Face Hallucination by Attentive Sequence Optimization with Reinforcement Learning

Yukai Shi, Guanbin Li, *Member, IEEE*, Qingxing Cao, Keze Wang, and Liang Lin, *Senior Member, IEEE*

Abstract—Face hallucination is a domain-specific super-resolution problem that aims to generate a high-resolution (HR) face image from a low-resolution (LR) input. In contrast to the existing patch-wise super-resolution models that divide a face image into regular patches and independently apply LR to HR mapping to each patch, we implement deep reinforcement learning and develop a novel attention-aware face hallucination (Attention-FH) framework, which recurrently learns to attend a sequence of patches and performs facial part enhancement by fully exploiting the global interdependency of the image. Specifically, our proposed framework incorporates two components: a recurrent policy network for dynamically specifying a new attended region at each time step based on the status of the super-resolved image and the past attended region sequence, and a local enhancement network for selected patch hallucination and global state updating. The Attention-FH model jointly learns the recurrent policy network and local enhancement network through maximizing a long-term reward that reflects the hallucination result with respect to the whole HR image. Extensive experiments demonstrate that our Attention-FH significantly outperforms the state-of-the-art methods on in-the-wild face images with large pose and illumination variations.

Index Terms—Face hallucination, reinforcement learning, recurrent neural network

1 Introduction

Face hallucination, a.k.a. facial image super-resolution, aims to generate a high-resolution (HR) face image from a given low-resolution (LR) input. Face hallucination is a fundamental problem in the field of face analysis and has drawn considerable research attention due to the need for solving this problem in various face-related tasks, including face attribute recognition [1], face alignment [2], [3] and face recognition [4], under complex low-quality real-world scenarios.

As a special form of general image super-resolution, obvious structural prior information exists in face images, which is therefore widely used in existing face hallucination algorithms [5], [6], [7]. Face prior information is usually embedded into the existing face hallucination models in the form of face component analysis [8], facial correspondence field [7] and facial landmark localization [6]. However, the calculation of this prior information requires additional calculations, and accurate parsing of the landmarks is difficult in low-resolution situations. Therefore, there is a series of work attempts to replace the fine-grained face prior computation with rough patch-wise super-resolving mapping, which can improve the efficiency of the algorithm while achieving comparable performance. Due to differences in appearance of facial organs and the natural symmetry of the facial region, existing patch-wise face hallucination methods either extract patches from detected facial landmarks or simply divide the face image into even patches and then independently perform LR to HR mapping on each detected patch [9], [10], [11], [12]. Specifically, end-to-end deep convolutional networks (CNNs) have recently achieved great success in learning discriminative patch-to-patch mapping from LR images to HR images [7], [13]. However, face structure priors and spatial configurations [14], [15] are often treated as external information, and the contextual dependencies among the super-resolution reconstruction of each patch are usually ignored during the hallucination processing.

The difficulty of super-resolution reconstruction also varies due to the inconsistency in the degree of detail deterioration of each facial part. On the other hand, the symmetry of the human face and the similarity in the appearance of the adjacent regions make previously hallucinated HR patches worthy of reference to the latter. Therefore, during the process of face hallucination, the reconstruction sequence of patches and the selection of patch locations at each step are crucial for global face hallucination, which is consistent with the human visual perception mechanism. When people observe a scene object, they usually start with perceiving the whole image and successively explore a sequence of regions via the attention shifting mechanism rather than separately processing the local regions. This finding motivates us to explore a new pipeline for face hallucination by sequentially searching for the local attention regions and considering their contextual dependency from a global perspective.

Inspired by the effectiveness of recurrent visual attention modeling for visual analysis and understanding [16], [17],

• *The authors are with the School of Data and Computer Science, Sun Yat-Sen University , Guangzhou 510006, China. E-mail: {shiyk3, caoqx}@mail2.sysu.edu.cn, liguanbin@mail.sysu.edu.cn, kezewang@gmail.com, linliang@ieee.org.*

Manuscript received 24 Apr. 2018; revised 29 Apr. 2019; accepted 30 Apr. 2019. Date of publication 7 May 2019; date of current version 1 Oct. 2020. (Corresponding author: Guanbin Li.)
Recommended for acceptance by T. Hassner.
Digital Object Identifier no. 10.1109/TPAMI.2019.2915301

[18], [19], we propose an attention-aware face hallucination (Attention-FH) framework that fully exploits the global contextual information of the face to recurrently discover and enhance a series of local face regions. Specifically, accounting for the diverse characteristics of face images in terms of blur, pose, illumination and facial appearance, we model the face hallucination problem as a strategy optimization problem for patch sequence selection and implement a search for an optimal enhancement route. We resort to the deep reinforcement learning (RL) model [20] to sequentially determine local patches for enhancement, as the RL technique has shown promising results on decision-making problems without the need for supervision information at each step.

Specifically, our Attention-FH framework jointly optimizes a recurrent policy network that learns to identify the facial region to be hallucinated at each time step and a local enhancement network for facial part super-resolution by considering the whole face image with previously enhanced parts. In our framework, rich correlation cues among different facial parts are explicitly exploited to guide the current region assignment, while past hallucination results are incorporated as a global reference during local enhancement in each step. In this way, the agent can make full use of the symmetry of the human face and the adjacent regions to assist in obtaining more accurate facial part hallucination reconstruction. For example, the agent can improve the enhancement of the right eye region by taking a clear version of the left eye region as reference.

Instead of performing supervision in each step, we employ a single global reward for RL, which measures the overall performance of the entire hallucinated HR face. The optimization of the recurrent policy network is updated following the RL algorithm [21], which can be treated as a Markov decision process (MDP) maximized with a long-term global reward. At each time step, the policy network learns to determine the location and the size of an optimal rectangular facial region by conditioning on the whole face image with all previously enhanced results and the encoded action history. A gated recurrent unit (GRU) layer is employed to encode the information from the previously attended facial regions. All the previously determined regions are also recorded to avoid duplicated selection of a region in a recurrent mode.

Given the attended facial region in each step, the local enhancement network is trained for hallucination reconstruction, and its loss is defined as the L_2 distance between the part hallucination result and the specific ground truth. Compared with whole-face super-resolution, the structure of facial components (attended parts) is stable and easy to restored. Notably, the supervision information from the enhancement of facial parts also effectively reduces unnecessary trial and error during the reinforcement optimization.

We compare the proposed Attention-FH approach with state-of-the-art face hallucination methods under both constrained and unconstrained settings. Extensive experiments show that our method substantially outperforms all the alternative methods. Moreover, our framework can explicitly generate a sequence of attentional regions during the hallucination, which finely accords with the human perception process and to some extent provides an interpretable mechanism in the hallucination recovery process.

A preliminary version of this work is published in [22]. In this work, we inherit the idea of exploring the interdependency of facial components and redevelop the policy network from the perspective of the attention mechanism and reinforcement optimization. The improvement upon the initial version includes size-free attention and a new reward function designed to maximize the stability of the RL. We have also added a comprehensive discussion of the design of the local enhancement network and greatly improved its efficiency while maintaining its performance. Moreover, we present more comparisons with state-of-the-art models and a more comprehensive ablation study on our proposed framework.

The rest of this paper is organized as follows. Section 2 reviews related work on face hallucination and deep RL. In Section 3, we introduce our proposed Attention-FH model. Section 4 provides extensive performance evaluation and comparisons with state-of-the-art models. Finally, we conclude this paper in Section 5.

2 Related Work

Face Hallucination and Image Super-Resolution. Face hallucination is a domain-specific image super-resolution problem proposed to map a LR facial image to its HR version. With obvious prior knowledge, face hallucination methods are required to handle extremely degraded faces and restore complex structural information. Early approaches hypothesized that corrupted faces are in a relatively controlled environment with small variations. Yang et al. [9] enforced LR and HR images to have similar sparse representations and implemented image super-resolution by taking into account the sparsity prior. Wang et al. [23] decomposed faces into different frequency bands and hallucinated faces by eigentransformation. By contrast, structured face hallucination (SFH) [10] pre-aligns faces and establishes the mapping between facial components. SFH not only achieves impressive results but also reveals that facial components are crucial in face hallucination. However, SFH relies heavily on pre-alignment, making it difficult to cope with situations where illumination and pose changes are considerable. More recently, deep neural networks have shown impressive performance in face hallucination and image restoration [13], [15], [24], [25], [26]. Dong et al. [15] employed a fully convolutional network (FCN) for image SR. Kim et al. [24] made a very deep CNN for image SR trainable by adopting a highway connection. Reed et al. [27] proposed an efficient sampling strategy to demonstrate high-quality image reconstruction. Ulyanov et al. [28] further demonstrated that deep neural networks can make full use of prior knowledge of the image itself to rebuild a corrupted image. Shocher et al. [26] employed a novel internal learning approach to fully explore LR inputs. UR-DGN [29] is claimed to be the first face SR method that uses a generative adversarial network. Tuzel et al. [25] claimed that global information is crucial to face hallucination and established a local and global network to restore faces.

