Attention-Aware Face Hallucination via Deep Reinforcement Learning

Qingxing Cao,   Liang Lin,*   Yukai Shi,   Xiaodan Liang,   Guanbin Li

School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

caoqx@mail2.sysu.edu.cn, linliang@ieee.org, shiyk3@mail2.sysu.edu.cn, xdliang328@gmail.com,
liguanbin@mail.sysu.edu.cn

Abstract

Face hallucination is a domain-specific super-resolution problem with the goal to generate high-resolution (HR) faces from low-resolution (LR) input images. In contrast to existing methods that often learn a single patch-to-patch mapping from LR to HR images and are regardless of the contextual interdependency between patches, we propose a novel Attention-aware Face Hallucination (Attention-FH) framework which resorts to deep reinforcement learning for sequentially discovering attended patches and then performing the facial part enhancement by fully exploiting the global interdependency of the image. Specifically, in each time step, the recurrent policy network is proposed to dynamically specify a new attended region by incorporating what happened in the past. The state (i.e., face hallucination result for the whole image) can thus be exploited and updated by the local enhancement network on the selected region. The Attention-FH approach jointly learns the recurrent policy network and local enhancement network through maximizing the long-term reward that reflects the hallucination performance over the whole image. Therefore, our proposed Attention-FH is capable of adaptively personalizing an optimal searching path for each face image according to its own characteristic. Extensive experiments show our approach significantly surpasses the state-of-the-arts on in-the-wild faces with large pose and illumination variations.

1. Introduction

Face hallucination refers to generating a high-resolution face image from a low-resolution input image, which is a very fundamental problem in face analysis field and can facilitate several face-related tasks such as face attribute recognition [16], face alignment [34], face recognition [35]

in the complex real-world scenarios in which the face images are often of very low quality.

Existing face hallucination methods [32, 31, 17, 23] usually focus on how to learn a discriminative patch-to-patch mapping from LR images to HR images. Particularly, recent great progresses are made by employing the advanced Convolutional Neural Networks (CNNs) [36] and multiple cascaded CNNs [37]. The face structure priors and spatial configurations [15, 3] are often treated as external information for enhancing faces / facial parts. However, the contextual dependencies among the facial parts are usually ignored during the hallucination processing. According to the studies of human perception process [19], humans start with perceiving the whole images and successively explore a sequence of regions with the attention shifting mechanism, rather than separately processing the local regions. This finding enlightens us to explore a new pipeline of face hallucination by sequentially searching for the attentional local regions and considering their contextual dependency from a global perspective.

Inspired by the recent successes of attention and recurrent models on a variety of computer vision tasks [22, 1, 6], we propose an Attention-aware face hallucination (Attention-FH) framework that recurrently discovers facial parts and enhance them by fully exploiting the global interdependency of the image, as shown in Fig. 1. In particular, accounting for the diverse characteristics of face images on blurriness, pose, illumination and face appearance, we explore to search for an optimal accommodated enhancement route for each face hallucination. And we resort to the deep reinforcement learning (RL) method [21] to harness the model learning since this technique has been demonstrated its effectiveness on globally optimizing the sequential models without supervision for every step.

Specifically, our Attention-FH framework jointly optimizes a recurrent policy network that learns the policies to select a preferable facial part in each step and a local enhancement network for facial parts hallucination, through considering the previous enhancement results of the whole face. In this way, rich correlation cues among different facial parts can be explicitly incorporated in the local en-
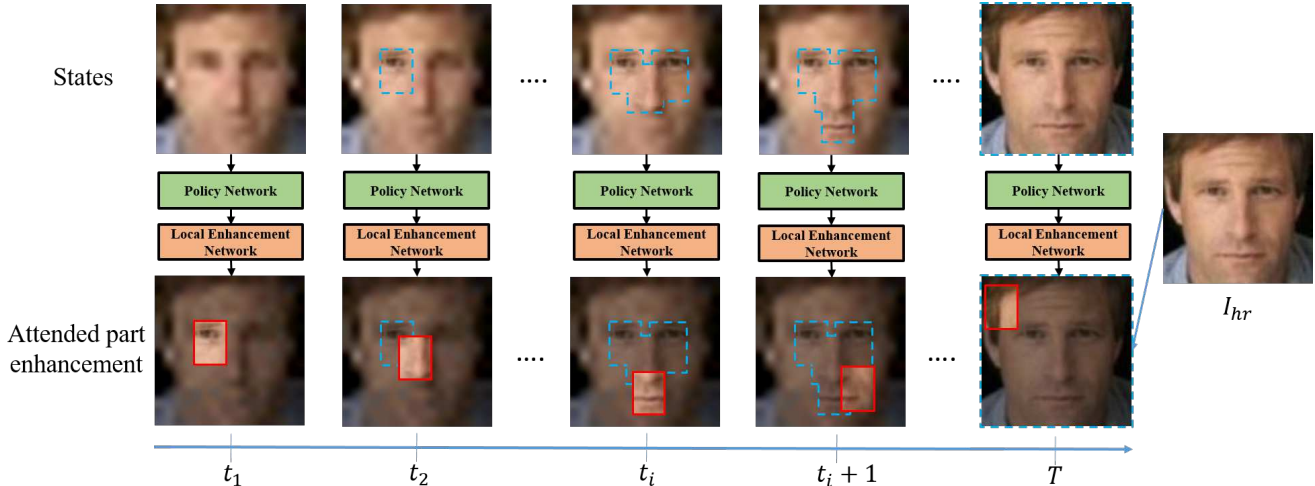
Figure 1: Sequentially discovering and enhancing facial parts in our Attention-FH framework. At each time step, our framework specifies an attended region based on the past hallucination results and enhances it by considering the global perspective of the whole face. The red solid bounding boxes indicate the latest perceived patch in each step and the blue dashed bounding boxes indicate all the previously enhanced regions. We adopt a global reward at the end of sequence to drive the framework learning under Reinforcement Learning paradigm.

hancement process in each step. For example, the agent can improve the enhancement of the mouth region by taking a more clear version of the eye region into account, as Fig. 1 illustrates.

