Attribute Augmented Convolutional Neural Network for Face Hallucination

Cheng-Han Lee[1] Kaipeng Zhang[1] Hu-Cheng Lee[1] Chia-Wen Cheng[2] Winston Hsu[1]
[1]National Taiwan University [2]The University of Texas at Austin
[1]{r05922077, r05944047, r05922174, whsu}@ntu.edu.tw [2]cwcheng@cs.utexas.edu

Abstract

Though existing face hallucination methods achieve great performance on the global region evaluation, most of them cannot recover local attributes accurately, especially when super-resolving a very low-resolution face image from 14 × 12 pixels to its 8 × larger one. In this paper, we propose a brand new Attribute Augmented Convolutional Neural Network (AACNN) to assist face hallucination by exploiting facial attributes. The goal is to augment face hallucination, particularly the local regions, with informative attribute description. More specifically, our method fuses the advantages of both image domain and attribute domain, which significantly assists facial attributes recovery. Extensive experiments demonstrate that our proposed method achieves superior visual quality of hallucination on both local region and global region against the state-of-the-art methods. In addition, our AACNN still improves the performance of hallucination adaptively with partial attribute input.

1. Introduction

Face hallucination is a domain-specific image super resolution technique which generates high resolution (HR) facial images from low-resolution (LR) inputs. Different from generic image super resolution methods, face hallucination exploits special facial structures and textures. In some applications such as face recognition in video surveillance system and image editing, face hallucination can be thought as a preprocessing step for these face-related applications.

Face hallucination has attracted great attention in the past few years [2, 8, 10, 12, 15, 7, 19, 16, 20]. All of previous works only utilize low resolution images as input to generate high resolution outputs without leveraging attribute information. Most of them cannot accurately hallucinate local attributes or accessories in ultra-low-resolution (i.e. 14 × 12 pixels). When downsampling a face image by 8× upscaling factor, almost 98.5% of the information is missing including some facial attributes (e.g. eyeglasses, beard etc.). Therefore, these methods achieve great performance only on the global region rather than local region.

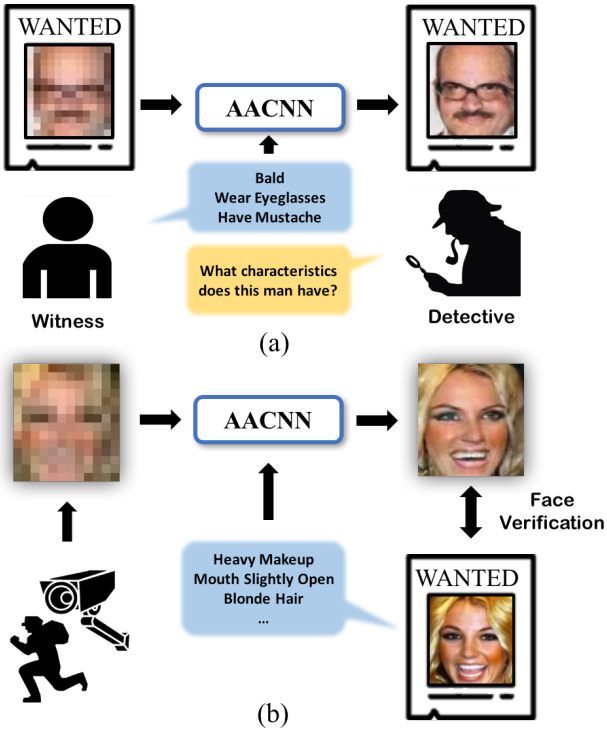


Figure 1. (a) Scenario I of AACNN : A detective questions a witness about more information of the suspect, because the suspect was only recorded by surveillance system with low resolution face. By the help of AACNN, the detective can obtain a more distinct wanted poster with clear facial attributes. (b) Scenario II of AACNN : We can get most facial attributes of the suspect from a high-resolution wanted poster to help hallucinate the low resolution face recorded by surveillance system. With this method, we can check if the recorded face is the suspect by face verification.

In this paper, we propose a novel Attribute Augmented Convolutional Neural Network (AACNN) which is the first method exploiting extra facial attribute information to overcome the above issue. Our model can be applied in two real-world scenarios. (i) A detective only has a wanted poster of the suspect with low-resolution face. He can obtain the de-