However, all these existing deep learning models attempt to improve the performance of face hallucination

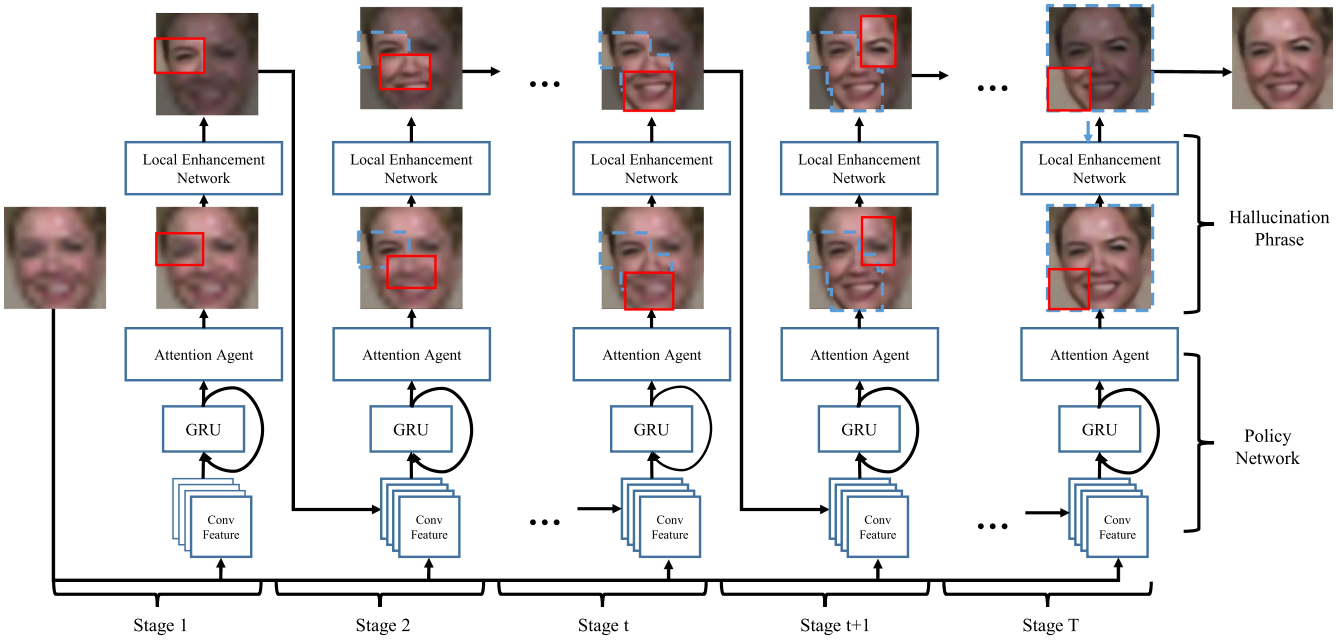


Fig. 1. Pipeline of Attention-FH. The proposed framework contains two modules: the policy network and the local enhancement network. In each step, the attention agent glimpses the whole image and provides actions. Each action indicates the center position of the next attended rectangular region and the size of the bounding box. The attended patch is further fed as input to a local enhancement network for super-resolution. We use the red solid bounding boxes and the blue dashed bounding boxes to indicate the attended patch and enhanced patch, respectively. A global reward is given at the end of the sequence to enforce the policy network to learn the optimal restoring route. With the two components, we can perform a coarse-to-fine face hallucination paradigm.

by designing deeper and more complex neural network structures. In this work, we start from the perspective of human cognition and model the face hallucination process as a patch-wise local reconstruction problem. We introduce a deep RL-based optimization method to learn a series of ordered patch hallucination sequences. For patch-wise hallucination,which is not our main focus, we draw on an existing FCN-based framework and incorporate it into our Attention-FH model for end-to-end training. We firmly believe that the Attention-FH framework proposed in this paper is compatible with any existing deep super-resolution models and will benefit from the future improvement of super-resolution algorithms.

Attention and Reinforcement Learning. Visual attention modeling is inspired by the human visual perception system. Visual attention modeling is widely embedded in existing deep neural networks in the form of adaptive feature weighting or salient region localization and has been proved to be effective in improving the performance of a series of computer vision tasks, including object proposal [30], object classification [31], relationship detection [32], image captioning [33] and visual question answering [34]. Some works have exploited RL to optimize the attention networks to address the problem that the coordinates of attended regions are not differentiable. For example [35] and [36], learned an agent that actively locates the target regions (face or objects) instead of exhaustively sliding subwindows on images. Goodrich et al. [35] defined 32 actions to shift the focal point and reward the agent when spotting the target. Caicedo et al. [36] defined an action set that contains several transformations of the bounding box and rewarded the agent if the bounding box became closer to the ground truth in each step. Both of these methods learned an optimal policy to locate the target through Q-learning.

3 Methodology

3.1 Inference Overview

We develop an Attention-FH to perform face hallucination in a coarse-to-fine manner. Specifically, Attention-FH is composed of two parts: a recurrent policy network that learns to adaptively locate a particular facial region for local hallucination, and a local enhancement network that directly learns to map a located facial patch with LR to its HR version, considering both the local and global perspective.

Given a face image I_{lr} with LR, the target of our proposed Attention-FH framework is to generate the corresponding HR version I_{hr} through a series of iterative local patch enhancements, which can be formulated as:

$$I_{hr} = F(I_{lr}|\theta), \tag{1}$$

where θ denotes the parameters of the Attention-FH model, and F is the whole hallucinating procedure.

Given a state s, the recurrent policy network is learned to predict the actions l_t, including the center position of the next attended rectangular region and the size of the bounding box. The attended patch is further cropped and fed as input to a local enhancement network for super-resolution. This process can be formulated as:

$$\begin{aligned} l_t &= f_\pi(s_{t-1};\theta_\pi), \\ \hat{I}^{l_t}_{t-1} &= g(l_t, I_{t-1}), \end{aligned} \tag{2}$$

where θ_π is the parameters of the recurrent policy network, f_π indicates the recurrent policy network and I_{t-1} denotes the restored face image at step $t-1$. The state s_{t-1} is a vector, encoded with the current state, and g refers to the cropping operation, which is applied to crop the corresponding patch $\hat{I}^{l_t}_{t-1}$ given action l_t.

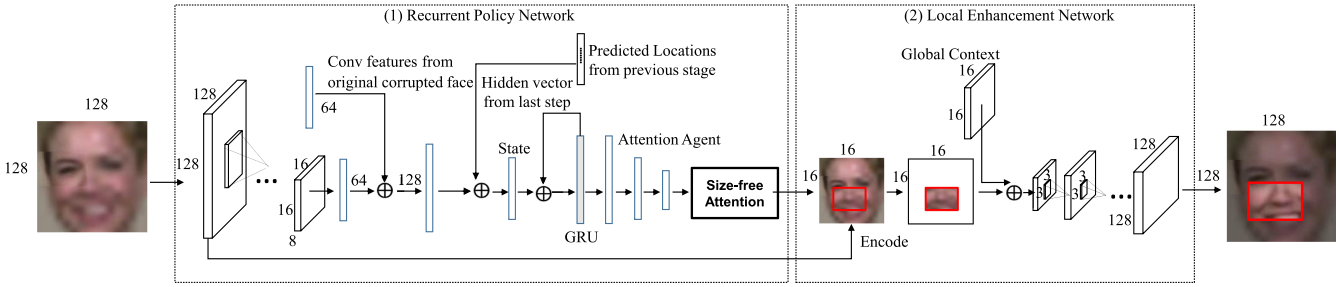


Fig. 2. Network architecture of our Attention-FH. At each time step t, the recurrent policy network outputs the actions by observing the original corrupted face image I_0, the current enhanced face image I_{t-1} and the historical actions v_l. Meanwhile, v_l is encoded by latent variables (e.g., 64 hidden states) of GRU. I_0 and I_{t-1} are first represented as high-level features from an output feature vector of a fully connected layer and are then concatenated to form a vector of 128 dimensions. The above three corresponding pieces of information constitute the state s. The GRU layer learns to infer the action probabilities by considering s. The output probabilities are formulated by a fully connected linear layer (128 neurons) for all candidate actions. Given the actions, we can obtain the local patch $\hat{I}_{t-1}^{l_t}$. Then, $\hat{I}_{t-1}^{l_t}$ is sent to the local enhancement network for hallucinating, which results in the enhanced patch $\hat{I}_t^{l_t}$. Finally, a new hallucinated face is generated by replacing $\hat{I}_{t-1}^{l_t}$ with $\hat{I}_t^{l_t}$.

Given attention patch $\hat{I}_{t-1}^{l_t}$, the local enhancement network f_e is adopted for hallucination.

$$\hat{I}_t^{l_t} = f_e(\hat{I}_{t-1}^{l_t}, I_{t-1}; \theta_e). \tag{3}$$

where θ_e is the parameter of the local enhancement network and $\hat{I}_t^{l_t}$ indicates the enhanced patch. After local enhancement, we replace $\hat{I}_{t-1}^{l_t}$ with $\hat{I}_t^{l_t}$. Specifically, the restored face image I_t for the next step t is produced by replacing the existing patch $\hat{I}_{t-1}^{l_t}$ with the enhanced patch $\hat{I}_t^{l_t}$. The overall coarse-to-fine face hallucination can be defined as:

$$\begin{cases} I_0 = I_{lr} \\ I_t = f(I_{t-1}; \theta) \quad 1 \le t \le T, \\ I_{hr} = I_T \end{cases} \tag{4}$$

where I_0 is the original corrupted face image, I_t indicates the enhanced full-sized facial image at each step t, T is the maximal recurrent step, $\theta = [\theta_\pi; \theta_e]$ and $f = [f_\pi; f_e]$. T is set to 18 according to our empirical analyses, which are presented in Section 4.