We define the global reward for reinforcement learning by the overall performance of the super-resolved face, which drives the recurrent policy network optimization. And the recurrent policy network is optimized following the reinforcement learning (RL) procedure [26] that can be treated as a Markov decision process (MDP) maximized with a long-term global reward. In each time step, we learn the policies to determine an optimal location of next attended region by conditioning on the current enhanced whole face and the history actions. One Long Short-Term Memory (LSTM) layer is utilized for capturing the past information of the attended facial parts. And the history actions are also memorized to avoid the inference trapped in a repetitive action cycle.

Given the selected facial part in each step, the local enhancement network performs the super-resolution operation. The loss of enhancement is defined based on the facial part hallucination quality. Notably, the supervision information from the enhancement of facial parts effectively reduces unnecessary trials and errors during the reinforcement optimization.

We compare the proposed Attention-FH approach with other state-of-the-art face hallucination methods under both constrained and unconstrained settings. Extensive experiments have shown that our method substantially surpasses all of them. Moreover, our framework can explicitly generate a sequence of attentional regions during the hallucina-

tion, which finely accord with human perception process.

2. Related Work

Face Hallucination and Image Super-Resolution. Face hallucination problem is a special case of image super-resolution, which requires more informative structure priors and suffers from more challenging blurring. Early techniques made an assumption that the faces are in a controlled setting with small variations. Wang *et al.* [25] implemented the mapping between low-resolution and high-resolution faces by an eigen transformation. Yang *et al.* [32] suggested that the low-resolution and high-resolution faces have similar sparse prior and the high-resolution faces can be accurately recovered from the low-dimensional projections. In particular, Yang *et al.* [31] replaced the patch-to-patch mapping with mapping between specific facial components, which incorporates priors on face. However, the matchings between components are based on the landmark detection results which are unavailable when the down-sampling factor is large. Recently, deep convolution neural network has been successfully applied to face hallucination [37, 36] as well as image super-resolution [13, 20]. Zhen *et al.* [2] advocated the use of network cascade for image SR with a local auto-encoder architecture. Similar FCNs were also used in [3], which formulated sparse-coding based SR method into 3 layers convolution neural network. Zhou *et al.* [36] addressed the importance of appearance invariant method, and adopted fully-connected layer to perform face hallucination. Ren *et al.* [20] suggested that the unevenly distributed pixels may have different influences. They used

Shepard interpolation to efficiently achieve translation invariant interpolation(TVI).

Reinforcement Learning and Attention Networks. Attention mechanism has been recently applied and has benefited various tasks, such as object proposal[12], object classification[18], relationship detection[27], image captioning[30] and visual question answering[28]. Since contextual information is important for computer vision problems, most of these works attempted to attend multiple regions by formulating their attention procedure as a sequential decision problem. Reinforcement learning technique was introduced to optimize the sequential model with delayed reward. This technique has been applied to face detection[5] and object localization[1]. These methods learned an agent that actively locates the target regions (objects) instead of exhaustively sliding sub-windows on images. For example, Goodrich *et al.* [5] defined 32 actions to shift the focal point and reward the agent when finding the goal. Caicedo *et al.* [1] defined an action set that contains several transformations of the bounding box and rewarded the agent if the bounding box is closer to the ground-truth in each step. These two methods both learned an optimal policy to locate the target through Q-learning.

3. Attention-Aware Face Hallucination

Given a face image I_{lr} with low-resolution, our Attention-FH framework targets on its corresponding high-resolution face image I_{hr} by learning a projection function F:

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

where θ denotes the function parameters. Our Attention-FH proposes to sequentially locate and enhance the attended facial parts in each step, which can be formulated as a deep reinforcement learning procedure. Our framework consists of two networks: the recurrent policy network that dynamically determines the specific facial part to be enhanced in current step and the local enhancement network which is further employed to enhance the selected facial part.

Specifically, the whole hallucination procedure of our Attention-FH can be formulated as follows. Given the input image I_{t-1} at the t-th step, the agent of the recurrent policy network selects one local facial part $\hat{I}_{t-1}^{l_t}$ with the location l_t:

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

where f_π represents the recurrent policy network, θ_π is the network parameters. s_{t-1} is the encoded input state of the recurrent policy network, which is constructed by the input image I_{t-1} and the encoded history action h_{t-1}. g denotes a cropping operation which crops a fixed-size patch from I_{t-1} at location l_t as the selected facial part. The patch size is set as 60×45 for all face images.

