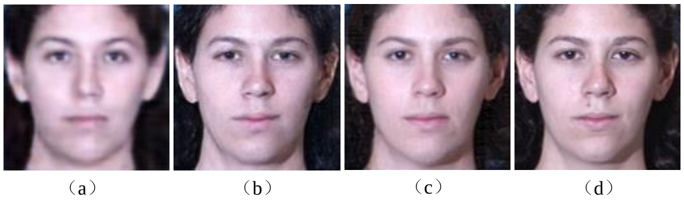


Fig. 8: Effects of the facial structure priors. (a) The result of V-TN model (25.784/0.870 in PSNR/SSIM). (b) The result of VividGAN model (26.289/0.876 in PSNR/SSIM). (c) The result of V+GT model (**28.405/0.902** in PSNR/SSIM). (d) The ground truth image.

IV. Prior knowledge for Joint Face SR and Frontalization

Any real-world object has distinct distribution in its structure, including human faces. Hence, we choose to model and leverage the facial structure prior to facilitate joint face SR and frontalization. From this viewpoint, there are two questions worthy of exploiting: (*i*) Is the facial structure prior knowledge really useful for joint face SR and frontalization? (*ii*) How much improvement does it bring?

To answer the above questions, we conduct subjective and objective experiments on the Multi-PIE database [49], which provides non-frontal/frontal face pairs of 337 individuals under various poses and illumination conditions. We introduce 48K face pairs, consisting of non-frontal faces (±15°, ±30°, ±45°, ±60°, ±75°, ±90°) and corresponding frontal faces (0°), to construct the training set. The rest 2K face pairs are used for testing.

A. Baseline Models

We formulate two baseline models to verify that the structure prior knowledge is significant for joint face SR and frontalization.

The baseline models are summarized as follows (Here, we denote VividGAN as V, touching-up network as TN, ground truth as GT):

- V-TN: we remove the touching-up subnetwork and construct the "V-TN" model, which consists of the coarse FH network and the fine-integration subnetwork.
- V+GT: we introduce the ground truth structure prior, *i.e.,* ground truth facial components, to replace the estimated prior of the touching-up subnetwork, formulating the "V+GT" model.

B. Effects of Prior Information

The performance of the compared models are illustrated in Fig. 8. Thus, we can answer the proposed two questions at the beginning of this section.

As we can see, the results in Figs. 8(b) and (c) show more vivid facial details than the result in Fig. 8(a). This clearly manifests the importance of the facial prior knowledge in hallucination process.

Then, how much improvement does prior information bring? According to Fig. 8, which presents the PSNR and Structural SIMilarity (SSIM) values of the compared models on the testing set, V+GT model (with the ground truth structure prior) outperforms VividGAN model (with the estimated structure prior) as well as V-TN model (without prior information) with 0.505 dB and 2.116 dB, respectively. Based on the PSNR value, the positive effect of the facial prior knowledge for joint face SR and frontalization is confirmed.

Specially, VididGAN provides a solution to hallucinate fine-grained facial components directly from non-frontal LR inputs, as shown in Fig. 7.

V. Experiments

A. Experimental Setup

VividGAN is trained and tested on multiple widely used benchmarks, *i.e.,* the Multi-PIE database [49], the MMI Facial Expression (MMI) database [50] and the Celebrity Face Attributes (CelebA) database [51].

Multi-PIE [49] has been described in Sec. IV. **CelebA** [51] is a large in-the-wild database that contains more than 200K face images under different pose, occlusion and background variations. **MMI** [50] contains 480 image sequences performed by 80 persons. The persons are of mixed ages, different genders, and various ethical backgrounds. We selected the first two images as neutral faces and the last two fifth part as emotional faces. Each of the face images has been manually annotated as one of the seven basic expression categories: "angry", "disgust", "fear", "happy", "sad", "surprise", or "neutral".

Note that the MMI and CelebA databases only provide frontal faces rather than frontal/non-frontal face pairs. Thus,

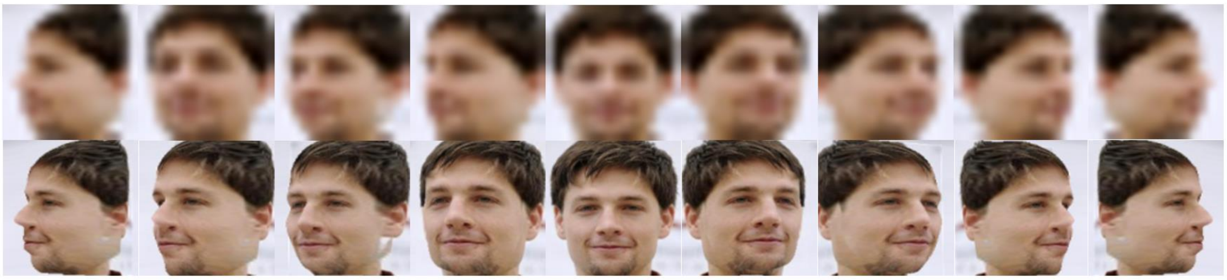


Fig. 9: Illustration of the synthesized non-frontal faces. First line: the synthesized non-frontal LR faces; second line: the corresponding HR ones.