3.2 Recurrent Policy Network

The recurrent policy network is designed to cooperate with a recurrent neural network to optimize a time-sequence of attended regions for local enhancement. As illustrated in Fig. 2, the Attention-FH framework is composed of a recurrent policy network and a local enhancement network. The recurrent policy network can be formulated as a decision-making process for optimal patch selection on time intervals. At each step, the policy network takes as input the concatenation feature vector (i.e., the state), which contains the current enhanced image, the original corrupted image and the encoded historical actions, and learns to determine the optimal image patch to be enhanced at each time step. In Table 1, we illustrate the architecture of feature extractor. At the final time step T, a global delayed reward, which is measured in terms of the accumulated attended rate and the mean squared error (MSE) between the hallucinated image and the corresponding ground truth, is used to guide the training of the policy network. The agent learns to predict the most appropriate restoring route for different identities by maximizing the global delayed reward.

TABLE 1
Detailed Setting of Each Component in the Feature Extractor

	1	2	3	4	5
layer	conv	conv	conv	conv	fc
stride	2	2	2	2	-
size	64	32	16	8	64
kernel	5	3	3	3	-
channel	8	8	16	32	1

State. To provide rich contextual information, the state of the agent is designed to contain the three following types of information. 1) The enhanced face I_t from previous steps, which enables the agent to determine the patch that needs to be repaired at the next time step by fully sensing the rich contextual information (e.g., the region that is still LR and needs to be enhanced). I_t is represented as a global feature vector v_t extracted from the output of a fully connected layer. 2) The original corrupted face image I_0, which is also encoded with a global feature vector v_0, as with the enhanced facial image. 3) The encoded history action vector v_l obtained by forwarding all previous action vectors $\{h_0, h_1, \ldots, h_{t-1}\}$ into the GRU network. The output hidden variable of the GRU thus encodes all previous action information and is denoted as v_l. We formulate the state s_t with size of 512×1 to encode multiple contextual information $\{v_t, v_0, v_l\}$. In this way, the target of the agent is to determine the region cropping action (including the location and the size) of the next attended local patch by considering the state s_t.

Action. Given state s_t, the agent attempts to generate an action indication, which represents the next attended local region $\hat{I}_{t-1}^{l_t}$ (cropping from the last hallucinated image $\hat{I}_{t-1}$) for enhancement. Due to the large differences in the orientation and size of faces in in-the-wild cases, the use of a fixed-size attention bounding box to capture all facial components is not optimal. To this end, we propose a content-adaptive and size-free attention mechanism for better local region extraction. Specifically, to fully capture one facial component, the agent needs to predict the (x, y, w, h) of a rectangular region at each time step, where $x, y, w,$ and h, respectively, refers to the center coordinates and the width and height of the bounding box. Let W and H be the width and height of the target hallucinated facial image, and let $x, y = (x, y | 1 \le x \le W, 1 \le y \le H)$. To reduce the search

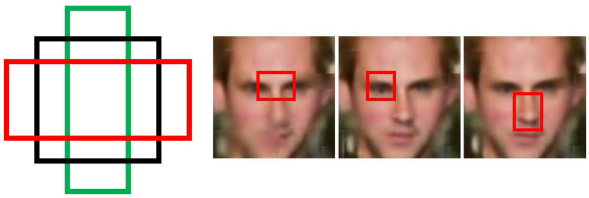


Fig. 3. A demonstration of size-free attention. As the facial part has different sizes with particular identities, our method locates the region with a flexible attention mechanism.

space, we give up the free search of the width and height and instead attempt to predict the ratio factor ID_{ratio} and the scale factor ID_{scale}. ID_{ratio} is used to control the length-width ratio. Empirical candidates for ID_{ratio} include $\{3:2, 1:1, 2:3\}$. As shown in Fig. 3, the bounding box is adapted w.r.t the ratio factor and is thus sufficiently flexible to handle various facial components. The scale factor ID_{scale} is adopted to represent the overall size of the attention box. With the sample factors, the size of the attention box can be obtained by:

$$\begin{aligned} L_h &= \mathrm{Z} * ID_{scale}^{h} \\ L_w &= L_h / ID_{ratio} * ID_{scale}^{w}, \end{aligned} \tag{5}$$

where L_h and L_w are the height and the width of the attention box, respectively. Z is a constant value that specifies the initial size of the attended region. Here, we set Z to 60. The action l_t consists of $\{x, y, ID_{ratio}, ID_{scale}, \}$. The GRU takes the state s_t as input and generates a 128×1 hidden vector v_c, which is fed to a fully connected layer to infer the action of the next step. We employ a tensor with the same size as the full image to ensure that the attended region can be accommodated.

Reward. The reward function is designed to guide the training of the recurrent policy network for the optimization of a time sequence of attended patches for local enhancement. We consider two factors when designing our reward function. 1) The MSE between the hallucinated image after T steps of local enhancement and the corresponding HR image. Specifically, let I_T be the enhanced image at the last step T, and let I_s be the image restored by the super-resolution generative adversarial network (SRGAN) [37]. We first compute their MSEs with respect to the corresponding HR ground truth, denoted as E_T and E_{SRGAN}. The first term of our reward function is defined as $\mathrm{E}_T - \mathrm{E}_{SRGAN}$. E_T measures the absolute peak signal-to-noise ratio (PSNR) of the restored image while the introduction of the second E_{SRGAN} replaces the reward function with the relative change in PSNR, which better reflects the evolution of each iteration and greatly enhances the stability of the model training. For instance, the PSNR of a restored image with simple details is usually much higher than that of an image with a relatively complex texture. By subtracting E_{SRGAN}, the reward function can better reflect the incremental situation of model training and thus apply more accurate rewards and punishments. 2) The attention rate, which is introduced to indicate whether the attended region has covered the whole image. In detail, a tensor T_0 of the same size as the full image is employed. Initially, T_0 is set with all zeros, and once a region is visited, its corresponding value is set to 1. We sum the tensor T_0 to $\mathrm{E_v}$ at the last step to reflect the attention ratio. The total reward function can be written as:

$$r_t = \begin{cases} 0 & t < T \\ \mathrm{E}_T - \mathrm{E}_{SRGAN} + \mathrm{E_V} & t = T. \end{cases} \tag{6}$$

During training, we set the reward in a global manner, i.e., the reward is assigned after step T is completed. The recurrent policy network is trained to maximize this reward via the REINFORCE algorithm [21].

Algorithm 1. Learning Algorithm of Attention-FH

Require: Training LR face images I_{lr}; HR face images I_{hr}; Initial actions h_0

while t = 1 **do**
  Represent I_{lr}, I_{lr}, h_0 as feature vectors v_0, v_0, v_l^0
  Obtain state $s_0 := \{v_t, v_0, v_l^0\}$
  Obtain actions $l_0 := f_\pi(s_0; \theta_\pi)$
  Crop out patch with respect to actions $\hat{I}_0^{l_0} := g(l_0, I_0)$
  Enhance the patch $\hat{I}_1^{l_0} := f_e(\hat{I}_0^{l_0}, I_0, I_t, I_G; \theta_e)$
  Replace $I_0^{l_0}$ with $\hat{I}_1^{l_0}$ to produce restored image I_1
  Update θ_e via $\hat{I}_0^{l_0}$ and ground truth patch $\hat{I}_{gt}^{l_0}$: $\frac{\partial L_e}{\partial f_e(\theta_e, \hat{I}_0^{l_0})}$
end while
while t $\leq$ T **do**
  Represent I_{lr}, I_{lr} as feature vectors v_t, v_0
  Forward all previous action vectors $\{h_0, h_1, \ldots, h_{t-1}\}$ and obtain history actions v_l
  Obtain state $s_t := \{v_t, v_0, v_l\}$
  Obtain actions $l_t := f_\pi(s_t; \theta_\pi)$
  Crop out patch with respect to actions $\hat{I}_t^{l_t} := g(l_t, I_t)$
  Enhance the patch $\hat{I}_{t+1}^{l_t} := f_e(\hat{I}_t^{l_t}, I_0, I_t, I_G; \theta_e)$
  Replace $I_t^{l_t}$ with $\hat{I}_{t+1}^{l_t}$ to produce restored image I_t
  Update θ_e via $\hat{I}_t^{l_t}$ and ground truth patch $\hat{I}_{gt}^{l_t}$: $\frac{\partial L_e}{\partial f_e(\theta_e, \hat{I}_t^{l_t})}$
  if t = T **then**
    Calculate reward $r_t := \mathrm{E}_T - \mathrm{E}_{SRGAN} + \mathrm{E}_V$
    Update θ_π via REINFORCE algorithm: $\frac{dr}{dl_t} := a * (r - b) * \frac{d \ln(f_\pi(x, l_t))}{dl_t}$
  end if
end while

3.3 Local Enhancement Network

Given an attended patch from the recurrent policy network, the local enhancement network f_π is employed for hallucination. To provide comprehensive contextual information, f_π takes the following three components as input: 1) The attended patch $\hat{I}_t^{l_t}$, which is represented by masking the outside area of the input image (setting the value of the region outside the selected patch to zero while keeping the pixels inside intact). 2) The current enhanced facial image I_t (with all previously hallucinated results pasted on), which provides global contextual information for the enhancement of the current patch. 3) The original corrupted image I_0, 4) and the global context I_G, which is calculated by expanding s_t to the dimension equal to the size of the input image I_0 by a fully connected layer followed by a reshaping operation (i.e., I_G is of the same shape of I_0). $\{\hat{I}_t^{l_t}, I_t, I_0, I_G\}$ are concatenated and further fed into the local enhancement network. To achieve a trade-off between performance and efficiency, we adopt a reduced version of LapSRN as our local enhancement network. The simplified settings of

TABLE 2
Detailed Setting of Each Component in the Local Enhancement Network

	1	2	3	4	5	6	7
layer	conv	conv	deconv	conv	deconv	conv	conv
stride	1	1	2	1	2	1	1
size	8	16	32	16	8	8	8
kernel	5	3	3	3	3	3	5
channel	64	32	32	32	8	8	1

LapSRN are listed in Table 2. We resize the three input components to the same size as the LR corrupted image without interpolation to improve the efficiency of our model. The local enhancement network is fully convolutional and is composed of 5 convolutional and 2 deconvolutional layers. The convolutional layers all have a stride of 1, the kernels of the head and tail layers are of size $5 * 5$ and the kernels of the other layers are of size $3 * 3$. By incorporating two deconvolutional layers with stride 2, the network is able to learn to reconstruct the resolution of the corrupted patch in a cascaded manner. At the end of the local enhancement network, the LR input image is upscaled to the target resolution. We crop out the attended region in the hallucinated result w.r.t its size and location. The cropped HR patch $\hat{I}_{t+1}^{l_t}$ is added to the accumulated enhanced result I_t. Finally, the residual between the attention patch $\hat{I}_t^{l_t}$ and the ground truth face image I_{hr} can be estimated by the well-trained local enhancement network.