We then enhance each local facial part $\hat{I}_{t-1}^{l_t}$ by our local enhancement network f_e. The resulting enhanced local patch $\hat{I}_t^{l_t}$ is computed as:

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

where θ_e is the local enhancement network parameters. The output image I_t at each t-th step is thus obtained by replacing the local patch of the input image I_{t-1} at location l_t with the enhanced patch $\hat{I}_t^{l_t}$. Our whole sequential Attention-FH procedure can be written as:

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

where T is the maximal number of local patch mining steps, $\theta = [\theta_\pi; \theta_e]$ and $f = [f_\pi; f_e]$. We set $T = 25$ empirically throughout this paper.

3.1. Recurrent Policy Network

The recurrent policy network performs the sequential local patch mining, which can be treated as a decision making process on discrete time intervals. At each time step, the agent takes action to determine an optimal image patch to be enhanced by conditioning on the current state it has reached so far. Given the selected location, the extracted local patch is enhanced through the proposed local enhancement network. During each time step, the state is updated by rendering the hallucinated face image with the enhanced facial part. The policy network recurrently selects and enhances local patches until the maximum time step is achieved. At the end of this sequence, a delayed global reward, which is measured by the mean squared error between the final face hallucination result and groundtruth high-resolution image, is employed to guide the policy learning of the agent. The agent can thus iterate to explore an optimal searching route for each individual face image in order to maximize the global holistic reward.

State: The state s_t at t-th step should be able to provide enough information for the agent to decide without looking back more than one step. It is thus composed of two parts: 1) the enhanced hallucinated face image I_t from previous steps, which enables the agent to sense rich contextual information for a new patch to be processed, e.g., the part which is still blur and requires to be enhanced; 2) the latent variable h_t obtained by forwarding the encoded history action vector h_{t-1} into the LSTM layer, which is used to incorporate all previous actions. In this way, the goal of the agent is to determine the location of the next attended local patch by sequentially observing state $s_t = \{I_t, h_t\}$ to generate a high-resolution image I_{hr}.

Action: Given a face image I with size $W \times H$, the agent targets on selecting one action from all possible locations
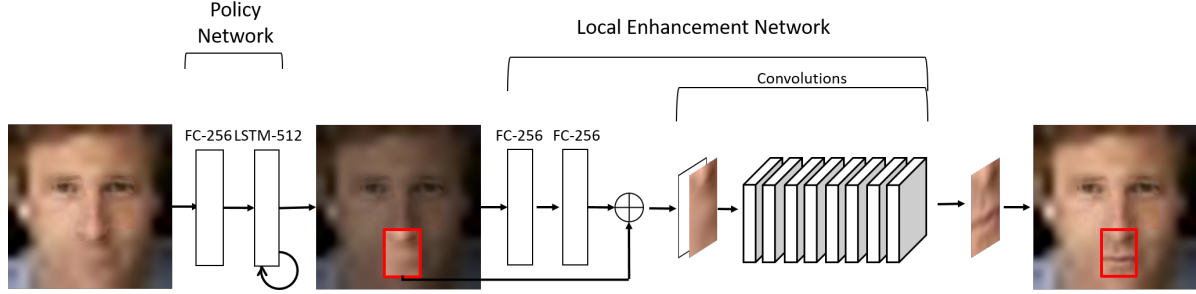
Figure 2: Network architecture of our recurrent policy network and local enhancement network. At each time step, the recurrent policy network takes a current hallucination result I_{t-1} and action history vector encoded by LSTM (512 hidden states) as the input and then outputs the action probabilities for all $W \times H$ locations, where W, H are the height and width of the input image. The policy network first encodes the I_{t-1} with one fully-connected layer (256 neurons), and then fuse the encoded image and the action vector with a LSTM layer. Finally a fully-connected linear layer is appended to generate the $W \times H$-way probabilities. Given the probability map, we extract the local patch, then pass the patch and I_{t-1} into the local enhancement network to generate the enhanced patch. The local enhancement network is constructed by two fully-connected layers (each with 256 neurons) for encoding I_{t-1} and 8 cascaded convolutional layers for image patch enhancement. Thus a new face hallucination result can be generated by replacing the local patch with an enhanced patch.

$l_t = (x, y | 1 \le x \le W, 1 \le y \le H)$. As shown in Fig. 2, at each time step t, the policy network f_π first encodes the current hallucinated face image I_{t-1} with fully-connected layer. Then the LSTM unit in policy network fuses the encoded vector with the history action vector h_{t-1}. Ultimately, a final linear layer is appended to produce a $W \times H$-way vector which indicates the probabilities of all available actions $P(l_t = (x, y) | s_{t-1})$, with each entry (x, y) indicating the probability of next attached patch located in position (x, y). The agent will then take action l_t by stochastically drawing an entry following the action probability distribution. During testing, we select the location l_t with the highest probability.

Reward: The reward is applied to guide the agent to learn the sequential policies to obtain the whole action sequence. Since our model targets on generating a hallucinated face image, we define the reward according to mean squared error (MSE) after enhancing T attended local patches at the selected locations with the local enhancement network. Given the fixed local enhancement network f_e, we first compute the final face hallucination result I_T by sequentially enhancing a list of local patches mined by $\mathbf{l} = l_{1,2,\dots,T}$. The MSE loss is thus obtained by computing $L_{\theta_\pi} = E_{p(\mathbf{l};\pi)}[\|I_{hr} - I_T\|_2]$, where $p(\mathbf{l};\pi)$ is the probability distribution produced by the policy network f_π. The reward r at t-th step can be set as:

$$r_t = \begin{cases} 0 & t < T \\ -L_{\theta_\pi} & t = T. \end{cases} \qquad (5)$$

By setting the discounted factor as 1, the total discounted reward will be $R = -L_{\theta_\pi}$.