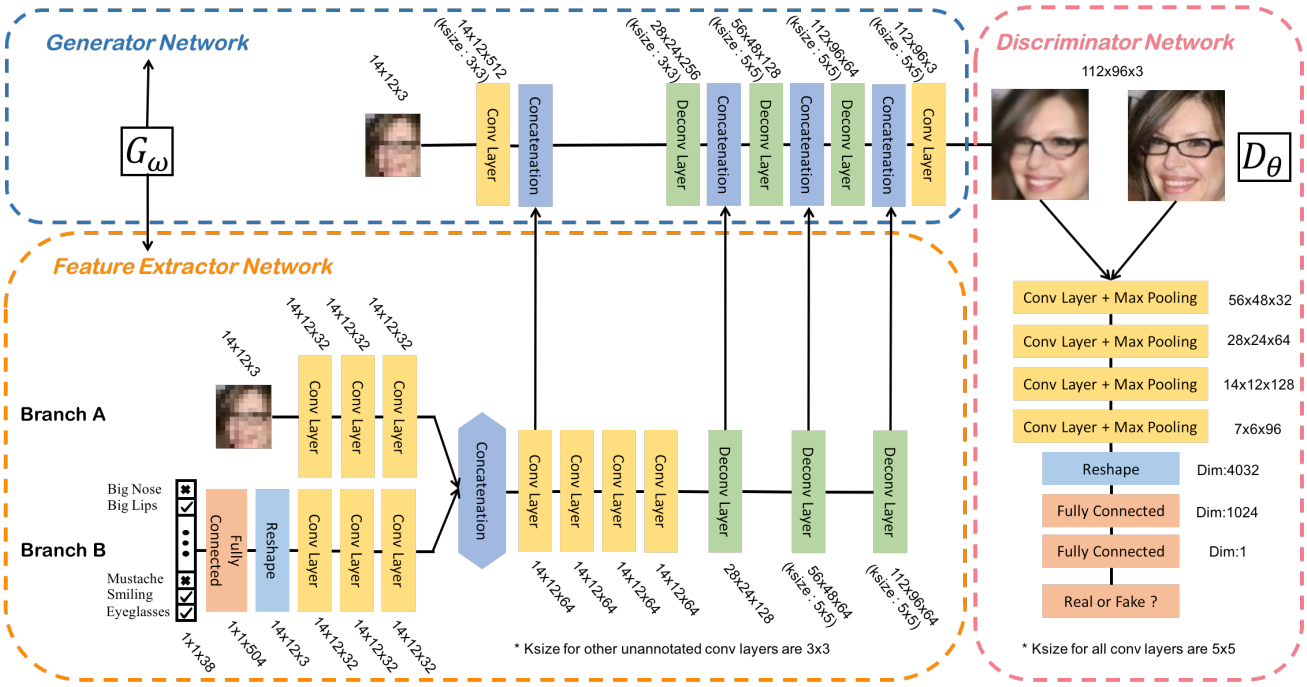


Figure 2. The network structure of Attribute Augmented Convolutional Neural Network (AACNN). AACNN contains three components : generator, feature extractor and discriminator. The generator network is responsible for learning mapping from LR image to HR image. The feature extractor network is responsible for extracting features, fusing two different feature domains, and guiding the generator towards the target HR image. Branch A can exploit more fine-grained information from low resolution facial image than the generator network. Branch B can extract high semantic features from input attributes and transform features into LR image shape which can perceptually learn the semantic from attributes. The discriminator is responsible for distinguishing real or fake of a input face.

tails of the suspect's facial attributes by questioning a witness . With the help of AACNN, the detective can receive a more distinct wanted poster with clear facial attributes as shown in Fig.1 (a). (ii) We can get most of the suspect's facial attributes from a high-resolution wanted poster to help hallucinate low resolution faces recorded by surveillance system. With this method, we can check if the recorded face is the suspect by face verification as shown in Fig.1 (b). Therefore, with the help of attribute information, our network can hallucinate low resolution images better in a novel way. AACNN utilizes both LR facial images and corresponding attributes as input to super-resolve a tiny ($i.e.$14 $\times$ 12 pixels) face image by a remarkable upscaling factor 8, where we reconstruct 64 pixels for each single pixel of the input LR image.

In the real world situation, since humans are impossible to know all the attributes of a face, we define a representation of unknown attribute. AACNN can still exploit partial information to help hallucinate LR faces and have superior visual quality. Details are shown in Sec. 3.3 and Sec. 4.4.

In Fig. 2, our network consists of three components: generator network, feature extractor network and discriminator network. The generator network is responsible for learning mapping from LR image to HR image. The feature extractor network is responsible for extracting features, fusing two different feature domains, and guiding the generator towards the target HR image. The discriminator is responsible for distinguishing real or fake of an input face. The compositions of LR images are essentially different from the compositions of attributes. For this reason, we develop a domain fusion method to solve this problem.

Overall, our main contributions are as following:

- We propose a brand new Attribute Augmented Convolutional Neural Network (AACNN) using attribute information to assist hallucinate low-resolution face images with 8$\times$ scaling factor. In particular, we propose a novel perceptual fusion method from image and attribute domains.
- Compared with previous state-of-the-art methods, our proposed method achieves superior visual quality of hallucination on both global and local regions.
- Our AACNN still improves the performance of hallucination adaptively with partial attribute inputs.

2. Related work

2.1. Face hallucination

Face hallucination is a special case of single-image super resolution which aims at recovering a high-resolution image for single low-resolution image. Generic image super-resolution does not take image class information into account. Face hallucination is a class-specific problem on human face which aims to exploit statistical information on facial images. Because face hallucintion super-resolves images of a specific class, it usually attains better results than generic methods. State-of-the-art face hallucination methods can be grouped into three categories: holistic face based methods, facial component based methods, and convolutional neural network (CNN) based methods.