3.4 Model Training

We illustrate the training strategy of our Attention-FH in Fig. 4. The recurrent policy network, which learns to obtain the attentional patch by maximizing the reinforced reward, is shown in Fig. 2 (1). Fig. 2 (2) shows the local enhancement network, which learns to enhance the attended patch in an end-to-end mode by minimizing the MSE.

In the training phase, the recurrent policy network is optimized by the REINFORCE algorithm [21], guided by the reward calculated at the end of sequential enhancement when the maximum time step T is reached. The local enhancement network is optimized to minimize the L_2 distance between the restored patch and its corresponding HR ground truth. The supervised loss is calculated at each time step and can be minimized based on backpropagation. After calculating the last step of attending patches for local enhancement, we obtain the global reward, which is leveraged to optimize the policy network.

The whole training algorithm of our Attention-FH is illustrated in Algorithm 1, which accords with the pipeline of our proposed framework shown in Fig. 1.

4 Experiments

To demonstrate the advantages of our Attention-FH, we have conducted extensive experiments on multiple widely used benchmarks, i.e., *CelebFaces Attributes Dataset* [1], *Public Figures Face Dataset* [38], *Labeled Faces in the Wild Dataset* [39], *Surveillance Cameras Face Dataset* [40] and *BioID Face Dataset* [41]. We first briefly introduce the evaluation datasets, the corresponding evaluation protocols and the implementation details. Then, we perform comprehensive comparisons to verify the superiority of our Attention-FH over all the compared state-of-the-art approaches. Note that because the degradation types of face hallucination are complex, we have employed several down-sampling factors to evaluate our Attention-FH under various challenging conditions. Finally, we have performed detailed ablation studies to demonstrate the contribution of each component within our Attention-FH.

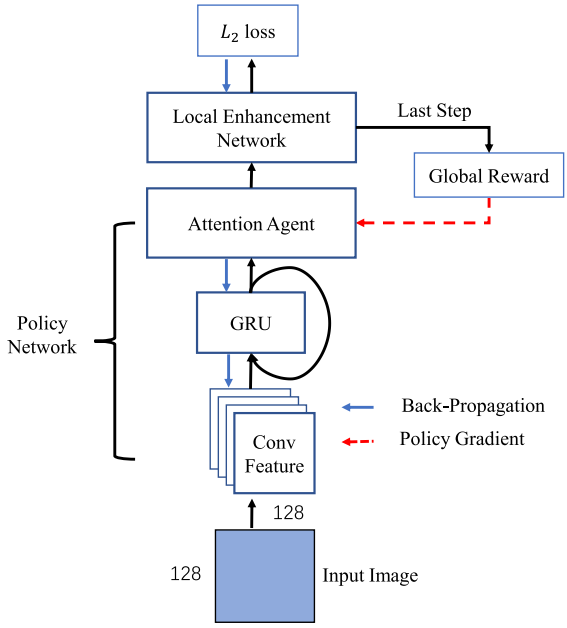


Fig. 4. Detailed training procedure of our Attention-FH. The local enhancement network is optimized via the L_2 loss between the restored patch and the corresponding HR ground truth. After performing the last step, we calculate the global reward and pass the policy gradient to optimize the recurrent policy network.

4.1 Datasets and Evaluation Protocols

We employ the following seven public datasets under various domains for a comprehensive evaluation to validate the robustness of our Attention-FH to in-the-wild faces.

- *CelebA* [1] is a large-scale dataset that contains 202,599 in-the-wild face images with 10,177 identities. Following the settings in [1], we adopt 188,311 images for training and use the remaining 14,288 for testing.
- *SCface* [40] consists of 4,160 images with 130 identities collected under an uncontrolled environment using five video surveillance cameras in various situations. As video surveillance is one of the main application scenarios for face hallucination, this dataset can be used to evaluate the compared methods from a practical perspective. We utilize 2,405 images for training and the rest for testing.
- *BioID* [41] is a public dataset with 1,521 gray face images, all taken in a laboratory from a frontal view. We use 1,028 images for training and the remaining 493 images for testing.
- *PubFig* [38] is a large dataset with 58,797 real-world face images collected from the Web. In our

TABLE 3
Quantitative Comparison of Our Model and the Competing Methods in Terms of the PSNR Index

Dataset	Scale	Bicubic	General SR		Face Hallucination		GAN Based	Our
			SRCNN	VDSR	BiCNN	GLN	SRGAN	
PubFig	×4	24.76	25.21	28.05	24.35	26.12	27.44	**28.87**
	×8	20.75	21.31	21.94	20.69	21.33	23.45	**23.59**
	×16	17.89	18.48	19.28	18.13	18.73	20.25	**20.31**
CelebA	×4	25.76	26.01	28.38	24.93	29.92	28.17	**30.58**
	×8	21.84	22.64	24.46	21.32	25.48	24.37	**26.14**
	×16	18.78	19.80	20.07	19.01	21.20	21.99	**22.63**
SCface	×4	26.15	26.54	31.59	26.04	29.84	30.10	**34.01**
	×8	20.83	21.68	24.12	20.67	24.10	24.72	**26.04**
	×16	17.15	18.45	20.32	17.21	18.28	20.63	**22.19**
BioID	×4	24.59	25.71	29.38	23.16	26.71	28.16	**33.38**
	×8	20.24	21.85	23.95	20.14	22.03	23.23	**27.81**
	×16	17.15	18.45	19.41	17.11	19.85	21.59	**23.48**
LFW	×4	26.79	28.94	32.11	26.60	30.34	31.41	**32.93**
	×8	21.92	23.92	24.12	22.62	24.51	25.49	**27.81**
	×16	19.95	21.34	22.40	20.82	22.44	23.01	**23.13**

*We use **bold face** and underline to indicate the first and second place in each dataset.*

TABLE 4
Quantitative Comparison of Our Model and the Competing Methods in Terms of the SSIM Index

Dataset	Scale	Bicubic	General SR		Face Hallucination		GAN Based	Our
			SRCNN	VDSR	BiCNN	GLN	SRGAN	
PubFig	×4	0.7600	0.7708	0.8262	0.7167	0.7793	0.8197	**0.8587**
	×8	0.5782	0.5819	0.5822	0.5651	0.5375	0.6747	**0.6805**
	×16	0.4581	0.4584	0.4761	0.4660	0.4203	0.5384	**0.5394**
CelebA	×4	0.7672	0.7717	0.8115	0.7663	0.8587	0.8214	**0.8711**
	×8	0.6107	0.6228	0.6747	0.6093	0.7141	0.6749	**0.7441**
	×16	0.5016	0.4995	0.5042	0.5019	0.5307	0.5946	**0.6316**
SCface	×4	0.8367	0.8435	0.9036	0.8396	0.8848	0.8907	**0.9356**
	×8	0.6303	0.6493	0.7009	0.6291	0.7132	0.7567	**0.7983**
	×16	0.4821	0.5018	0.5355	0.4823	0.4962	0.6175	**0.6835**
BioID	×4	0.8056	0.8244	0.8618	0.6719	0.8333	0.8600	**0.9418**
	×8	0.6189	0.6196	0.6676	0.5657	0.5996	0.6806	**0.8568**
	×16	0.4821	0.5018	0.4908	0.4572	0.5305	0.6546	**0.7376**
LFW	×4	0.8469	0.8686	0.8917	0.8329	0.8922	0.9247	**0.9418**
	×8	0.6712	0.6927	0.7031	0.6801	0.7109	0.7314	**0.8568**
	×16	0.5617	0.5742	0.5879	0.5838	0.6132	0.6449	**0.6542**

*We use **bold face** and underline to indicate the first and second place in each dataset.*

experiment, 11,041 images are utilized for training, and the remaining 6,425 images are used for testing.

- *LFW* [39] contains 13,233 in-the-wild face images with 5,749 identities. We split the dataset w.r.t the partitions mentioned in [39].
- *Multi-PIE* [42] contains images of 337 subjects captured from complicated perspectives and under complex illumination conditions in four different sessions. We use 126,093 images for training and 31,524 images for testing.
- *Extended Yale-B* [43] is a large face dataset that contains 16,128 images of 28 identities under 9 poses and 64 illumination conditions. We randomly choose 12,908 images for training and 3,220 images for evaluation.

In terms of the evaluation metrics, we exploit PSNR and structural similarity (SSIM) to measure the performance of the compared methods. For a comprehensive evaluation, the feature similarity is also compared by adopting FSIM [44].