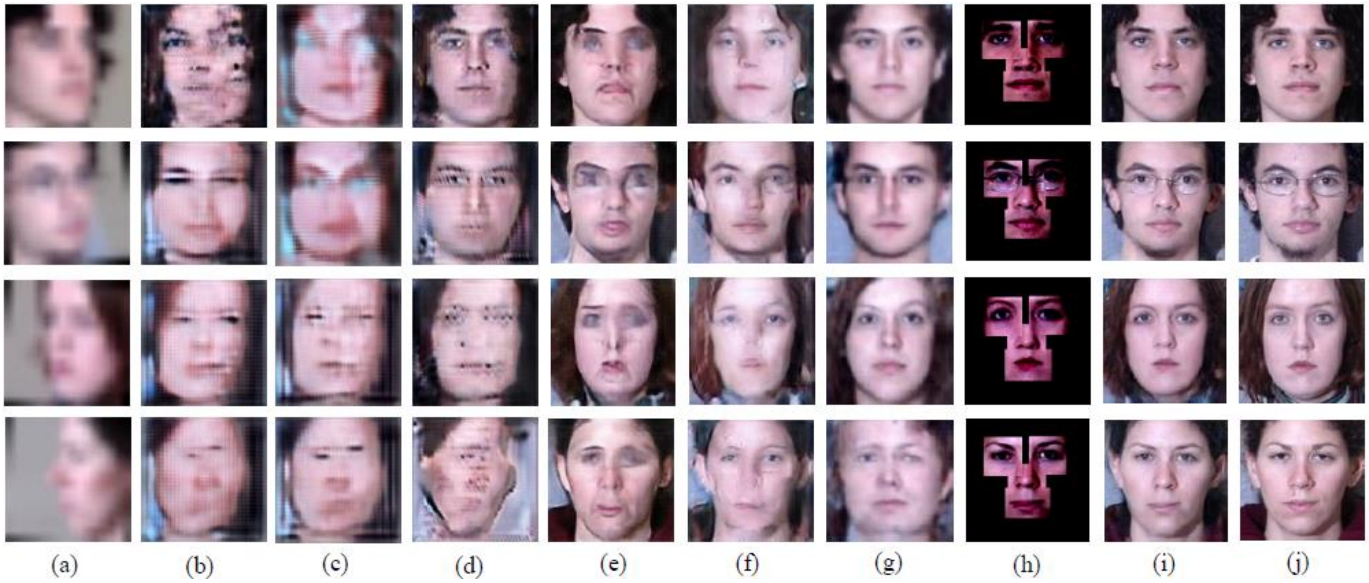


Fig. 10: Qualitative comparisons of state-of-the-art methods for the Multi-PIE database. Columns: (a) Unaligned LR inputs under various poses (from top row to bottom row, +60^o, +45^o, -75^o, -90^o). (b) Bicubic + [3] (c) [3] + [7]. (d) [3] + [59]. (e) [56] + [3]. (f) [55] + [3]. (g) [5]. (h) Fine-grained facial components. (i) VividGAN. (j) Ground truth.

for the training purpose, we use a popular 3D face model [35] to synthesize the non-frontal HR face images under eight different poses (±22°, ±40°, ±55°, ±75°). Afterwards, we manually transform and downsample these synthesized non-frontal HR faces to formulate the unaligned non-frontal LR images (16×16 pixels). Fig. 9 illustrates some synthesized non-frontal faces.

In this manner, we generate 18K and 7K frontal/non-frontal face pairs (frontal HR face/unaligned non-frontal LR face) for the CelebA and MMI databases, respectively. We choose 80 percent of the face pairs for training and 20 percent of the face pairs for testing. Under this setting, the training and testing sets do not overlap.

1) Compared methods: We conduct comparative experiments in three fashions.

- F+SR: face frontalization techniques (DRGAN [3], Hassner*et al.* [12]) + face SR methods (SRGAN [56], CBN [7], WaveletSRnet [59], TDAE [55]) (we use bicubic interpolation to adjust image sizes);
- SR+F: face SR methods (SRGAN [56], CBN [7], WaveletSRnet [59], TDAE [55]) + face frontalization techniques (DRGAN [3], Hassner*et al.* [12]);
- Joint SR+F: TANN [5] and our VividGAN.

In the first fashion (F+SR), we first frontalize the non-frontal LR faces, and then super-resolve the frontalized ones by state-of-the-art SR methods. In the second fashion (SR+F), we first super-resolve the non-frontal LR faces, and then frontalize the upsampled ones. In the third fashion (Joint SR+F), TANN [5] jointly achieves face SR and frontalization in a whole framework.