3.2. Local Enhancement Network

The local enhancement network f_e is used to enhance the extracted low-resolution patches. Its input contains the whole face image I_{t-1} that is rendered by all previous enhanced results and the selected local patch $\hat{I}_{t-1}^{l_t}$ at current step. As shown in Fig. 2, we pass the input I_{t-1} into two fully-connected layers to generate a feature map that has the same size of the extracted patch $\hat{I}_{t-1}^{l_t}$ in order to encode the holistic information of I_{t-1}. This feature map is then concatenated with the extracted patch $\hat{I}_{t-1}^{l_t}$ and go through convolution layers to obtain the enhanced patch $\hat{I}_t^{l_t}$.

We employ the cascaded convolution network architecture similar to general image super-resolution methods [3, 13]. No pooling layers are used between convolution layers, and the sizes of feature maps are kept fixed throughout all convolution layers. We follow the detailed setting of the network employed by Tuzel *et al.*[24]. Two fully-connected layers contain 256 neurons. The cascaded convolution network is composed of eight layers. Conv1 and conv7 layers have 16 channels of 3×3 kernels; conv2 and conv6 layers have 32 channels of 7×7 kernels; conv3, conv4 and conv5 layers have 64 channels of size 7×7 kernel; conv8 has kernel of size 5×5 and outputs the enhanced image patch with the same size and channel as the extracted patch.

In the initialization, we first up-sample the image I_{lr} to the same size as high-resolution image I_{hr} with Bicubic method. Our network first generates a residual map and then combines the input low-resolution patch with the residual map to produce the final high-resolution patch. Learning from the residual map has been verified to be more effective than directly learning from the original high-resolution

images [13] [7].

3.3. Deep Reinforcement Learning

Our Attention-FH framework jointly trains the parameters θ_π of the recurrent policy network f_π and parameters θ_e of the local enhancement network f_e. We introduce a reinforcement learning scheme to perform joint optimization.

First, we optimize the recurrent policy network with REINFORCE algorithm [26] guided by the reward given at the end of sequential enhancement. The local enhancement network is optimized with mean squared error between the enhanced patch and the corresponding patch from the ground truth high-resolution image. This supervised loss is calculated at each time step, and can be minimized based on back-propagation.

Since we jointly train the recurrent policy network and local enhancement network, the change of parameters in local enhancement network will affect the final face hallucination result, which in turn causes a non-stationary objective for the recurrent policy network. We further employ the variance reduction strategy as mentioned in [18] to reduce variance due to the moving rewards during the training procedure.

4. Experiments

4.1. Datasets and implementation details

Extensive experiments are evaluated on BioID [11] and LFW [9] datasets. The BioID dataset contains 1521 face images collected under the lab-constrained settings. We use 1028 images for training and 493 images for evaluation. The LFW dataset contains 5749 identities and 13233 face images taken in an unconstrained environment, in which 9526 images are used for training and the remaining 3707 images are used for evaluation. This train/test split follows the split provided by the LFW datasets. In our experiment, we first align the images on BioID dataset with SDM method [29] and then crop the center image patches with size of 160×120 as the face images to be processed. For LFW dataset, we use aligned face images provided in LFW funneled[8] and extract the centric 128×128 image patches for processing. We evaluate two scaling factors of 4 and 8, denoted as $4\times$ and $8\times$ in the following figures and tables. The input low resolution image is generated by resizing the original image with fixed scaling factors. Thus the input images in BioID are resized as 40×30 and 20×15, and the input images in LFW are resized as 32×32 and 16×16 respectively.

We set the maximum time steps $T = 25$ in our Attention-FH model for both datasets. And the face patch size is $H \times W = 60 \times 45$ for all experiments. The network is updated using ADAM gradient descent [14]. The learning rate and the momentum term are set to 0.0002 and 0.5 respectively.

4.2. Evaluation protocols and comparisons

We adopt the widely used Peak Signal-to-Noise Ratio (PSNR), structural similarity(SSIM) as well as feature similarity (FSIM) [33] as our evaluation measurements.

We compare our method with several state-of-the-art face hallucination and image super-resolution methods. For face hallucination approaches, we compare with SFH [31], MZQ [17], BCCNN [36], GLN [24]. We also compare with two general image super-resolution methods: SRCNN [3] and VDSR [13]. For the results of VDSR and SRCNN, we carefully re-implement their models for both scaling factors 4 and 8. We first pre-train the models proposed in VDSR and SRCNN using 7,000 images from PASCAL VOC2012 [4] and then finetune them on the training sets of LFW and BioID.

4.3. Quantitative and qualitative comparisons

Table 1 shows the performance of our model and comparisons with other state-of-the-art methods. Our model substantially beats all compared methods on LFW and BioID datasets in terms of PSNR, SSIM and FSIM metrics. Specifically, the average gains achieved by our model in terms of PSNR are 2.59dB, 1.66dB, 2.43dB and 1.8dB compared with the second best method. The largest improvement by our method is 2.59dB over the the second best method in LFW on scaling factor of 4.