4.2 Implementation Details

All the datasets are first aligned with two points of the eye regions by CFSS [45]. Then, we simply crop the images to a size of 160 × 120 by prefetching the centric region, except for the LFW dataset, the images of which are cropped to 128 × 128. To ensure a fair comparison,

TABLE 5
Quantitative Comparison of Our Model and the Competing Methods in Terms of the FSIM Index

Dataset	Scale	Bicubic	General SR		Face Hallucination		GAN Based	Our
			SRCNN	VDSR	BiCNN	GLN	SRGAN	
PubFig	×4	0.8454	0.8581	0.8942	0.8748	0.8748	0.8892	**0.9099**
	×8	0.7337	0.7604	0.7587	0.7669	0.7669	0.8081	**0.8121**
	×16	0.6360	0.7872	0.7207	0.7151	0.7151	0.7297	**0.7346**
CelebA	×4	0.8565	0.8634	0.8890	0.9148	0.9148	0.8956	**0.9211**
	×8	0.7514	0.7762	0.8077	0.8290	0.8290	0.8066	**0.8410**
	×16	0.6511	0.7098	0.6980	0.7331	0.7331	0.7552	**0.7628**
SCface	×4	0.8813	0.8906	0.9321	0.9185	0.9185	0.9205	**0.9513**
	×8	0.7518	0.7893	0.8195	0.8318	0.8318	0.8437	**0.8675**
	×16	0.6487	0.7055	0.7236	0.6955	0.6955	0.7586	**0.7977**
BioID	×4	0.8538	0.8767	0.9066	0.8926	0.8926	0.9012	**0.9571**
	×8	0.7405	0.7800	0.7977	0.7803	0.7803	0.8162	**0.9072**
	×16	0.6487	0.7055	0.7104	0.7379	0.7379	0.7849	**0.8336**
LFW	×4	0.8947	0.9069	0.9300	0.8982	0.9151	0.9384	**0.9571**
	×8	0.7824	0.8314	0.8391	0.7903	0.8405	0.8278	**0.9072**
	×16	0.6771	0.7454	0.7595	0.7243	0.7788	0.7722	**0.7824**

*We use **bold face** and underline to indicate the first and second place in each dataset.*

all the methods are trained on only the corresponding training set, without using the other datasets for pre-training. We evaluate our method with scaling factors of ×4, ×8 and ×16 to model different types of situations. In addition, we also normalize the input images into $[-1, 1]$. The recurrent time step T of the policy network is set to 18 to achieve a trade-off between efficiency and accuracy. The setting of the recurrent time step T is also investigated in Section 4.5. Our Attention-FH is trained using ADAM gradient descent [46] with a base learning rate of 3×10^{-4}, a weight decay of $1e^{-7}$, and a momentum term of 0.5. The training batch size is 16. Considering the absolute free attention region can lead to unstable performance, we impose some empirical constraints on the size-free attention mechanism, i.e., the length and width of the attention region are customized with respect to the ScaleID and RatioID, which are evaluated in Section 4.5.

4.3 Competing Methods

We compare our method with several state-of-the-art methods, including SRCNN [15], VDSR [24], SFH [10], BiCNN [13], GLN [25], and SRGAN [37]. These methods can be categorized into three groups: (i) general image super-resolution: SRCNN and VDSR; (ii) face hallucination: SFH, BiCNN and GLN; and (iii) generative adversarial learning: SRGAN. The first and second types are commonly applied to address regular image and face image restoration, respectively, while the third one is widely used in image generation and has achieved impressive results.

4.4 Quantitative and Qualitative Comparisons

As illustrated in Tables 3, 4, 5, and 6, our Attention-FH consistently outperforms all the compared state-of-the-art methods, with clear margins in terms of all evaluation metrics. Attention-FH outperforms the best of the competing methods with 2.42 dB, 1.32 dB, and 1.56 dB on the SCface dataset with respect to the PSNR index, respectively. Moreover, our Attention-FH surpasses all the competing methods by large margins on all datasets when the scaling factor is small (e.g., 4). These results confirm the significant superiority of our Attention-FH. Note that we do not present the quantitative results of SFH [10], which relies heavily on face alignment and thus may fail to handle some testing images.

The visual comparison on the SCface dataset is presented in Fig. 5. Since the SCface dataset is similar to a real-world scenario, the dataset can be employed to explicitly validate the performance of each hallucination approach in terms of practicability. As shown in Fig. 5, regular super-resolution methods produce hallucinated facial images with blurry predictions. By contrast, our Attention-FH can generate faces with well-maintained facial structure. This result demonstrates that our Attention-FH is capable of deblurring and anti-aliasing facial images to preserve the structural information.

The qualitative results shown in Figs. 6, 7 and 8 demonstrate that our Attention-FH achieves significant improvements in restoration quality compared with all the competing methods. In addition, the attention mechanism also benefits our Attention-FH when addressing variation in pose, illumination and facial appearance. As depicted in Fig. 6, the facial expression of the woman in the third row is a 'smile' with her mouth opened, and the latter woman has a 'smile' with her mouth closed. Our Attention-FH outperforms all the competing methods in successfully addressing these two case with corrupted inputs. Furthermore, our Attention-FH can even hallucinate the man in the eighth row with 'glasses', which is extremely challenging for all the compared state-of-the-art approaches. As demonstrated in Fig. 8, our Attention-FH can recover naturally acceptable facial images even after substantial information has been lost by downsampling.

To further verify the effectiveness of our model, we also compare our proposed Attention-FH with some methods [47], [48], [49] proposed after the conference version of

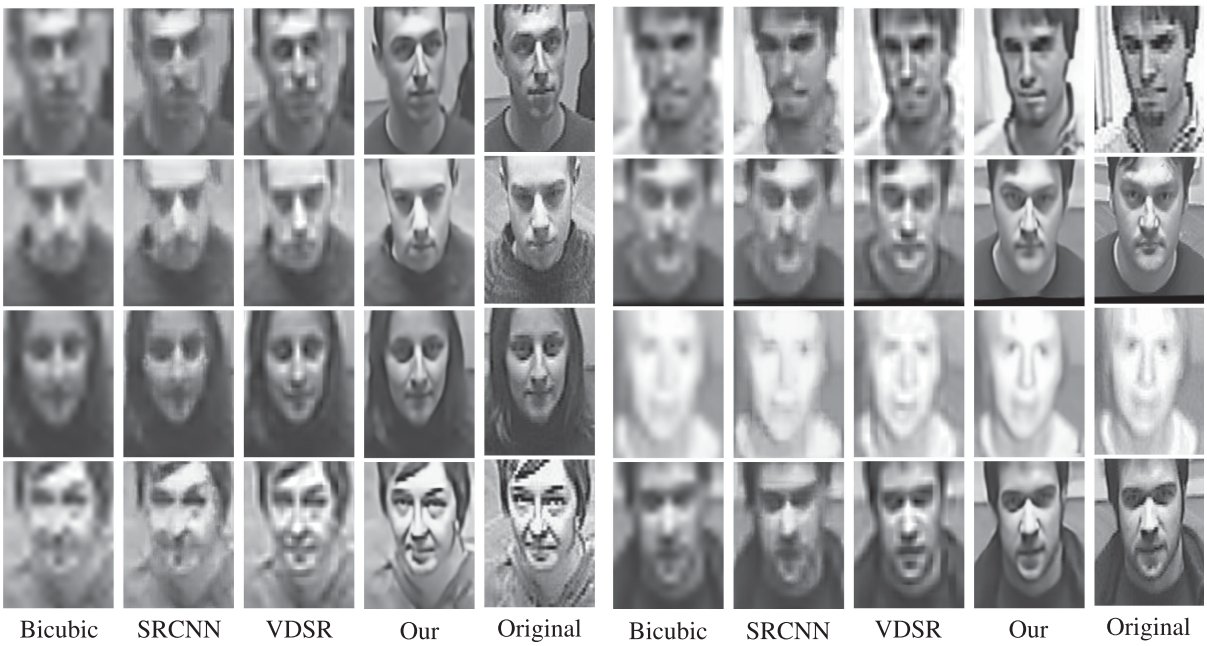


Fig. 5. Qualitative comparisons on the SCface [40] dataset with a scaling factor of 8. The testing faces are captured by surveillance cameras, similarly to practical scenarios. This figure is best viewed by zooming in on the electronic version.

this paper. Pixel-SR [48] and Image Transformer [49] exhibit good performance in general object hallucination, and enhanced deep super-resolution network (EDSR) is famous for generating clear structures in general image SR. Since Pixel-SR and Image Transformer aim to hallucinate small-size objects (e.g., 32×32 pixels), the target resolution of face hallucination (e.g., 128×128 pixels) may lead to GPU memory explosion. Hence, following the official implementation, we conduct patch-based learning and combine the restored patches into a full image for evaluation. As shown in Table 7, our model produces better results than the other methods, except for EDSR. Nevertheless, Attention-FH is sufficiently flexible to incorporate EDSR as the local enhancement network to improve performance. We reduce the number of recurrent steps to 4 and implement EDSR for local enhancement. "Our-EDSR" achieves superior performance to EDSR as we expected, which well illustrates the flexibility and effectiveness of the proposed model. Furthermore, we average the results of three model snapshots with different iterations (after convergence) of "Our-EDSR" and "EDSR" for comparison. As illustrated in Table 7, "Our-EDSR ensemble" achieves superior performance.