For a fair comparison, we retrained these models on our training dataset. Since SRGAN [56], CBN [7] and WaveletSRnet [59] cannot achieve face alignment during their upsampling procedure, we employ a STN [4] network to align the input unaligned LR faces. However, TANN [5], TDAE [55] and our VividGAN generate upright results automatically. Furthermore, compared to DRGAN [3] and Hassner*et al.* [12], our VividGAN does not need accompanying pose or facial landmark labels to serve as guidance information for face frontalization.

B. Subjective Evaluation

Fig.10 and 11 illustrate the visual results of the compared methods. The poses of input LR faces span across a spectrum from slight to large angles.

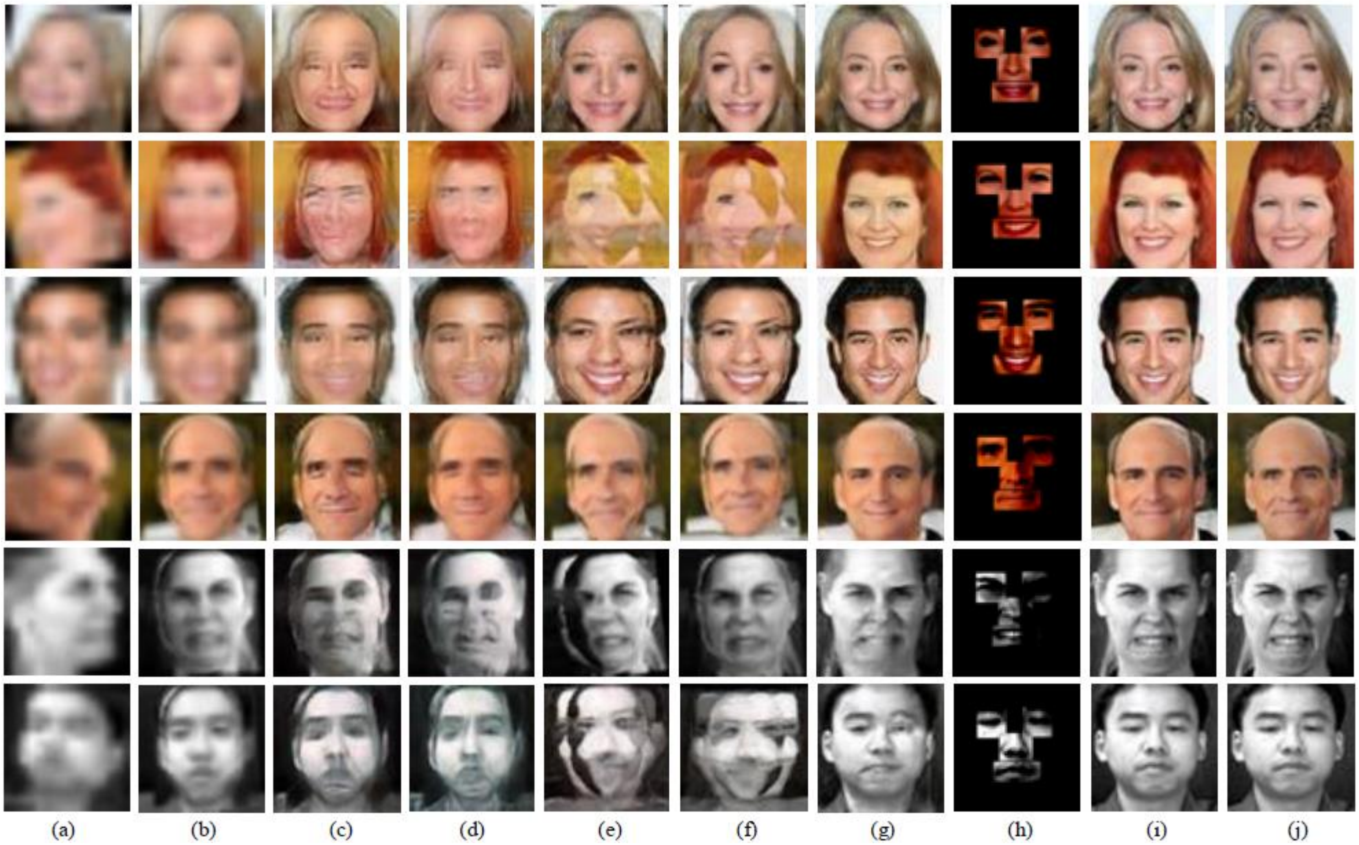


Fig. 11: Qualitative comparisons of state-of-the-art methods for the CelebA and MMI databases. Columns: (a) Unaligned LR inputs under various poses (from top row to bottom row, -40^o, -75^o, -22^o, +75^o, +75^o, -40^o). (b) Bicubic + [12]. (c) [12] + [7]. (d) [12] + [59]. (e) [56] + [12]. (f) [55] + [12]. (g) [5]. (h) Fine-grained facial components. (i) VividGAN. (j) Ground truth.