Holistic face based methods learn a global face model. Wang et al. [13] develop an eigen-transformation method to generate HR face by finding a linear mapping between LR and HR face subspaces. Liu et al. [8] employs a global face model learning by Principal Component Analysis (PCA). Ma et al. [10] samples LR exemplar patches from aligned HR face images to hallucinate faces. Holistic face based methods require precisely aligned reference HR and LR facial images with the same pose and facial expression.

Facial component based methods resolve facial parts rather than the entire face, and thus can address various poses and expressions. Tappen et al. [12] exploits SIFT flow to align facial parts of LR images, and then reconstruct LR face images by warping corresponding HR face images. However, the global structure is not preserved because of using local mapping. Yang et al. [15] proposes a structured face hallucinated method to maintain the facial structure. However, it needs accurate facial landmark to assist.

Convolutional neural networks based methods have claimed the state-of-the-art performance recently. Zhou et al. [19] presents a bi-channel CNN to hallucinate blurry face images. They firstly use CNN to extract facial features. Zhu et al. [20] jointly learns face hallucination and face spatial configuration estimation. However, the results of these methods look over-smooth due to using pixel-wise Euclidean distance loss.

2.2. Generative adversarial network

Goodfellow et al. [3] introduce the GAN framework to simultaneously train generator and discriminator that compete with each other. This model can generate realistic images form random noise. Radford and Metz et al. [11] propose a set of constraints on the architectural topology of Convolutional GANs (DCGAN) that make them stable to train in most settings. Arjovsky et al. [1] introduce a new method to measure the distance of two data distribution called Wasserstein GAN which makes training processes of GAN more stable. GAN is generally a popular generative model recently which can generate realistic images.

2.3. Face hallucination with adversarial training

For pre-aligned faces, Yu et al. [16] first introduces Generative Adversarial Network (GAN) to solve face hallucination. This method jointly uses the pixel-wise Euclidean distance loss and the adversarial loss, which aims to generate a realistic facial image closet to the average of all potential faces. For un-aligned faces, Yu et al. [17] which is a continuation of [16] proposes Transformative Discriminative Neural Network (TDN) by concatenating the spatial transformation layers for solving deficient results because of unaligned tiny input. Given noisy and unaligned tiny input, Yu et al. [18] introduce Transformative Discriminative Autoencoders (TDAE) which uses autoencoder architecture and discriminator network by concatenating the spatial transformation layers to solve deficient results. By leveraging adversarial training, we can make hallucinated images more realistic. However, these works have weak ability to recover detailed facial attributes.

3. Proposed Method

3.1. Overall Framework

The problem we have to solve is to hallucinate a very low-resolution face image from 14×12 pixels to its $8 \times$ larger one. We first recover such low-resolution images with assist of additional facial attribute information. The inputs of our framework are tiny ($i.e.14 \times 12$ pixels) face images and discrete attribute vectors with 38 elements. We also define a representation of unknown attribute, and replace each attribute vector with specific unknown proportion in unknown attribute experiment (see Sec. 3.3). The outputs are clear face images with 112×96 pixels. By using convolution neural network, we can fuse different domain features and super-resolve low resolution images (see Sec. 3.4). Our framework contains three components which are generator network, feature extractor network and discriminator network (see Sec. 3.2).

3.2. Network Architecture

Our AACNN contains three components : generator, feature extractor and discriminator.

Generator network. In Fig. 2 , the structure of our generator network uses learnable transposed convolution layer for super-resolution due to its superior performance. It is responsible for learning a mapping between low resolution image and high resolution image and receiving the features from feature extractor. We use PReLU [4] activation function after each layer in convolution and deconvolution stage except for image reconstruction which utilizes tanh.

Feature extractor network. In our model, we introduce feature extractor network (Fig. 2) to extract feature from

both low resolution image and attribute, and fuse them together. The extractor injects guidance to the generator at every upsampling scale, and assists the generator to learn the features from image and attribute. The feature extractor consists of two sub branches as shown in Fig. 2. These two branches will concatenate together before upsampling layer. Branch A uses three convolution layers to extract fine-grained features of low resolution faces before upsampling. Branch B takes attribute as input, expands its dimension from 38 to 504 by fully connected layer, and then reshapes it to the same size of LR image ($14 \times 12 \times 3 = 504$). The following convolutional process is the same as Branch A. We use PReLU [4] activation function after all layers.

Discriminator network. The discriminator network is responsible for distinguishing real or fake of a input face. In Fig. 2, the structure of our discriminator is a 6-layer CNN network. The inputs are generated images and ground truth images and the output is the probability of input being realistic image. We follow the setting of DCGAN [11] which uses LeakyReLU [14] as activation function except for the last layer which uses a sigmoid function, and batch normalization [5] added to all convolutional layers.