Additionally, we compare Attention-FH on general image super-resolution with state-of-the-art image SR approaches (i.e., VDSR, LapSRN, MemNet [50] and IDN [51]). We follow the training scheme of IDN [51] and conduct experiments on Set14 [52] to demonstrate the performance of our method on arbitrary domain images. With the overall parameter number fixed, we adaptively decrease the recurrent step and build up the local enhancement network for better image SR. As shown in Table 8, our method consistently outperforms all the compared methods on the $\times 2$, $\times 3$ and $\times 4$ settings. Though Attention-FH is indeed capable of restoring the general image well, our model is still a face hallucination framework. By incorporating recurrent attention mechanism, Attention-FH is specialized in low-resolution faces and achieves much greater advantages in the task of face hallucination.

4.5 Ablation Study on the Policy Network

To demonstrate the effectiveness of the policy network, we compare several variants of our Attention-FH as baseline methods, i.e., "CNN-16", "Our w/o attention", "Our w/ random", "Our w/ sequences", "Our w/o size-free" and "Our w/ I_0 agent". "CNN-16" indicates a plain convolution network with 16 layers. "Our w/o attention" refers to that the whole face image is recursively enhanced via a recurrent model from a holistic perspective instead of attending patches. "Our w/ random" denotes that we randomly select the attention region to perform random patch-based enhancement. "Our w/ sequences" scans through the whole image in sequence. "Our w/o size-free" uses a fixed attention box inside the policy network. "Our w/ I_0 agent" replaces I_t with I_0 as the input for the policy network. Moreover, we also consider the complicated transformation mechanism that excludes RL, e.g., Attention-FH with spatial transform net (STN) [53] (denoted as "STN"), which is capable of learning to select informative patches for face hallucination by estimating the subgradients w.r.t the location of the captured patch.

To demonstrate the superiority of the policy network, we conduct a comparison with the STN [53]. Specifically, our Attention-FH and the STN both employ a paramount patch mining strategy. STN chooses the patch by learning auxiliary terms by minimizing the MSE, and our Attention-FH employs RL to iteratively identify the correct attention region. Therefore, we conduct STN [53] for comparison to illustrate the strengths of RL. For STN, the outputs of the policy network are replaced with a vector $\{x, y\}$, which indicates the corresponding coordinates to extract the patch. We fix the transform scale to

Fig. 6. Qualitative comparison on the PubFig [38] dataset with a scaling factor of 8. The image is best viewed by zooming in on the electronic version.

ensure an attended patch size of 60×45. As shown in Table 9, our full model outperforms STN by a clear margin. This result confirms that RL is beneficial in crucial region mining for face hallucination.

4.5.1 *Effectiveness of Patch-Wise Enhancement*

As shown in Table 9, our full model surpasses "Our w/o attention" by 0.73 dB and 0.28 dB on the LFW dataset with the scaling factors of 4 and 8, respectively. This justifies that

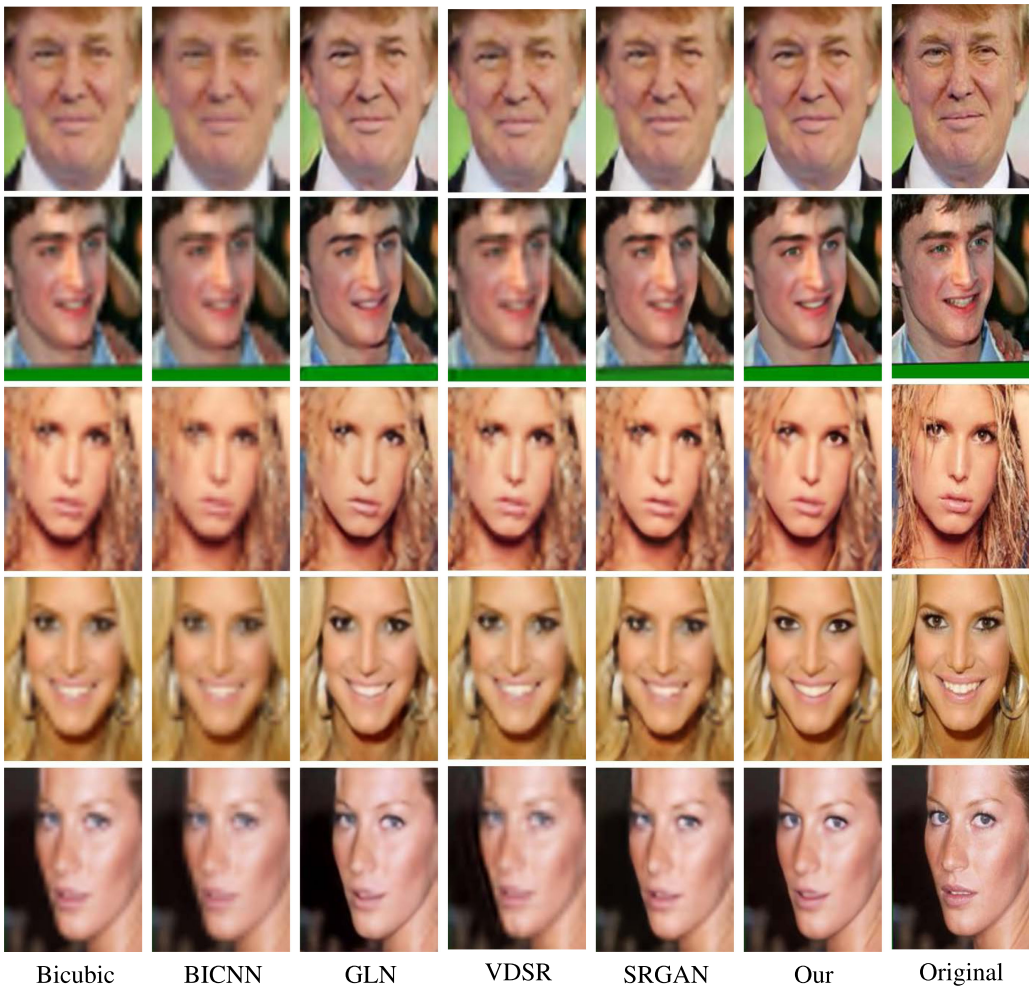


Fig. 7. Qualitative comparison on the PubFig [38] dataset with a scaling factor of 4. The image is best viewed by zooming in on the electronic version.

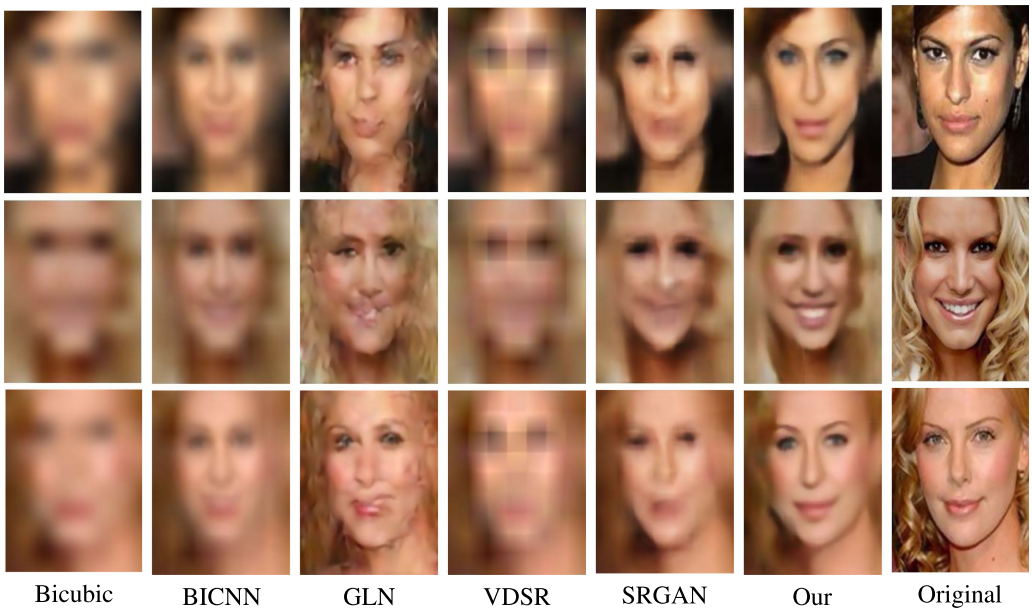


Fig. 8. Qualitative comparison on the PubFig [38] dataset with a scaling factor of 16. The image is best viewed by zooming in on the electronic version.

TABLE 6
Comparison on Multi-PIE [42] and Extended Yale-B [43]

Algorithm	Multi-PIE ×8	Yale-B ×8
Bicubic	21.11	24.67
SRCNN	24.53	25.69
VDSR	26.12	26.57
GLN	28.17	25.98
SRGAN	28.26	27.32
Our	27.81	28.89

TABLE 8
Comparison on General Image Super-Resolution

Algorithm	×2	×3	×4
Bicubic	30.24	27.55	26.00
VDSR	33.03	29.77	28.01
LapSRN	32.99	29.79	28.09
MemNet [50]	33.28	30.00	28.26
IDN [51]	33.30	29.99	28.31
Our	33.34	33.07	28.33

TABLE 7
Comparison of Our Proposed Model with Deeper Inferences

Algorithm	LFW ×8	LFW ×16
Pixel-SR	21.11	20.47
Image Transformer	24.53	22.22
DRCN	26.43	22.39
EDSR	28.17	24.10
EDSR ensemble	28.26	24.21
Our	27.81	23.13
Our-EDSR	28.25	24.23
Our-EDSR ensemble	28.30	24.28

TABLE 9
Comparison of Our Proposed Model Under Different Settings

Algorithm	LFW ×4	LFW ×8
CNN-16	29.11	24.02
Our w/o attention	32.29	25.69
Our w/ random	31.63	25.74
Our w/ sequences	31.78	25.98
Our w/o size-free	32.89	26.12
Our w/ I_0 agent	32.12	25.97
STN [53]	28.13	25.75
Our	33.02	26.24

in-the-wild faces are too changeable to restore, however individual facial parts are relatively stable and can be exploited for partial enhancement. Besides, "Our w/ random" obtains a significant improvement over "CNN-16", which indicates that the multiple recurrent enhancements itself help to improve the image recovery. Furthermore, our full model achieves consistently higher PSNR values than "Our w/ random" due to the crucial patch sequence optimization implemented via RL.