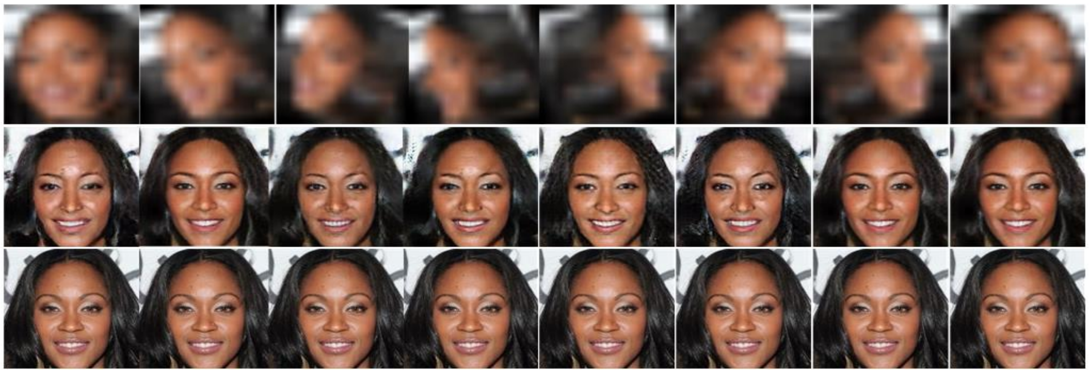


Fig. 12: Results of our VividGAN. First row: the input unaligned LR faces under various poses; second row: our hallucinated results; third row: the ground truth frontal HR face images.

As Fig. 10(b) and 11(b) show, the combination of bicubic interpolation and frontalization methods [3], [12] are handicapped to generate photo-realistic facial details. Since bicubic upsampling only interpolates new pixels from neighboring pixels without training, the produced non-frontal HR images always lack details. Thus, the following frontalization results are accompanied with more artifacts. In addition, the results also suffer from obvious skew artifacts. This indicates that it is difficult to align LR faces accurately by a simple STN network. Benefiting from the multiple STN layers embedded in our coarse FH network, along with the supervision of the class-wise discriminative loss, VividGAN generates upright frontal HR faces, as seen in Figs. 10(i) and 11(i).

As discussed in Sec. III, face frontalization for LR inputs is a challenging task. This is reflected by the results of the first combination fashion (F+SR) (see Figs. 10(c), 10(d), 11(c) and 11(d)), where the face appearances are accompanied with distorted contours and ghosting artifacts. This phenomenon is caused by the failed SR procedure on the defective frontalized faces. However, benefiting from the mirror symmetry loss, our coarse FH network succeeds in preserving the content-integrity of the frontalized faces, which benefits the following hallucination step. Thus, our VividGAN achieves satisfied results (see Figs. 10(i) and 11(i)) with well-maintained facial

TABLE I: Average PSNR [dB] and SSIM values of the compared methods on the whole testing set.

SR Method	Multi-PIE		MMI		CelebA		Multi-PIE		MMI		CelebA	
	F+SR						SR+F					
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	18.991	0.730	19.125	0.738	20.956	0.799	21.352	0.810	21.670	0.812	21.055	0.802
SRGAN [56]	19.235	0.733	19.958	0.742	20.072	0.794	21.562	0.819	21.328	0.809	20.996	0.800
CBN [7]	20.251	0.752	20.006	0.749	19.784	0.783	21.974	0.796	20.846	0.792	20.073	0.789
WaveletSRnet [59]	21.164	0.798	20.752	0.742	21.079	0.805	22.839	0.837	22.005	0.824	21.831	0.817
TDAE [55]	20.032	0.749	20.353	0.753	20.353	0.792	22.649	0.824	22.124	0.836	20.708	0.796
TANN [5]	24.426	0.831	23.748	0.845	25.690	0.870	24.426	0.831	23.748	0.845	25.690	0.870
VividGAN	**26.289**	**0.876**	**26.452**	**0.881**	**26.952**	**0.893**	**26.289**	**0.876**	**26.452**	**0.881**	**26.952**	**0.893**

TABLE II: Quantitative evaluations on different out-of-plane rotation degrees on CelebA and MMI testing sets.