3.3. Problem Formulation

In vanilla experiment, for a LR face input I_i^{LR}, its corresponding attribute input is $I_i^A = \{I_{i_1}^A, I_{i_2}^A, ..., I_{i_38}^A\}$ and $I_{i_n}^A \in \{-1, +1\}, n = 1, 2, ..., 38$ where $\{+1\}$ means that the face contains target attribute and $\{-1\}$ means that the face doesn't contain target attribute.

In the real world situation, since humans are impossible to know all the attributes of a face, we define a representation of unknown attribute. In unknown attribute experiment, the corresponding attribute input of the LR image is $I_{i_n}^A \in \{-1, 0, +1\}, n = 1, 2, ..., 38$ where $\{0\}$ means that the person providing attribute input doesn't know if the target attribute classes exist or not. We randomly change some known attributes into unknown one. More details are shown in Sec. 4.4.

We use pixel-wise Euclidean distance loss, called super-resolution (SR) loss, to constrain the overall appearance between a hallucinated facial image and its corresponding high-resolution facial image, and adversarial loss to make hallucinated facial image more realistic.

We penalize pixel-wise Euclidean distance between hallucinated face and the corresponding HR face:

$$L^{SR}(I_i^A, I_i^{LR}, I_i^{HR}) = \left\| G_\omega(I_i^A, I_i^{LR}) - I_i^{HR} \right\|_2^2, \quad (1)$$

where I_i^A, I_i^{LR} and I_i^{HR} are ith attribute vector, LR facial image and HR facial image respectively in the training data, and $G_\omega(I_i^A, I_i^{LR})$ is the hallucination model output for I_i^A and I_i^{LR}.

The objective function is represented as:

$$\min_\omega \frac{1}{N} \sum_{i=1}^{N} L^{SR}(I_i^A, I_i^{LR}, I_i^{HR}), \quad (2)$$

We also further use adversarial training strategy to encourage $G_\omega(I_i^A, I_i^{LR})$ to construct high-quality results. The GAN simultaneously trains a generator network , G, and discriminator network, D. The training process alternates optimizing the generator and discriminator, which compete with each other. The generator learns to generate samples that can fool the discriminator. The discriminator learns to distinguish real data and samples from generator. The loss function we use is as following:

$$\begin{aligned} &L^{adv}(G_\omega(I_i^A, I_i^{LR}), I_i^{HR}) \\ &= logD_\theta(I_i^{HR}) + log(1 - D_\theta(G_\omega(I_i^A, I_i^{LR}))), \end{aligned} \quad (3)$$

The objective function with adversarial loss is represented as:

$$\begin{aligned} \max_\theta \min_\omega \frac{1}{N} \sum_{i=1}^{N} & L^{SR}(I_i^A, I_i^{LR}, I_i^{HR}) \\ & + \lambda L^{adv}(G_\omega(I_i^A, I_i^{LR}), I_i^{HR}), \end{aligned} \quad (4)$$

where λ is trade-off weight, ω denotes the parameters of hallucination model G_ω which consists of generator network and feature extractor network and θ denotes the parameters of D_θ which consists of discriminator network. All parameters are optimized using stochastic gradient descent (SGD) with standard backpropagation.

3.4. Perceptual fusion from image and attribute domain

The information containing in LR images mostly dissimilates to the one in attributes. As this reason, we propose a method to fuse low resolution image features and attribute features, and design a feature extractor which consists of two sub branches. In Fig. 2, each sub branch extracts complementary features. Branch A can exploit more fine-grained information from low resolution facial image than that from the generator network. Branch B can extract high semantic features from input attributes and transform those features into LR image shape which can perceptually learn the meaning from attributes without knowing the information of Branch A. We can see an example in Fig. 3. After extracting two complementary of feature maps, we choose concatenation to fuse features, because the overlap of two different domain features is small. Finalliy, we expand and inject those features to every scale of the generator network.

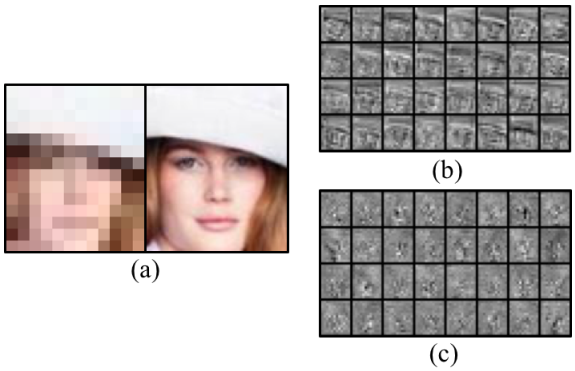


Figure 3. Visualization of feature maps in the concatenated layer. (a) Reference LR and HR face pair. (b) Visualization of the first half concatenated layer (Branch A). Branch A can exploit more fine-grained information from low resolution facial image than the generator network. (c) Visualization the last half concatenated layer (Branch B). Branch B can extract high semantic features from input attributes and transform features into LR image shape which can perceptually learn the semantic from attributes.