4.5.2 Effectiveness of Sequentially Attending Patches

We demonstrate the contribution of sequentially attending patches from the perspective of the attention agent. As illustrated in Table 9, we first input the face image without enhancements into the policy network, e.g., "Our w/ I_0 agent", and observe that the performance (PSNR) degrades on the LFW dataset with scaling factors of 4 and 8. By contrast, our full model produces better scores. This result proves that the restored patch not only improves the final face image but also leads the agent to select more accurate region sequence for restoration.

4.5.3 Effectiveness of Increasing Recursion Depth

To demonstrate the sensitivity of our Attention-FH in terms of the recursive step T, we explore the effect of different recursive steps T for sequentially enhancing facial parts. Specifically, we conduct the experiment under five different settings ($T = 5, 15, 18, 25, 35$) of recursive step on the LFW dataset with scaling factors of 4 and 8. Fig. 9 shows that the face hallucination performance gradually increases with increasing number of attention steps. The PSNR measure improves dramatically when the number of recursion steps is small, as the extracted patches are still not enough to cover the whole image. When the number of recursion steps reaches more than 15, the extracted patches are generally enough to cover the whole image. Beyond 15 steps, the stepwise performance improvement in PSNR becomes negligible. This phenomenon becomes more obvious as the number of steps approaches 25. Owing to the size-free attention strategy and stabilized reward function, we can obtain considerable restored quality under the PSNR metric when T is set to 18. In our experiment, we empirically set $T = 18$ considering the acceptable computational costs under practical scenarios.

4.5.4 Effectiveness of the Stabilized Reward Function

We conduct an experiment to validate the contribution of the proposed reward function. Compared with former reward function, our stabilized reward function incorporates PSNR gain value instead of absolute PSNR value as the reward. Given the renewal reward, Attention-FH demonstrates an accurate attention agent towards face halluci-

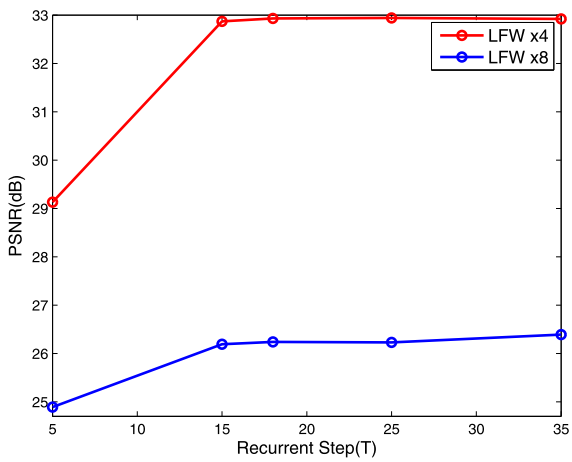


Fig. 9. PSNR comparison on the variants of our Attention-FH using different numbers of steps for sequentially enhancing facial parts on the LFW dataset. $T = 18$ achieves a balance between accuracy and efficiency.

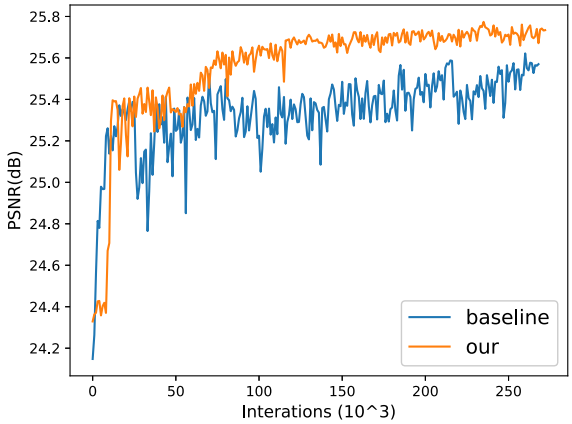


Fig. 10. We enhance the reward function by improving its stability. We redefine the previous reward function from our preliminary conference version by means of a more accurate baseline and achieve superior performance.

nation. As shown in Fig. 10, our reward function achieves more stable performance with lower variance. Furthermore, compared with the reward function from our preliminary conference version, this new reward function leads to higher and more stable PSNR results for our Attention-FH.

4.5.5 *Effectiveness of the Size-Free Attention Mechanism*

As depicted in Fig. 3, our facial component has a different size for each identity. We improve our policy network by implementing a flexible attention mechanism to attend the facial part accurately. We conduct an ablation study by adopting a fixed attention box in the policy network, named "Our w/o size-free", to validate the effectiveness. This model has the same settings as our full model, except for using a 60×60 attention box. Table 9 shows that although "Our w/o size-free" produces favorable results, our Attention-FH achieves 0.13 dB and 0.12 dB improvements with scaling factors of 4 and 8, respectively. These results confirm the contribution of the proposed size-free attention mechanism.

4.6 Ablation Study on the Local Enhancement Network

Since the pipeline of our Attention-FH is flexible and extensional, we investigate the use of different network architectures as the local enhancement network. To make the investigation comprehensive, we consider several methods, namely, VDSR [24], Sub-pixel [54], GLN [25], LapSRN [55], U-Net [56] and FSRCNN [57]. For U-Net [56], we reduce the parameters to avoid the case that the model is too large to train. As shown in Table 10, LapSRN [55] achieves the best results while U-Net [56] is the most efficient method for generating hallucinated faces. However, neither method exhibits distinct differences in terms of PSNR under the recurrent attention mechanism. We choose FSRCNN [57] as the implementation of our local enhancement network based on the trade-off between efficiency and accuracy.

TABLE 10
Experimental Study of the Trade-off between Efficiency and Accuracy on Different Local Enhancement Network Architectures

Algorithm	PSNR	Time (millisecond)
VDSR [24]	23.56	0.312
Sub-pixel [54]	23.59	0.066
GLN [25]	23.63	0.092
LapSRN [55]	23.64	0.081
U-Net [56]	23.60	0.057
FSRCNN [57]	23.64	0.075

This analysis is conducted on images of size 120×160.

TABLE 11
Efficiency Comparison on the PubFig Dataset with a Scaling Factor of 8

Method	Parameter Number	PSNR	Time (frame/second)
VDSR [24]	664,704	21.94	0.025
EDSR [47]	2,463,217	24.21	0.045
Our	1,706,850	23.62	0.092

4.7 Efficiency Analysis

We have also conducted an experimental comparison to verify the efficiency of Attention-FH. The results in Table 11 demonstrate that our Attention-FH requires very little time cost to achieve the superior performance. Compared with EDSR, our model achieves remarkable parameter

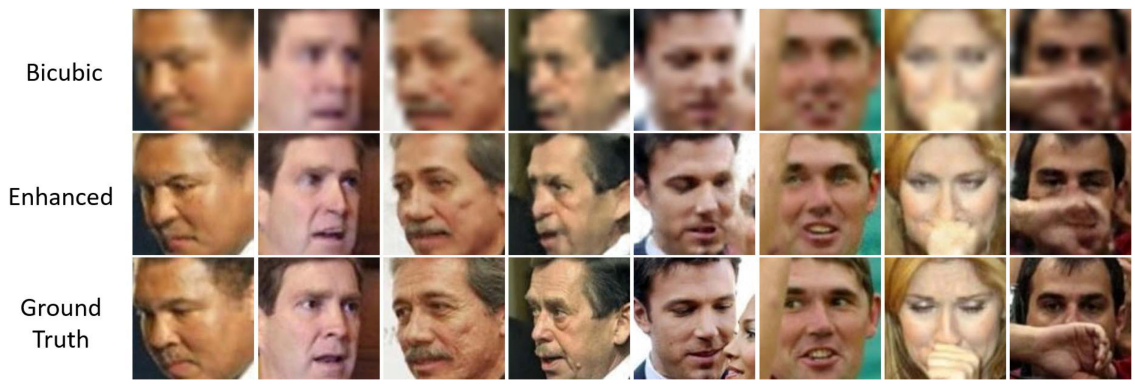


Fig. 11. Visual results on profile \ occlusion samples. We conduct this comparison on LFW [39] with $\times 8$ factor. Best viewed by zooming in the electronic version.

advantages. As Attention-FH is less efficient than VDSR, the proposed model achieves significant improvement over restoration quality. With a single GPU, Attention-FH is able to perform real-time efficiency. However, Attention-FH still meets an efficiency bottleneck when it is deployed on the mobile platform.

4.8 Limitations

In this section, we discuss the limitation of Attention-FH. With the novel attention mechanism, our model is capable of hallucinating profile \ occlusion faces well. However, Attention-FH may hallucinate the incorrect facial part if the occluded content is close to facial component. As shown in the last row of Fig. 11, Attention-FH hallucinate a mouth in the occlusion area, which occurred by hand. This failure case illustrates the limitation that Attention-FH meets an upper bound on complex occlusion samples. We will further improve Attention-FH by enhancing the generalization ability towards complex occlusion cases.