	SR Method	±22°		±40°		±55°		±75°	
		CelebA	MMI	CelebA	MMI	CelebA	MMI	CelebA	MMI
F+SR	Bicubic	21.849/0.818	20.052/0.746	21.032/0.802	19.233/0.729	20.901/0.793	18.036/0.711	20.452/0.786	18.021/0.702
	SRGAN [56]	21.080/0.786	20.991/0.785	20.562/0.773	20.148/0.756	20.073/0.796	19.200/0.731	20.066/0.789	19.005/0.724
	CBN [7]	20.683/0.792	20.984/0.721	19.991/0.785	20.571/0.713	19.392/0.779	20.352/0.708	19.183/0.771	20.003/0.699
	WaveletSRnet [59]	22.004/0.812	22.341/0.798	21.358/0.804	21.860/0.775	20.996/0.799	21.241/0.739	20.742/0.794	20.975/0.726
	TDAE [55]	21.758/0.813	21.042/0.767	21.041/0.796	20.248/0.735	20.425/0.789	20.015/0.724	20.002/0.781	19.997/0.718
SR+F	Bicubic	21.648/0.827	22.341/0.800	21.167/0.804	21.976/0.788	21.004/0.798	21.278/0.773	20.982/0.793	20.904/0.767
	SRGAN [56]	21.877/0.825	22.109/0.787	21.032/0.803	21.893/0.787	20.989/0.791	21.241/0.762	20.335/0.784	21.036/0.756
	CBN [7]	21.004/0.804	22.843/0.776	20.838/0.799	22.132/0.755	20.104/0.791	21.947/0.776	19.893/0.779	21.895/0.733
	WaveletSRnet [59]	22.951/0.839	24.005/0.806	22.004/0.823	23.346/0.797	21.752/0.815	22.901/0.788	21.004/0.808	22.536/0.778
	TDAE [55]	21.728/0.804	22.903/0.802	21.003/0.799	22.142/0.788	20.652/0.795	22.096/0.781	20.126/0.785	22.002/0.774
Joint	TANN [5]	26.109/0.880	24.423/0.808	25.710/0.870	23.752/0.798	25.362/0.870	23.004/0.791	25.025/0.870	22.821/0.782
SR+F	**VividGAN**	**27.948/0.908**	**26.797/0.885**	**27.763/0.899**	**26.301/0.869**	**26.841/0.887**	**26.012/0.850**	**26.065/0.886**	**25.979/0.839**

structure as well as identity-preserving details.

Since current face SR methods [7], [25], [55], [56] fail to super-resolve the input faces under large pose variations and produce results with false artifacts and broken structures, which largely deteriorate the following face frontalization procedure. In this way, the second combination fashion (SR+F) also fails to recover authentic face details (see Figs. 10(e), 10(f), 11(e) and 11(f)). However, as verified in Sec. IV, benefiting from the estimated facial structure prior (see Fig. 10(h) and 11(h)), VividGAN is able to reconstruct fine facial details, especially when the LR inputs are accompanied with large poses (*e.g.,* -90^o, the forth line in Fig. 10(i)).

TANN [5], as the first proposed framework to jointly perform face SR and frontalization in a whole framework, can generate relative satisfactory results, shown in Figs. 10(g) and 11(g). However, due to its single-stage mechanism, when the input LR faces are accompanied with extreme poses (*e.g.,* the forth line in Fig. 10(a)) or complex expressions (*e.g.,* the fifth and sixth lines in Fig. 11(a)), TANN fails to preserve realistic facial details. In contrast, our VividGAN is able to revise facial details and recover visually appealing faces (*e.g.,* "weird mouth" (the fifth line in Fig. 11(i)) and "slight closed eyes" (the sixth line in Fig. 11(i))).

Above all, our VividGAN can generate photo-realistic frontal HR faces from very LR inputs span across various poses and expressions, showing in Figs. 10(i), 11(i) and 12.

C. Objective Evaluation

The above qualitative performance are confirmed by quantitative evaluations. We report average PSNR and SSIM values for the compared methods on the testing set, in two different perspectives (see Tabs. I and II).

As can be seen from Tab. I, our VividGAN achieves remarkable results than the rest in both indoor and in the wild databases. Specially, on the MMI testing set, VividGAN outperforms the second best technique, TANN, with a large margin of approximate 2.7 dB in PSNR. This result corroborates the subjective finding for the inputs under complex expressions. Thanks to our coarse-to-fine Vivid-FHnet, which recovers photo-realistic facial details, resulting in superior quantitative performance.

In Tab. II, the first and second numbers denote PSNR and SSIM scores, respectively. It is apparent from Tab. II that VividGAN achieves impressive performance in all the rotation degrees. Closer inspection of the table shows that the results of VividGAN in extremely high rotation degrees still get large-scale increases than the state-of-the-art methods. This also verifies that VividGAN can attain photo-realistic hallucinated results when the input LR faces are accompanied with extreme poses. This satisfied performance can be attributed to the facial structure prior knowledge. As a result, VividGAN does not degrade seriously as the rotation degree increases.

D. Ablation Analysis

We report the performance of different VividGAN variants, which are trained with various loss combinations, on MultiPIE and CelebA (see Tab. III and Fig. 6). With a slight abuse of notation, we denote the compared VividGAN variants as: (*i*) L_G^-: L_{mse}, L_h; (*ii*) $L_G^\dagger$: L_{mse}, L_{id} and L_h; (*iii*) $L_G^\ddagger$: L_{mse}, L_{id}, L_h and L_{sys}; (*iv*) $L_G^\star$: L_{mse}, L_{id}, L_h, L_{sys} and L_{adv}^c; (*v*) L_G: L_{mse},

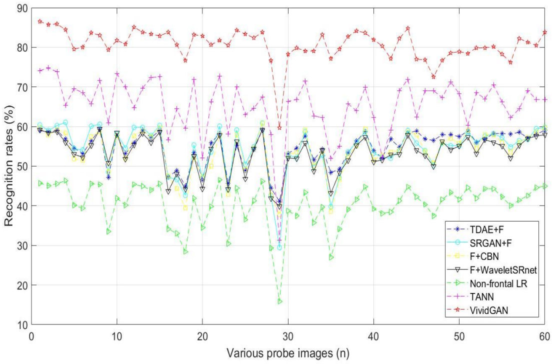


Fig. 13: Face recognition rates for the compared methods on varying gallery sets.