Method	PSNR	SSIM
Baseline - L^{SR}	26.8585	0.7535
A - L^{SR}	27.3134	0.8001
B - L^{SR}	27.1243	0.7949
A + B (AACNN - L^{SR})	**27.4007**	**0.8036**

Table 1. Quantitative comparison on the global region with the combinations of different sub branch. A + B can make the performance improve a lot due to combining attribute information and fined-grand features of LR faces.

Method	PSNR	SSIM
Bicubic	24.2669	0.6700
Ma et al. [10]	23.8438	0.7119
LapSRN [6]	25.6547	0.7212
UR-DGN [16]	24.0931	0.6843
Baseline - L^{SR}	26.8585	0.7825
AACNN - L^{SR}	**27.4007**	**0.8036**
AACNN - $L^{SR} + L^{adv}$	25.3428	0.7118

Table 2. Quantitative comparison on the global region with the state-of-the-art methods. AACNN - L^{SR} have superior performance on both PSNR and SSIM than other state-of-the-arts.

4. Experiments

4.1. Implementation details

Training data. We use CelebA dataset [9] to learn our model. It consists of 202599 face images, and each image uses similarity transformation based on five landmarks (two eyes, nose and mouth corners) to align facial images to 96 × 112 pixels images. Every image in CelebA goes with 40 attribute classes. We use only 38 classes, because there are 2 classes out of the region we cropped and aligned in data preprocessing. We select 100000 images in CelebA as training set, and generate LR face images by downsampling without aliasing. In experiments of unknown attribute, we replace specific proportion of known attribute with unknown attribute.

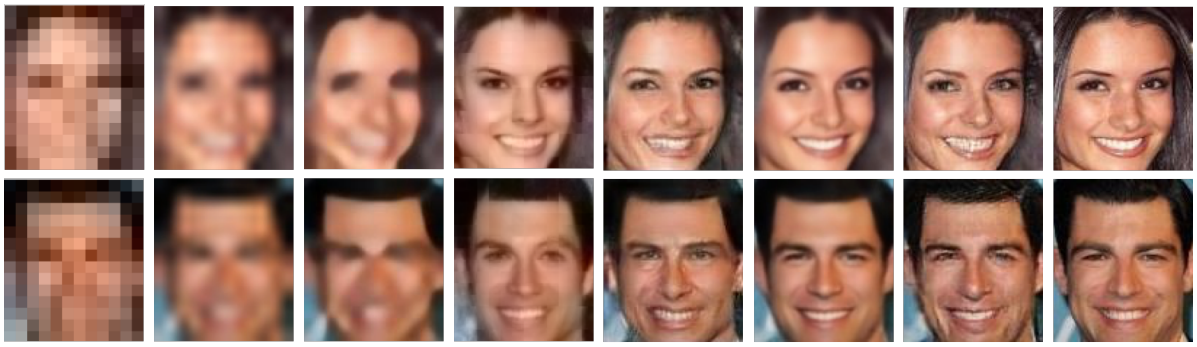
Testing data. We also use CelebA dataset with the same preprocessing as training data to evaluate our model. In global evaluation, we randomly choose 10000 images in the remaining images of CelebA as global region testing set. In local evaluation, we randomly select 20000 images in remaining images of CelebA as local region testing set. We use 8 specific attribute classes which perform significant improvement in restoration, and constitute 8 subsets from local testing set. Each class-specific subset contains 1000 images, and we make overlap region of 8 subset images as large as possible. In experiments of unknown attribute, we also replace specific proportion of known attribute with unknown attribute, and we make sure that target attribute for local evaluation will not be replaced.

Training details. As shown in Fig. 2, we implement the AACNN model by using the Caffe library with our modification. For the model training, we use the batch size of 64. The learning rate is started from 0.0005, and is divide by 1.25 after each 3000 iterations. The optimization algorithm we used is RMSProp. We set the decay rate to 0.99 and weight decay rate to 0.0005. For AACNN - $L^{SR} + L^{adv}$, we set λ (see Eq. 4) to 0.01.

Evaluation on combinations of different sub branch. In Table 1, we discuss about the performance of different sub branch combinations. AACNN have all two sub branches, and get the best performance. Our baseline model is purely generator network which uses pixel-wise Euclidean distance loss without feature extractor network and discriminator network. Branch A is important for extracting fined-grand LR face features. Branch B extract purely attribute information without LR image features, and get lower performance than using only Branch A. A + B can make the performance improve a lot due to combining attribute information and fined-grand features of LR faces.

4.2. Evaluation on global region of face hallucination

In global region evaluation, we evaluate the face image with complete size (96 × 112 pixels) by image super resolution evaluation metrics : Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM). Firstly, we investigate different combinations of feature extractor's sub branch. Then, we compare AACNN with other



(a) LR (b) Bicubic (c) LapSRN (d) Ma et al. (e) UR-DGN (f) AACNN - L^{SR} (g) AACNN - $L^{SR} + L^{adv}$ (h) HR

Figure 4. Comparison with the state-of-the-art methods on hallucination global test dataset. (a) Low-resolution inputs images. (b) Bicubic interpolation. (c) LapSRN [6]. (d) Ma et al. [10]. (e) UR-DGN [16]. (f) AACNN - L^{SR}. (g) AACNN - $L^{SR} + L^{adv}$. (h) High-resolution images. (f) and (g) both shows superior hallucinated effect on visual results. (g) is more realistic than (f) with adversarial training.

state-of-the-arts in recent years. More results are shown in supplementary material.