The traditional face hallucination methods, SFH [31] and MZQ [17], are highly dependent on the performance of facial landmarks detection. When the scaling factor is set to 8, the landmarks detection results are not reliable given the very low-resolution input, thus their performances are not as good as that of our model. As for deep-learning based methods, our proposed method outperforms the best general image super-resolution method VDSR [13] by 3.68dB, 2.05dB, 3.04dB and 2.25dB on different experiments respectively. Furthermore, our model outperforms the second best face hallucination method GLN [24] by a significant margin. Noted that GLN [24] shares similar model architecture with our local enhancement network. Therefore, the highly improved performance achieved by our model confirms the effectiveness of utilizing the attention agent.

The qualitative comparisons of face hallucination results on LFW dataset and BioID dataset are shown in Fig. 3 and Fig. 4. As can be observed from these results, our model produces much clearer images than GLN and VDSR, despite the large variations. For example, in Fig. 3, the eyes of man on the leftmost column can only be successfully recovered by our Attention-FH, which demonstrates the effectiveness of our recurrent enhanced model.

4.4. Ablation studies

We perform extensive ablation studies and demonstrate the effects of several important components in our frame-
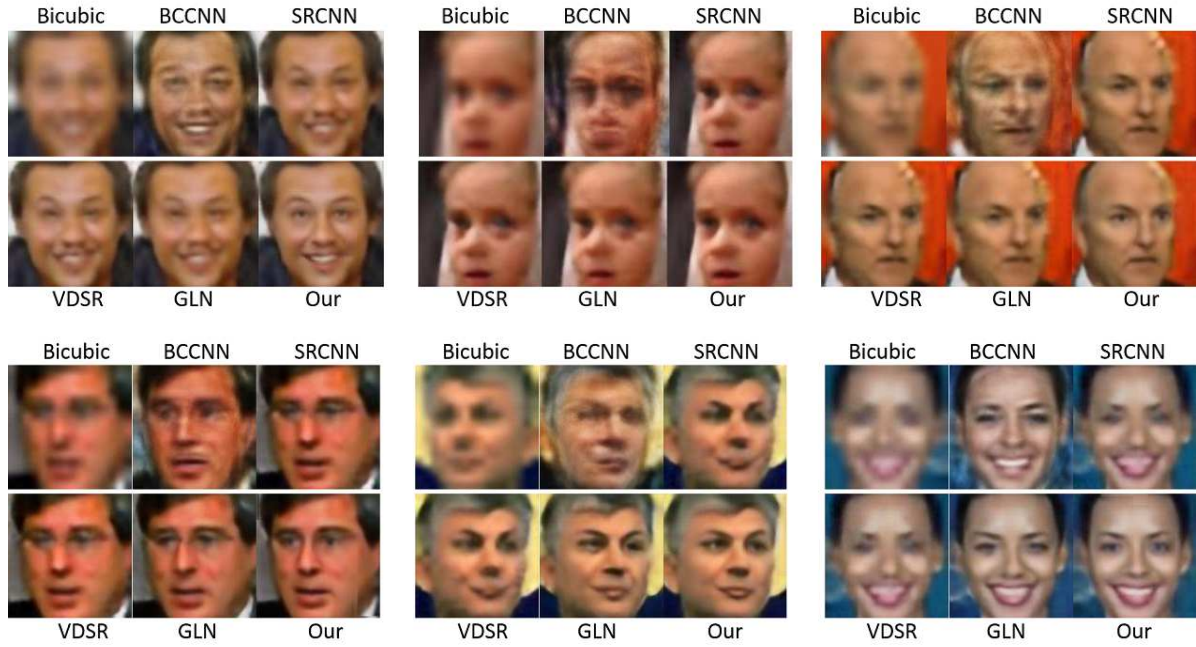
Figure 3: Qualitative results on LFW-funneled with scaling factor of 8. Results of SFH and MZQ methods are not displayed as they depends on facial landmarks which are often failed to detect in such low-resolution images.

(a) Bicubic   (b) BCCNN   (c) SFH   (d) MZQ   (e) SRCNN   (f) VDSR   (g) GLN   (h) Our   (i) Original