L_{id}, L_h, L_{sys}, L^c_{adv} and L^f_{adv}. Note that L_h is the prerequisite constrain for our facial component-aware module.

As demonstrated in Tab. III, only adopting L_{mse} leads to unsatisfied quantitative results (L^-_G in Tab. III). Then, the identity similarity loss L_{id} not only improves the visual quality (see Fig. 6(e)) but also increases the quantitative performance ($L^\dagger_G$ in Tab. III). This is due to the factor that L_{id} forces the high-order moments of the hallucinated faces, *i.e.,* feature maps of faces, to be similar to their ground-truths. In addition, we also verify the effectiveness of the mirror symmetry loss L_{sys}, which are formulated based on the domain-specific knowledge on human faces. As indicated in Tab. III ($L^\ddagger_G$), the introduction of L_{sys} leads to better quantitative performance. Moreover, since the adversarial loss L^c_{adv} promotes the face alignment and frontalization performance for the coarse hallucinated faces, the results of $L^\star_G$ are more superior than the former ones. Finally, we confirm that the two-level discriminators boost the hallucination performance significantly, resulting the highest quantitative performance (L_G in Tab. III).

TABLE III: Ablation study of various loss combinations

	Multi-PIE		CelebA	
	PSNR	SSIM	PSNR	SSIM
L^-_G	23.017	0.779	22.945	0.773
$L^\dagger_G$	24.332	0.803	23.981	0.798
$L^\ddagger_G$	25.998	0.845	25.057	0.804
$L^\star_G$	26.055	0.874	26.128	0.885
L_G	26.289	0.876	26.952	0.893

VI. Discussions

In this section, we demonstrate that our VividGAN benefits the downstream tasks, *i.e.,* face recognition and expression classification.

A. Comparisons with SoA on Face Recognition

We employ a state-of-the-art pretrained face recognition model (SphereFaceNet [60]) to conduct face recognition experiments on the Multi-PIE database. We choose the images of 50 different individuals, including twelve poses and five severe illumination conditions, to form the probe set. Then, all the face images in the probe set are down-sampled and then hallucinated by the compared methods. The frontal HR image under normal illumination of each individual is selected as the ground truth to construct the gallery set. For each face image, we extract the deep features (SphereFace) [60] from the output of the FC_1 layer in SphereFaceNet model. Tab. IV indicates the face recognition results of the compared methods.

Specially, for the sake of fair comparison, we also verify the compared techniques based on varying gallery sets to get the average recognition rates (see Fig. 13). In this setting, the gallery face image for each individual is varied from the first hallucinated face to the last one. The number of experiment times is equal to that of the hallucinated faces per individual in the probe set.

Above all, our VividGAN improves the face recognition performance significantly compared to the other methods. Specially, the face recognition rate of our hallucinated faces outperforms that of the original non-frontal LR faces by a large margin. Thus, our VividGAN has remarkable identity preservation ability, which substantially satisfies the need of the downstream face recognition task.

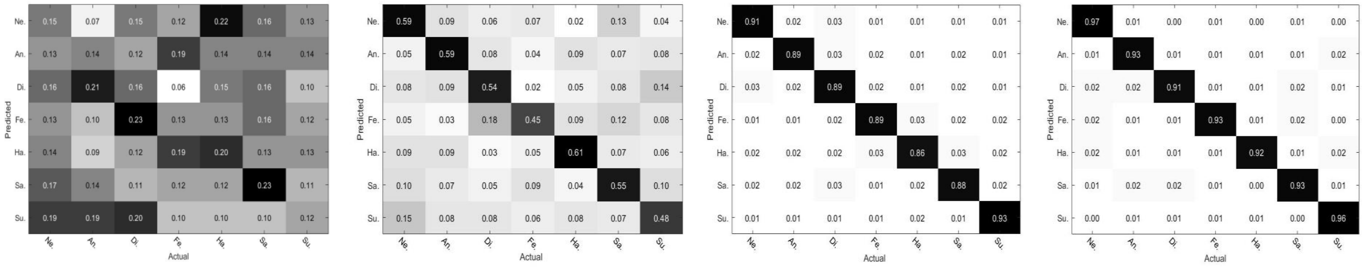
TABLE IV: Face recognition results for different methods on the Multi-PIE database.