Comparing with state-of-the-arts.
We compare our AACNN with bicubic interpolation, our baseline model, and other three state-of-the-art methods. Our AACNN have superior visual results as shown in Fig. 4. For Ma et al. [10], LapSRN [6], and UR-DGN [16], we use their released source code. In the case of UR-DGN, we especially retrain on our aligned face images which size is different from original setting. The quantitative comparison are shown in Table 2.

Ma et al. [10] samples LR exemplar patches from aligned HR face images to hallucinate faces. It suffers from obvious blocking artifacts especially on large pose.

LapSRN [6] is design to solve general super resolution problem. It jointly optimizes the upsampling filters with deep convolution neural layers to predict sub-band residuals and progressive reconstruct multiple intermediate SR prediction by using Laplacian pyramid. We retrained LapSRN with CelebA. However, it shows blurry results on facial image with a remarkable upscaling factor 8.

UR-DGN [16] exploits generative adversarial networks (GAN) framework for face hallucination. It jointly uses the pixel-wise euclidean distance loss and the adversarial loss, which aims to generate a realistic facial image closet to the average of all potential faces. Although GAN can generate realistic face images, the results of UR-DGN sometimes looks distorted or disappeared for specific attributes.

In quantitative results, we compare our AACNN with other methods by using average PSNR and SSIM. LapSRN [6] gets great performance, but it seems not clear enough and loses lots of details on visual results. The results of UR-DGN [16] and AACNN - $L^{SR} + L^{adv}$ shows lower performance on PSNR, because the objective of adversarial loss is to make hallucinated images more realistic but close the distance of hallucinated images and HR images. Table 2 shows that AACNN - L^{SR} have superior performance on both PSNR and SSIM than other state-of-the-arts because of introducing attribute information to low-resolution face hallucination.

Method	Eyeglasses PSNR / SSIM	Narrow eyes PSNR / SSIM
Bicubic	20.46 / 0.457	21.79 / 0.533
Ma et al. [10]	19.75 / 0.488	21.43 / 0.588
LapSRN [6]	22.81 / 0.551	24.33 / 0.621
UR-DGN [16]	19.75 / 0.438	21.62 / 0.568
Baseline - L^{SR}	21.77 / 0.551	24.12 / 0.670
A - L^{SR}	22.12 / 0.579	24.68 / 0.696
B - L^{SR}	23.63 / 0.632	26.44 / 0.764
AACNN - L^{SR}	**23.77 / 0.643**	**26.81 / 0.779**
AACNN - $L^{SR} + L^{adv}$	21.66 / 0.514	24.81 / 0.689

Table 3. Quantitative comparison on local region - "eye" part with the state-of-the-art methods on the class specific test dataset. We can observe that eyeglasses is the hardest one to recover among this part. AACNN - L^{SR} still has superior performance in this region.

4.3. Evaluation on local region of face hallucination

Since global region evaluation is hard to reflect improvement of facial detail enhacement, we crop smaller regions from original images to enlarge the evaluation effect of attribute recovery. In local region evaluation, we evaluate the face image by image super resolution evaluation metrics (PSNR and SSIM) with 3 different cropped sizes and locations as shown in Fig. 5.

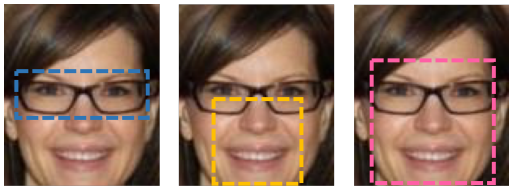
In Table 3, we discuss two attributes in eye part. Eye-

Method	Mouth slightly open PSNR / SSIM	Goatee PSNR / SSIM	Mustache PSNR / SSIM	Big nose PSNR / SSIM
Bicubic	22.96 / 0.507	22.34 / 0.480	22.48 / 0.486	23.00 / 0.506
Ma et al. [10]	23.06 / 0.603	21.79 / 0.530	22.04 / 0.545	22.78 / 0.584
LapSRN [6]	24.98 / 0.570	24.52 / 0.544	24.60 / 0.550	24.95 / 0.566
UR-DGN [16]	22.75 / 0.554	20.81 / 0.467	20.77 / 0.474	22.19 / 0.537
Baseline - L^{SR}	25.45 / 0.669	23.85 / 0.598	24.08 / 0.605	25.04 / 0.648
A - L^{SR}	26.01 / 0.700	24.40 / 0.634	24.67 / 0.642	25.58 / 0.678
B - L^{SR}	27.65 / 0.757	26.18 / 0.690	26.28 / 0.696	27.20 / 0.735
AACNN - $\boldsymbol{L^{SR}}$	**27.98 / 0.773**	**26.40 / 0.704**	**26.55 / 0.711**	**27.49 / 0.750**
AACNN - $\boldsymbol{L^{SR} + L^{adv}}$	25.55 / 0.671	23.97 / 0.577	24.09 / 0.584	24.88 / 0.629

Table 4. Quantitative comparison on local region - "mouth & nose" part with the state-of-the-art methods on the class specific test dataset. We can observe that beard (i.e. Goatee and Mustache) is the hardest one to recover among this part. AACNN - L^{SR} still has superior performance in this region.