(j) Bicubic   (k) BCCNN   (l) SFH   (m) MZQ   (n) SRCNN   (o) VDSR   (p) GLN   (q) Our   (r) Original
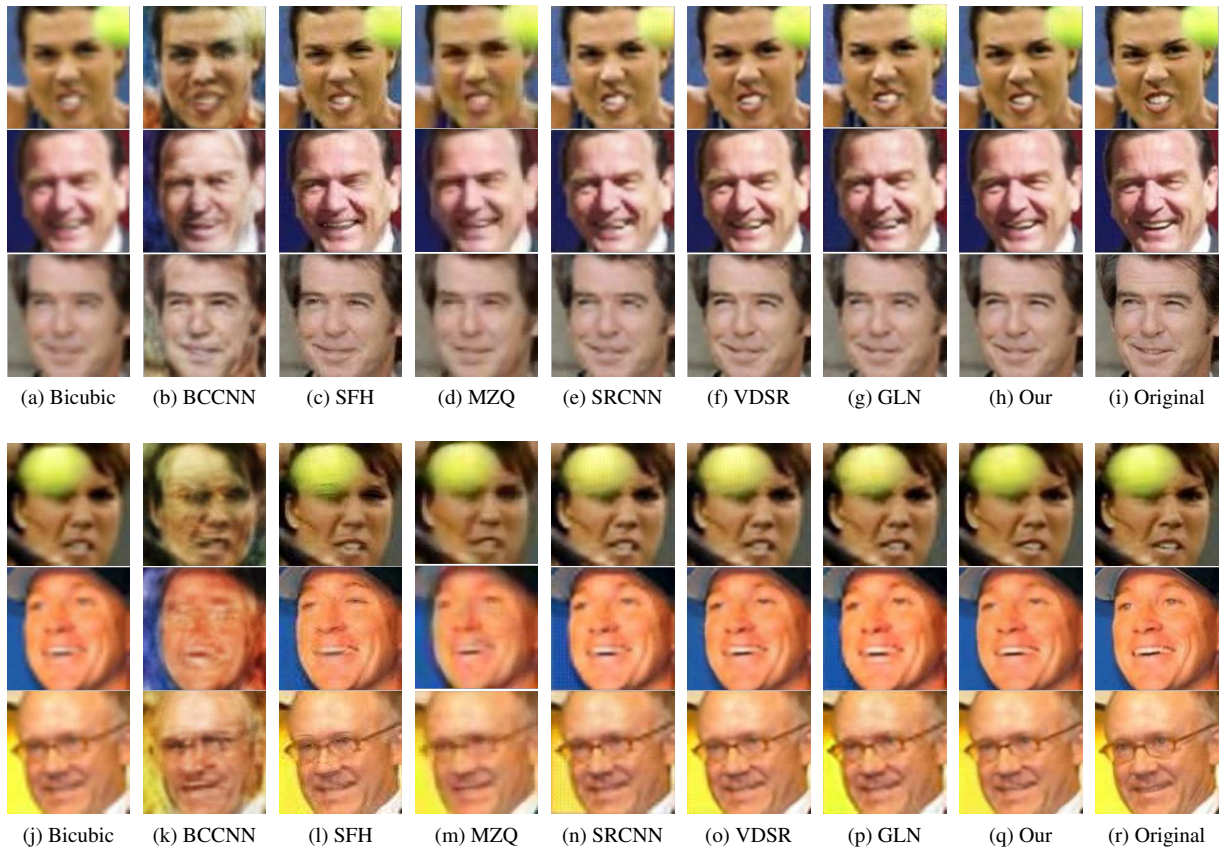
Figure 4: Qualitative results on LFW-funneled with scaling factor of 4.

Methods	LFW-funneled 4×			LFW-funneled 8×			BioID 4×			BioID 8×		
	PSNR	SSIM	FSIM	PSNR	SSIM	FSIM	PSNR	SSIM	FSIM	PSNR	SSIM	FSIM
Bicubic	26.79	0.8469	0.8947	21.92	0.6712	0.7824	25.18	0.8170	0.8608	20.68	0.6434	0.7539
SFH [31]	26.59	0.8332	0.8917	22.12	0.6732	0.7832	25.41	0.8034	0.8494	20.31	0.6234	0.7238
BCCNN [36]	26.60	0.8329	0.8982	22.62	0.6801	0.7903	24.77	0.8034	0.8421	21.40	0.6504	0.7621
MZQ [17]	25.93	0.8313	0.8865	22.12	0.6771	0.7802	24.66	0.8003	0.8573	21.11	0.6481	0.7594
SRCNN [3]	28.94	0.8686	0.9069	23.92	0.6927	0.8314	27.02	0.8517	0.8771	22.34	0.6980	0.8274
VDSR [13]	29.25	0.8711	0.9123	24.12	0.7031	0.8391	28.52	0.8627	0.8914	24.31	0.7321	0.8465
GLN [24]	<u>30.34</u>	<u>0.8922</u>	<u>0.9151</u>	<u>24.51</u>	<u>0.7109</u>	<u>0.8405</u>	<u>29.13</u>	<u>0.8794</u>	<u>0.8966</u>	<u>24.76</u>	<u>0.7421</u>	<u>0.8525</u>
Our	**32.93**	**0.9104**	**0.9427**	**26.17**	**0.7604**	**0.8630**	**31.56**	**0.9002**	**0.9343**	**26.56**	**0.7864**	**0.8748**

Table 1: Comparison between our method and others in terms of PSNR, SSIM and FSIM evaluate metrics. We employ the **bold face** to label the first place result and <u>underline</u> to label the second in each column.

	LFW 4×	LFW 8×
T = 5	28.89	23.55
T = 15	31.51	25.25
T = 25	32.93	26.17
T = 35	32.91	26.31

Table 2: Comparison of the variants of our Attention-FH using different number of steps for sequentially enhancing facial parts on LFW dataset.

	LFW 4×	LFW 8×
CNN-16	29.11	24.02
Our w/o attention	32.26	25.71
Random Patch	31.60	25.76
I_0	32.10	25.92
Spatial Transform	28.13	25.75
Our	**32.93**	**26.17**

Table 3: Comparison of our model with different architecture settings, including the 16-layers convolution network "CNN-16", the model "Our w/o attention" that recurrently enhance the whole image, the model that randomly picks patches, the model with original low-resolution input image I_0 as input for policy network, the end-to-end trainable spatial transform network, and our Attention-FH.