SR method	Accuracy	
	F+SR	SR+F
Bicubic	52.33%	51.27%
SRGAN [56]	53.48%	52.49%
CBN [7]	52.89%	51.21%
WaveletSRnet [59]	55.24%	54.74%
TDAE [55]	54.31%	55.30%
TANN [5]	68.52%	
Non-frontal LR	40.26%	
Frontal HR	97.89%	
VividGAN	82.28%	

B. Comparisons with SoA on Face Expression Classification

Furthermore, our VividGAN also achieves significant improvement in face expression classification performance.

We perform a standard 10-fold subject-independent cross-validation on the MMI database. First, the synthesized frontal/non-frontal MMI face pairs are split into 10 subsets according to the identity information via an ascending order. Thus, the individuals in any two subsets are mutually exclusive. For each run, 9 subsets are employed for training and the remaining one subset for testing. Here, we retrain all the compared hallucination models and a state-of-the-art expression classification model, VGG-VD-16 [61]. We perform such 10 runs by enumerating the subset used for testing. Then, in each validation branch, we use the fine-tuned VGG-VD-16 [61] model to get the expression classification results for all the testing face samples. Here, the testing face samples include the hallucinated frontal HR faces by the compared hallucination models, the non-frontal LR faces (upsampled by bicubic interpolation) and the ground truth frontal HR faces. Finally, the expression classification results are computed as



(a) Non-frontal LR images (b) TANN results (c) VividGAN results (d) Frontal HR images

Fig. 14: The confusion matrices between the true labels and different predicted labels on the MMI database.

the average of the 10 runs (see Tab. IV). Fig. 14 also provides the corresponding confusion matrices.

Obviously, as can be clearly seen in Tab. IV and Fig. 14, the results of our VividGAN achieve superior facial expression classification performance, which exceeds the other methods. This indicates that VividGAN recovers photo-realistic facial details faithfully.

TABLE V: Facial expression classification results for different methods on the MMI database

SR method	Accuracy	
	F+SR	SR+F
Bicubic	26.08%	27.95%
SRGAN [56]	31.26%	33.02%
CBN [7]	29.30%	30.17%
WaveletSRnet [59]	35.58%	37.34%
TDAE [55]	34.72%	35.69%
TANN [5]	56.70%	
Non-frontal LR	23.46%	
Frontal HR	94.62%	
VividGAN	88.53%	

C. User Study

To demonstrate the effectiveness and superiority of VividGAN, we conduct a user study.

We choose fifty CelebA female as well as fifty CelebA male LR faces with various poses as testing samples. For each testing sample, six different hallucinated HR frontal face images are shown at the same time where one is generated by VividGAN and the others are generated by the other compared methods, *i.e.,* TANN and other four combination methods. The participants are required to choose the most similar one with respect to the ground-truth image. Twenty participants are invited to finish the user study. Finally, we collect 2000 votes from the participants and show the percentage of votes for each compared method in Fig. 15. The result shows that the hallucinated faces obtained by our VividGAN are preferred more often than those of other methods.

VII. Conclusion

This paper presents an end-to-end trainable VividGAN framework to jointly super-resolve and frontalize tiny non-frontal face images, in a coarse-to-fine fashion. VividGAN fully utilizes the facial prior knowledge and achieves face SR (an upscaling factor of 8×) along with frontalization (with pose variations up to 90^o). Moreover, we also provide a solution for fine-grained facial component hallucination directly from non-frontal LR faces. With the aid of the two-level discriminators, our network realizes a preview and revise mechanism to generate visually appealing results. Experimental results validate the effectiveness of VividGAN, which yields identity-preserving faces and substantially boosts the performance of downstream tasks, *i.e.,* face recognition and expression classification.

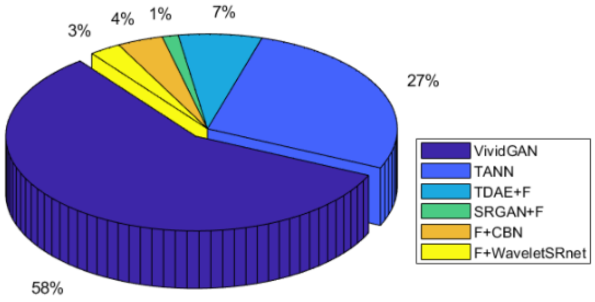


Fig. 15: Evaluation of user study on the test samples.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No.61871123), Key Research and Development Program in Jiangsu Province (No.BE2016739) and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions. This work was supported in part by ARC under Grant DP180100106 and DP200101328.

References

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, “Face recognition: A literature survey,” *ACM computing surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[2] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, “Face super-resolution guided by facial component heatmaps,” in *ECCV*, 2018, pp. 217–233.

[3] L. Q. Tran, X. Yin, and X. Liu, “Representation learning by rotating your faces,” *IEEE TPAMI*, 2018.

[4] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, “Spatial transformer networks,” in *NIPS*, 2015, pp. 2017–2025.

[5] 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*, 2019.

[6] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *ECCV*, 2016, pp. 318–333.