Method	Heavy Makeup PSNR / SSIM	Chubby PSNR / SSIM
Bicubic	22.20 / 0.582	22.51 / 0.534
Ma et al. [10]	22.20 / 0.664	22.05 / 0.585
LapSRN [6]	24.49 / 0.653	24.75 / 0.602
UR-DGN [16]	22.23 / 0.635	21.52 / 0.539
Baseline - L^{SR}	24.94 / 0.730	24.20 / 0.639
A - L^{SR}	25.46 / 0.751	24.65 / 0.668
B - L^{SR}	27.17 / 0.814	26.42 / 0.724
AACNN - $\boldsymbol{L^{SR}}$	**27.55 / 0.826**	**26.61 / 0.734**
AACNN - $\boldsymbol{L^{SR} + L^{adv}}$	25.35 / 0.735	24.42 / 0.620

Table 5. Quantitative comparison on local region - "face" part with the state-of-the-art methods on the class specific test dataset. Heavy makeup distribute with a large area on face region. AACNN - L^{SR} still has superior performance in this region.



(a) Eye part (b) Mouth & nose part (c) Face part

Figure 5. 3 types of local regions cropped from original size (96 $\times$ 112 pixels). (a) Cropped size: 90 $\times$ 30 pixels. (b) Cropped size: 50 $\times$ 50 pixels. (c) Cropped size: 74 $\times$ 75 pixels. Since global region evaluation is hard to reflect improvement of facial detail enhacement, we crop smaller regions from original images to enlarge the evaluation effect of attribute recovery.

glasses are the hardest one to recover. It can be divided into two types - sunglasses and common eyeglasses. Sunglasses remain information on LR images, but common eyeglasses only remain a little. In the case of common eyeglasses, target attribute on results of most methods may be disappeared or distorted. Some examples are shown in Fig. 6. In Table 4, we discuss four attributes in mouth & nose part. Beard (i.e. Goatee and Mustache) is the most difficult one to recover, because it gets inferior performance among four attributes. In Table 5, we discuss two attributes on face part. We crop a face size square to evaluate face region, because some attributes distribute with a large area on face like heavy makeup. Our AACNN - L^{SR} achieves superior quantitative results on three local regions than other state-of-the-art methods. Different from global evaluation, Branch B gets higher performance than using only Branch A due to enhancing local region with attribute information.

Region	100 / 100 PSNR	50 / 100 PSNR	50 / 50 PSNR	50 / 25 PSNR
Global region	27.40	27.35	27.36	27.34
Eyeglasses	23.77	23.74	23.73	23.72
Goatee	26.40	26.36	26.35	26.32

Table 6. Quantitative comparison on global and local region with different proportion of known attribute in training and testing. In the first row, left number denotes the proportion of known attribute in training data, and right number denotes the proportion of known attribute in testing data. Our AACNN still improves the performance of hallucination adaptively with model trained and tested by partial attribute inputs.

For visual results showing in Fig. 6, we can see some samples compared with previous methods where our AACNN has superior visual quality especially on eyeglasses. Both (g) and (h) can hallucinate specific attribute accurately in visual results. (h) is more realistic than (g) with adversarial training.

4.4. Evaluation on unknown attribute situation

In this section, we do an auxiliary experiment for unknown attribute situation. We randomly change some