	SRCNN(3-layers)	VDSR(20-layers)	OURS(8-layers)
Time	20ms	100ms	81ms

Table 4: Computational cost for testing on 128×128 images.

work:

Effectiveness of increasing recursion depth. Firstly, we explore the effect of using different recursive steps T for sequentially enhancing facial parts. We train our model with four different settings $T = 5, 15, 25, 35$. Table 2 shows that the face hallucination performance gradually increases with more attention steps. The PSNR measurement improves dramatically when the number of recursion steps is low, since the extracted patches are unable to cover the whole image. When the number of recursion steps gets greater than 15, the extracted patches can cover the whole image, the step-wise performance improvement on PSNR becomes minor. This phenomenon becomes more obvious when the number of steps gets close to 25.

Effectiveness of patch-wise enhancement manner. We further evaluate the effectiveness of our patch-wise enhancement manner. In table 3, "CNN-16" indicates the results of a 16-layered fully convolution neural network. By comparing our model with "CNN-16", there are 3.82 dB and 2.14 dB improvements in terms of PSNR on LFW of factor 4 and factor 8. We also conduct another ablation study by recurrently enhancing the whole image at each step without extracting patches, named as "Ours w/o attention". This model has the same architecture as our model, and the number of recurrent steps is set to 5, which nearly covers the same area of overlapping regions as our full model. From table 3, we can see that although the recurrent model without attention can achieve promising results, our model with attention still promotes 0.67 dB and 0.46 dB on LFW $4\times$ and LFW $8\times$ respectively, which demonstrates the effectiveness of using attention-aware model and reinforcement learning.

Effectiveness of sequentially attending patches. To demonstrate the ability of our proposed model to sense meaningful attention sequence, we conduct two experiments: 1) The patch is randomly picked at each step instead of being chosen by the agent; 2) use original LR image as input for policy network instead of image that has been locally enhanced at previous steps. As reported in table 3, randomly picking patches drops 1.33dB and 0.41dB on LFW $4\times$ and $8\times$ respectively, indicating the effectiveness of the agent in our model in locating meaningful attention sequence, which can further be verified in Fig. 5. The second experiment shows that the performance will respectively drop 0.83dB and 0.25dB without previous enhanced information, which confirms that the contextual interdependency between patches can help to mine the next patches by our attention agent.

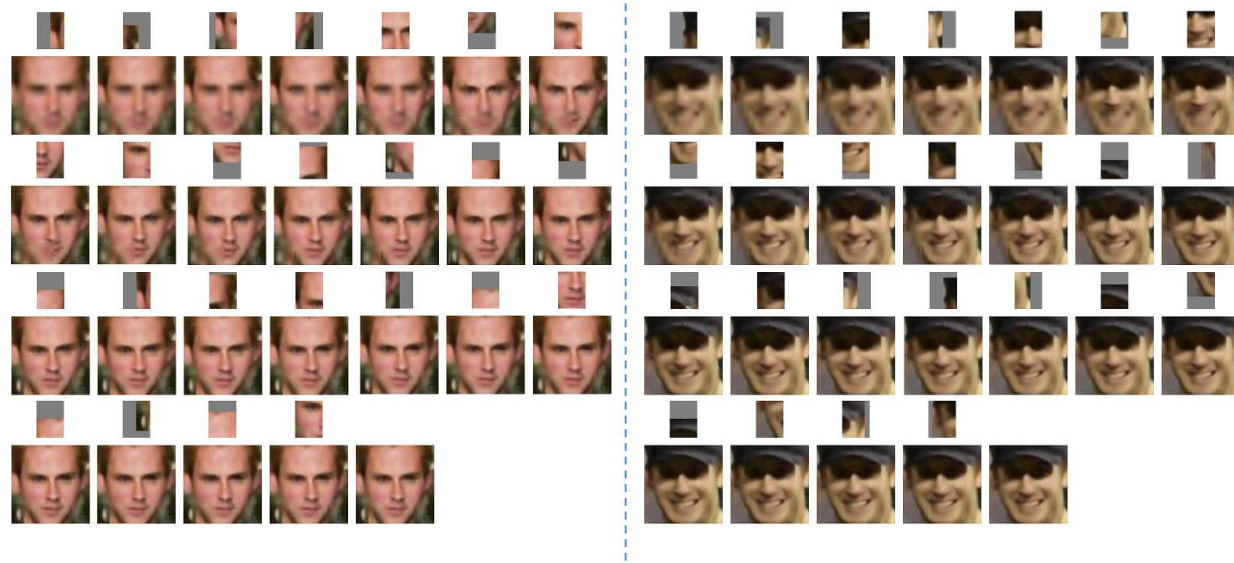**Effectiveness of reinforcement learning.** We also

Figure 5: Example results of enhancement sequences and corresponding patches selected by the agent. The gray in some patches indicates area outside of original image. Best viewed by zooming in the electronic version.

compare our reinforcement learning with end-to-end back-propagation method. Spatial transform network proposed by Jaderberg *et al.*[10] is capable of computing the sub-gradients corresponding to the location of the extracted patch. The comparison results are given in table 3.