[7] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *ECCV*, 2016, pp. 614–630.

[8] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li, "Attention-aware face hallucination via deep reinforcement learning," in *CVPR*, 2017, pp. 690–698.

[9] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "Fsrnet: End-to-end learning face super-resolution with facial priors," in *CVPR*, 2018, pp. 2492–2501.

[10] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *CVPR*, 2016, pp. 2387–2395.

[11] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust statistical face frontalization," in *CVPR*, 2015, pp. 3871–3879.

[12] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *CVPR*, 2015, pp. 4295–4304.

[13] 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 *ICCV*, 2017, pp. 2439–2448.

[14] V. Blanz, T. Vetter *et al.*, "A morphable model for the synthesis of 3d faces." in *Siggraph*, vol. 99, no. 1999, 1999, pp. 187–194.

[15] 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 *CVPR*, 2015, pp. 787–796.

[16] 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.

[17] 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.

[18] 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.

[19] C. Liu, H.-Y. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *IJCV*, vol. 75, no. 1, pp. 115–134, 2007.

[20] S. Kolouri and G. K. Rohde, "Transport-based single frame super resolution of very low resolution face images," in *CVPR*, 2015, pp. 4876–4884.

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

[22] 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.

[23] M. F. Tappen and C. Liu, "A bayesian approach to alignment-based image hallucination," in *ECCV*, 2012, pp. 236–249.

[24] C.-Y. Yang, S. Liu, and M.-H. Yang, "Hallucinating compressed face images," *IJCV*, vol. 126, no. 6, pp. 597–614, 2018.

[25] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution," in *ICCV*, 2017, pp. 1689–1697.

[26] 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 the IEEE International Conference on Computer Vision*, 2017, pp. 251–260.

[27] R. Dahl, M. Norouzi, and J. Shlens, "Pixel recursive super resolution," in *ICCV*, 2017, pp. 5439–5448.

[28] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," *arXiv preprint arXiv:1601.06759*, 2016.

[29] 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 the European Conference on Computer Vision (ECCV)*, 2018, pp. 185–200.

[30] F. Yang, J. Wang, E. Shechtman, L. Bourdev, and D. Metaxas, "Expression flow for 3d-aware face component transfer," *ACM transactions on graphics (TOG)*, vol. 30, no. 4, p. 60, 2011.

[31] D. Jiang, Y. Hu, S. Yan, L. Zhang, H. Zhang, and W. Gao, "Efficient 3d reconstruction for face recognition," *Pattern Recognition*, vol. 38, no. 6, pp. 787–798, 2005.

[32] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith, "Fully automatic pose-invariant face recognition via 3d pose normalization," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 937–944.

[33] T. Hassner, "Viewing real-world faces in 3d," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3607–3614.

[34] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[35] I. Masi, A. T. Tran, T. Hassner, J. T. Leksut, and G. G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" 2016.

[36] V. Blanz and T. Vetter, "Face recognition based on fitting a 3d morphable model," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003.

[37] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Recover canonical-view faces in the wild with deep neural networks," *arXiv preprint arXiv:1404.3543*, 2014.

[38] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, "Rotating your face using multi-task deep neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 676–684.

[39] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman, "Face synthesis from facial identity features," *arXiv preprint arXiv:1701.04851*, 2017.

[40] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8398–8406.

[41] J. Zhao, J. Li, X. Tu, F. Zhao, Y. Xin, J. Xing, H. Liu, S. Yan, and J. Feng, "Multi-prototype networks for unconstrained set-based face recognition," *arXiv preprint arXiv:1902.04755*, 2019.

[42] M. Kan, S. Shan, H. Chang, and X. Chen, "Stacked progressive autoencoders (spae) for face recognition across poses," in *CVPR*, 2014, pp. 1883–1890.

[43] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni, "Regressing robust and discriminative 3d morphable models with a very deep neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5163–5172.

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

[45] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun, "Towards high fidelity face frontalization in the wild," *International Journal of Computer Vision*, pp. 1–20, 2019.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[47] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[48] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *CVPR*, 2018, pp. 3291–3300.

[49] 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.

[50] M. F. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in *Proceedings of Int'l Conf. Language Resources and Evaluation, Workshop on EMOTION*, Malta, May 2010, pp. 65–70.

[51] 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.

[52] Y. Zhang, P. Lv, X. Lu, and J. Li, "Face detection and alignment method for driver on highroad based on improved multi-task cascaded convolutional networks," *Multimedia Tools and Applications*, pp. 1–19, 2019.

[53] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016, pp. 483–499.

[54] 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*, 2018, pp. 7794–7803.

[55] X. Yu and F. Porikli, "Face hallucination with tiny unaligned images by transformative discriminative neural networks," in *AAAI*, 2017.

[56] 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 *CVPR*, 2017, pp. 4681–4690.

[57] 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.

[58] 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 the IEEE International Conference on Computer Vision*, 2017, pp. 1021–1030.

[59] 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.

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

[61] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.