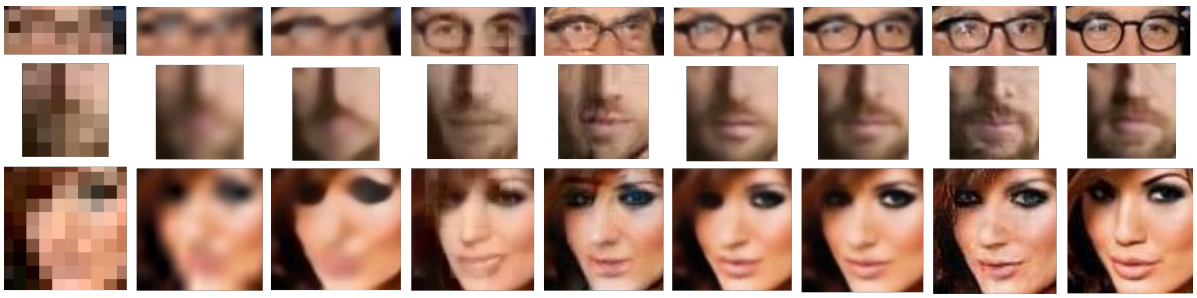


Figure 6. Comparison with the state-of-the-art methods on hallucination local test dataset. The first row is eyeglasses on "eye" part, the middle row is goatee on "mouth & nose" part, and the rest is heavy makeup on "face" part. (a) Low-resolution inputs images. (b) Bicubic interpolation. (c) LapSRN [6]. (d) Ma et al. [10]. (e) UR-DGN [16]. (f) Baseline - L^{SR}. (g) AACNN - L^{SR}. (h) AACNN-$L^{SR} + L^{adv}$. (i) High-resolution images. Both (g) and (h) can hallucinate specific attribute accurately in visual results. (h) is more realistic than (g) with adversarial training.

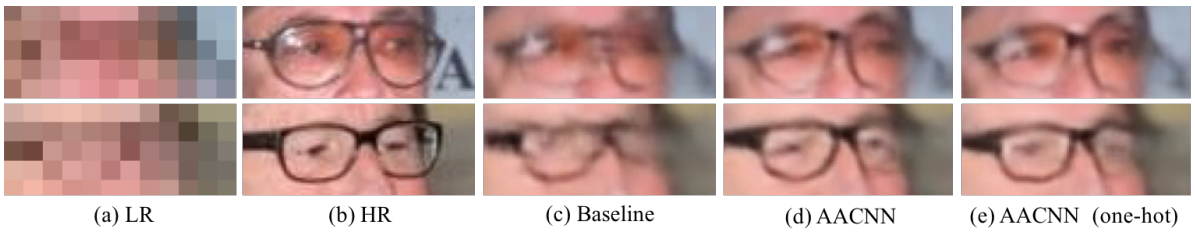


Figure 7. (a) Low-resolution inputs images. (b) High-resolution images. (c) Baseline - L^{SR}. (d) AACNN - L^{SR} with all attributes are known. (e) AACNN - L^{SR} with one-hot attribute input (only eyeglasses is known). From visual results, our method can significantly recover the target attribute with specific one-hot attribute vector (eyeglasses), and the recovery effect is close to AACNN with all attribute known input.

known attributes into the unknown one and train a model by attribute vectors with each only 50% information known. Finally, we test the model with different known proportion of attribute vectors. In Table 6, we do this experiment on global and local evaluation (i.e. eyeglasses and goatee).

In the all-attribute-known situation, If testing on the model which train with 50% known attributes, we can still have great performance on global and local evaluation as shown in the first two column of Table 6.

In the partial-attribute-known situation, we can still have great performance (as shown in the last two column of Table 6) by using the model which train with 50% known attributes.

In Fig. 7, we further use class specific one-hot attribute vector (eyeglasses) to test on the model which train with 50% known attributes. From the visual results, our method can significantly recover the target attribute, and the effect is close to AACNN with all attribute known input. As a result, AACNN still improves the performance of hallucination adaptively, even if we only know partial attribute input.

5. Conclusions

In face hallucination, most of previous methods cannot accurately hallucinate local attributes or accessories in ultra-low-resolution. We propose a novel Attribute Augmented Convolutional Neural Network (AACNN) to assist face hallucination by exploiting facial attributes. More specifically, our method fuses the advantages of both image domain and attribute domain and achieves superior visual quality than other state-of-the-art methods. In addition, our AACNN still improves the performance of hallucination adaptively with partial attribute input.

6. Acknowledgement

This work was supported in part by MediaTek Inc and the Ministry of Science and Technology, Taiwan, under Grant MOST 107-2634-F-002-007. We also benefit from the grants from NVIDIA and the NVIDIA DGX-1 AI Supercomputer.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] S. Baker and T. Kanade. Hallucinating faces. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 83–88. IEEE, 2000.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[5] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[6] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. *arXiv preprint arXiv:1704.03915*, 2017.

[7] Y. Li, C. Cai, G. Qiu, and K.-M. Lam. Face hallucination based on sparse local-pixel structure. *Pattern Recognition*, 47(3):1261–1270, 2014.

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

[9] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

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

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

[12] M. F. Tappen and C. Liu. A bayesian approach to alignment-based image hallucination. In *European Conference on Computer Vision*, pages 236–249. Springer, 2012.

[13] 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.

[14] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[15] C.-Y. Yang, S. Liu, and M.-H. Yang. Structured face hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1099–1106, 2013.

[16] X. Yu and F. Porikli. Ultra-resolving face images by discriminative generative networks. In *European Conference on Computer Vision*, pages 318–333. Springer, 2016.

[17] X. Yu and F. Porikli. Face hallucination with tiny unaligned images by transformative discriminative neural networks. In *AAAI*, pages 4327–4333, 2017.

[18] X. Yu and F. Porikli. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3760–3768, 2017.

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

[20] S. Zhu, S. Liu, C. C. Loy, and X. Tang. Deep cascaded bi-network for face hallucination. In *European Conference on Computer Vision*, pages 614–630. Springer, 2016.