Specifically, for "spatial transform" model, we replace the softmax output of the policy network with a 2-dimensional vector, which contains coordinate offsets x and y of the extracted patch. The transform scale s is fixed so that the size of extracted patch is also 60×45. The spatial transform layer takes $\{x, y, s\}$ as well as the face image generated at previous step as inputs, and extracts the patch for enhancement. We expand the enhanced patch with padding 0s by the spatial transform layer with a transform vector $\{-s*x, -s*y, 1/s\}$. Then, the expanded patch can be added to the low-resolution face image. We calculate the MSE loss between the groundtruth and the current enhanced result at each step and train the whole recurrent model in an end-to-end manner. Except for the spatial transform layer, all other settings and model architecture remain the same as our method.

4.5. Visualization of attended regions

We visualize the detailed steps of how the agent works in our Attention-FH framework. We show the sequences of intermediate local enhancement results, as well as the attended regions located by the agent. As shown in Fig. 5, our Attention-FH is able to first locate the corner of the images, these regions are usually flat background and are easy to be enhanced without the knowledge of particular face characteristic. Secondly, the model turns to attend the facial

components such as ear, eye, nose and mouth. Finally, the model refines detailed and high-frequency areas at last few steps.

4.6. Complexity of recurrent hallucination

We compare the computation cost of our method with other one-pass full image SR methods. The testing time of a single 128×128 image running on TITANX is reported in table 4. SRCNN[3] is a 3-layers CNN and it is the fastest among the compared methods. VDSR[13] use a very deep CNN with 20-layers, though it achieve state-of-art performance, it requires longer processing time. Our local enhancement net is an 8 layers CNN. Though it requires multiple pass, the running time is still comparable. Importantly, the extra recurrent iteration in our method is performed on patches, i.e. one forward pass per patch, it performs faster than other SR models[13].

5. Conclusion

In this paper, we propose a novel Attention-aware Face Hallucination (Attention-FH) framework and optimize it using deep reinforcement learning. In contrast to existing face hallucination methods, we explicitly incorporate the rich correlation cues among different facial parts by casting the face hallucination problem as a Markov decision process. Extensive experiments demonstrate that our model not only achieves state-of-the-art performance on popular evaluation datasets, but also demonstrates better visual results. In future, we will try to extend our model to a more general form, which can handle other low-level vision problems.

References

[1] J. C. Caicedo and S. Lazebnik. Active object localization with deep reinforcement learning. In *ICCV*, pages 2488–2496, 2015.

[2] Z. Cui, H. Chang, S. Shan, B. Zhong, and X. Chen. Deep network cascade for image super-resolution. In *ECCV*, pages 49–64, 2014.

[3] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199, 2014.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.

[5] B. Goodrich and I. Arel. Reinforcement learning based visual attention with application to face detection. In *CVPR*, pages 19–24, 2012.

[6] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. DRAW: A recurrent neural network for image generation. In *ICLR*, pages 1462–1471, 2015.

[7] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang. Convolutional sparse coding for image super-resolution. In *ICCV*, pages 1823–1831, 2015.

[8] G. B. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *ICCV*, 2007.

[9] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[10] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu. Spatial transformer networks. In *NIPS*, pages 2017–2025. 2015.

[11] O. Jesorsky, K. J. Kirchberg, and R. Frischholz. Robust face detection using the hausdorff distance. In *AVBPA*, pages 90–95, 2001.

[12] Z. Jie, X. Liang, J. Feng, X. Jin, W. Lu, and S. Yan. Tree-structured reinforcement learning for sequential object localization. In *NIPS*, pages 127–135. 2016.

[13] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. 2016.

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[15] C. Liu, H.-Y. Shum, and W. T. Freeman. Face hallucination: Theory and practice. *International Journal of Computer Vision*, 75(1):115–134, 2007.

[16] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015.

[17] X. Ma, J. Zhang, and C. Qi. Hallucinating face by position-patch. *Pattern Recognition*, 43(6):2224–2236, 2010.

[18] V. Mnih, N. Heess, A. Graves, and k. kavukcuoglu. Recurrent models of visual attention. In *NIPS*, pages 2204–2212. 2014.

[19] J. Najemnik and W. S. Geisler. Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391, 2005.

[20] J. S. Ren, L. Xu, Q. Yan, and W. Sun. Shepard convolutional neural networks. In *NIPS*, pages 901–909, 2015.

[21] 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*, 529:484–503, 2016.

[22] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015.

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

[24] O. Tuzel, Y. Taguchi, and J. R. Hershey. Global-local face upsampling network. *arXiv preprint arXiv:1603.07235*, 2016.

[25] X. Wang and X. Tang. Hallucinating face by eigentransformation. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(3):425–434, 2005.

[26] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.

[27] L. Xiaodan, L. Lisa, and P. X. Eric. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *CVPR*, 2017.

[28] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *ICML*. 2016.

[29] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013.

[30] 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 *ICML*, pages 2048–2057, 2015.

[31] C.-Y. Yang, S. Liu, and M.-H. Yang. Structured face hallucination. In *CVPR*, pages 1099–1106, 2013.

[32] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

[33] L. Zhang, L. Zhang, X. Mou, and D. Zhang. Fsim: a feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

[34] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):918–930, 2016.

[35] 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.

[36] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Learning face hallucination in the wild. In *AAAI*, pages 3871–3877, 2015.

[37] S. Zhu, S. Liu, C. C. Loy, and X. Tang. Deep cascaded bi-network for face hallucination. *arXiv preprint arXiv:1607.05046*, 2016.