5 Conclusion

In this paper, we have proposed a deep RL-based attention mechanism to address the problem of face hallucination. In contrast to traditional patch-wise face hallucination models that usually neglect the interdependency between facial components, our framework implements a deep RL model and jointly optimizes a recurrent policy network, which learns to determine an ordered patch hallucination sequence, and a local enhancement network for facial part super-resolution. Our Attention-FH fully reflects the human visual perception mechanism and is capable of adaptively inferring an optimal search path for each facial image according to its unique appearance features. Extensive experiments show that Attention-FH outperforms state-of-the-art face hallucination methods and achieves leading performance on both widely used evaluation protocols and visual quality comparisons.

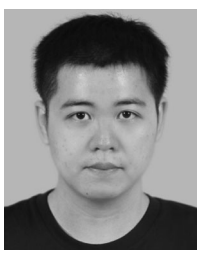
Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant No.2018YFC0830103 and No.2016YFB1001004, in part by the NSFC-Shenzhen Robotics Projects (U1613211), in part by the National Natural Science Foundation of China under Grant No.61702565, in part by National High Level Talents Special Support Plan (Ten Thousand Talents Program) and in part by the Fundamental Research Funds for the Central Universities under Grant No.18lgpy63.

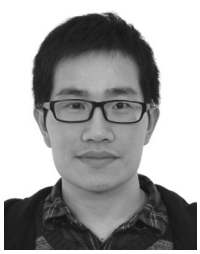
References

[1] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3730–3738.
[2] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 918–930, May 2016.
[3] L. Liu, G. Li, Y. Xie, Y. Yu, Q. Wang, and L. Lin, "Facial landmark machines: A backbone-branches architecture with progressive representation learning," *IEEE Trans. Multimedia*, 2019.
[4] E. Zhou, Z. Cao, and Q. Yin, "Naive-deep face recognition: Touching the limit of LFW benchmark or not?" *arXiv preprint arXiv:1501.04690*, 2015.
[5] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 105–114.
[6] 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 Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2492–2501.
[7] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 614–630.
[8] Y. Song, J. Zhang, S. He, L. Bao, and Q. Yang, "Learning to hallucinate face images via component generation and enhancement," in *Proc. Twenty-Sixth Int. Joint Conf. Artif. Intell.*, 2017, pp. 4537–4543.
[9] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.
[10] C.-Y. Yang, S. Liu, and M.-H. Yang, "Structured face hallucination," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1099–1106.
[11] X. Ma, J. Zhang, and C. Qi, "Hallucinating face by position-patch," *Pattern Recognit.*, vol. 43, no. 6, pp. 2224–2236, 2010.
[12] M. F. Tappen and C. Liu, "A bayesian approach to alignment-based image hallucination," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 236–249.
[13] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Learning face hallucination in the wild," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 3871–3877.
[14] C. Liu, H.-Y. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *Int. J. Comput. Vis.*, vol. 75, no. 1, pp. 115–134, 2007.
[15] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.
[16] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015.
[17] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 464–472.
[18] T. Chen, Z. Wang, G. Li, and L. Lin, "Recurrent attentional reinforcement learning for multi-label image recognition," *Thirty-Second AAAI Conf. Artif. Intell.*, 2018.
[19] G. Li, Y. Gan, H. Wu, N. Xiao, and L. Lin, "Cross-modal attentional context learning for RGB-D object detection," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1591–1601, Apr. 2019.
[20] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, pp. 484–503, 2016.
[21] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Mach. Learn.*, vol. 8, no. 3, pp. 229–256, 1992.
[22] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li, "Attention-aware face hallucination via deep reinforcement learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 690–698.
[23] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)*, vol. 35, no. 3, pp. 425–434, Aug. 2005.
[24] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646–1654.
[25] O. Tuzel, Y. Taguchi, and J. R. Hershey, "Global-local face upsampling network," *arXiv preprint arXiv:1603.07235*, 2016.
[26] A. Shocher, N. Cohen, and M. Irani, ""Zero-shot" super-resolution using deep internal learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3118–3126.
[27] S. Reed, A. v. d. Oord, N. Kalchbrenner, S. G. Colmenarejo, Z. Wang, D. Belov, and N. de Freitas, "Parallel multiscale autoregressive density estimation," in *Proc. 34th Int. Conf. Mach. Learn. Vol. 70*, 2017, pp. 2912–2921.
[28] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9446–9454.
[29] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 318–333.

[30] Z. Jie, X. Liang, J. Feng, X. Jin, W. Lu, and S. Yan, "Tree-structured reinforcement learning for sequential object localization," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 127–135.
[31] V. Mnih, N. Heess, A. Graves, and k. kavukcuoglu, "Recurrent models of visual attention," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.
[32] L. Xiaodan, L. Lisa, and P. X. Eric, "Deep variation-structured reinforcement learning for visual relationship and attribute detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4408–4417.
[33] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
[34] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 2397–2406.
[35] B. Goodrich and I. Arel, "Reinforcement learning based visual attention with application to face detection," in *Proc. IEEE Comput. Society Conf. Comput. Vis. Pattern Recognit. Workshops*, 2012, pp. 19–24.
[36] J. C. Caicedo and S. Lazebnik, "Active object localization with deep reinforcement learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2488–2496.
[37] C. Ledig, Z. Wang, W. Shi, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, and A. Tejani, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 105–114, 2016.
[38] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Proc. IEEE 12th Int. Conf. Comput. Vision*, 2009, pp. 365–372.
[39] G. B. Huang, V. Jain, and E. Learned-Miller, "Unsupervised joint alignment of complex images," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[40] M. Grgic, K. Delac, and S. Grgic, "Scface–surveillance cameras face database," *Multimedia Tools Appl.*, vol. 51, no. 3, pp. 863–879, 2011.
[41] O. Jesorsky, K. J. Kirchberg, and R. Frischholz, "Robust face detection using the hausdorff distance," in *Proc. Int. Conf. Audio- Video-Based Biometric Person Authentication*, 2001, pp. 90–95.
[42] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "The CMU multi-pose, illumination, and expression (multi-pie) face database," Robot. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. TR-07–08, 2007.
[43] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.
[44] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.
[45] S. Zhu, C. Li, C. L. Chen, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4998–5006.
[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2015.
[47] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1132–1140.
[48] R. Dahl, M. Norouzi, and J. Shlens, "Pixel recursive super resolution," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5439–5448.
[49] N. Parmar, A. Vaswani, J. Uszkoreit, Ł. Kaiser, N. Shazeer, A. Ku, and T. Dustin, "Image transformer," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4052–4061.
[50] Y. Tai, J. Yang, X. Liu, and C. Xu, "Memnet: A persistent memory network for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4549–4557.
[51] Z. Hui, X. Wang, and X. Gao, "Fast and accurate single image super-resolution via information distillation network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 723–731.
[52] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surfaces*, 2010, pp. 711–730.
[53] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.
[54] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.
[55] W. S. Lai, J. B. Huang, N. Ahuja, and M. H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5835–5843.
[56] O. Ronneberger, P. Fischer, and T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Berlin, Germany: Springer, 2015.
[57] C. Dong, C. L. Chen, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 391–407.

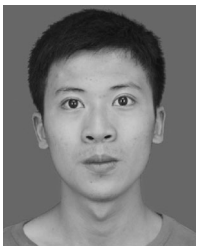


Yukai Shi received the BE degree from Heilongjiang University, Harbin, China, in 2014, and is currently working towards the PhD degree in the School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China. His research interests includes computer vision and machine learning.



Guanbin Li (M'15) received the PhD degree from the University of Hong Kong, in 2016. He is currently an associate professor with the School of Data and Computer Science, Sun Yat-sen University. He was a recipient of Hong Kong PhD Fellowship. His current research interests include computer vision, image processing, and deep learning. He has authorized and co-authorized on more than 30 papers in top-tier academic journals and conferences. He serves as an area chair for the conference of VISAPP. He has been serving as a reviewer for numerous academic journals and conferences such as the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, the *IEEE Transactions on Image Processing*, the *IEEE Transactions on Multimedia*, the *IEEE Transactions on Computers*, CVPR, AAAI and IJCAI. He is a member of the IEEE.

Qingxing Cao received the BS degree in software engineering from Sun Yat-Sen University, Guangzhou, China, in 2013. He is currently working toward the PhD degree in computer science and technology at Sun Yat-Sen University, advised by Professor Liang Lin. His current research interests include deep learning and computer vision.



Keze Wang received the BS degree in software engineering from Sun Yat-Sen University, Guangzhou, China, in 2012. He is currently working toward the dual PhD degrees at Sun Yat-Sen University and Hong Kong Polytechnic University, advised by Prof. Liang Lin and Lei Zhang. His current research interests include computer vision and machine learning. More information can be found on his personal website http://kezewang.com.

Liang Lin (M'09, SM'15) is a full professor of Sun Yat-sen University. He is the Excellent Young Scientist of the National Natural Science Foundation of China. From 2008 to 2010, he was a post-doctoral fellow at the University of California, Los Angeles. From 2014 to 2015, as a senior visiting scholar, he was with the Hong Kong Polytechnic University and the Chinese University of Hong Kong. He currently leads the SenseTime R&D teams to develop cutting-edge and deliverable solutions on computer vision, data analysis and mining, and intelligent robotic systems. He has authored and co-authored more than 100 papers in top-tier academic journals and conferences. He has been serving as an associate editor of IEEE Trans. Human-Machine Systems, The Visual Computer and Neurocomputing. He served as Area/Session chair for numerous conferences, including ICME, ACCV, and ICMR. He was the recipient of the Best Paper Runners-Up Award at ACM NPAR 2010, a Google Faculty Award in 2012, the Best Paper Diamond Award at IEEE ICME 2017, and the Hong Kong Scholars Award in 2014. He is a fellow of IET. He is a senior member of the IEEE.