Stochastic Attribute Modeling
for Face Super-Resolution

Hanbyel Cho¹, Yekang Lee¹, Jaemyung Yu¹, and Junmo Kim¹

School of Electrical Engineering, KAIST, South Korea
{tlr14658, askhow, jaemyung, junmo.kim}@kaist.ac.kr

Abstract. When a high-resolution (HR) image is degraded into a low-resolution (LR) image, the image loses some of the existing information. Consequently, multiple HR images can correspond to the LR image. Most of the existing methods do not consider the uncertainty caused by the stochastic attribute, which can only be probabilistically inferred. Therefore, the predicted HR images are often blurry because the network tries to reflect all possibilities in a single output image. To overcome this limitation, this paper proposes a novel face super-resolution (SR) scheme to take into the uncertainty by stochastic modeling. Specifically, the information in LR images is separately encoded into deterministic and stochastic attributes. Furthermore, an Input Conditional Attribute Predictor is proposed and separately trained to predict the partially alive stochastic attributes from only the LR images. Extensive evaluation shows that the proposed method successfully reduce the uncertainty in the learning process and outperforms the existing state-of-the-art approaches.

Keywords: Face Super-Resolution, Stochastic Attributes, Uncertainty

1 Introduction

Face super-resolution (SR), also known as face hallucination, is a task in which an input low-resolution (LR) face image is restored to a high-resolution (HR) face image. Face SR is commonly used as a tool to preprocess the degraded input images for other face-related tasks, such as face verification [15,19,16] and face detection [23,7], as a high performance can be attained when high-quality input images are used.

In recent years, there have been significant achievements in deep neural network (DNN) based SR methods. SRCNN [3] firstly introduced convolutional neural network (CNN) based SR method, which applied patch extraction and non-linear mapping on the SR processing. VDSR [9] proposed residual architecture so that the network only learns the difference between LR and HR images, and also utilized gradient clipping for fast training. As a special case of the general SR, there exists face related prior knowledge in face images, which can be major clue for face SR problem. To utilize such rich information, GLN [20] extracted the global and local spatial information from an LR image, and the HR
image was restored using both of information. Wavelet-SRNet [8] did the same thing in frequency domain. Furthermore, to improve the performance of face SR, some researchers [1,2] utilized various facial priors, which involve semantic meanings such as facial landmarks [4,25] and parsing maps [18,12].

Despite these efforts, the face SR approaches encounter a major challenge of the restored images still being blurry. As shown in Figure 2, the existing approaches [3,9,11] lead to blurry results in the distinguishing parts of a person such as the eyebrows, because several possible HR images can be inferred from the LR image.

In the face domain, the attributes (such as the shape of the eyebrows) that distinguish the identity of a person can be considered as key features that a network should restore. Such attributes can be classified into two main categories, namely, deterministic and stochastic attributes, depending on the degree of information loss, as shown in Figure 1.

The deterministic attributes are the attributes of an HR image, which can be clearly determined from the LR even if the image is degraded. These attributes do not contribute to the blurriness of the output because the HR corresponding to the LR has a one-to-one correlation when using a conventional face SR framework. In contrast, the stochastic attributes refer to the attributes that have partially or completely lost their information and thus are subject to probabilistic prediction from the LR image. Depending on the degree of information loss, the stochastic attributes can be divided into two subcategories. In the partially alive case shown in Figure 1, the attributes such as the eyebrow shape can be inferred probabilistically in the form of a Gaussian from the LR image. In the fully lost case, the probability of the attributes can be assumed to follow a uniform distribution because the attributes such as a wrinkle on a cheek cannot be inferred.

Specifically, for the stochastic attributes, several HR images exist that correspond to the LR image, and these aspects are accepted by the network as the uncertainty. In conventional face SR, there was no means of removing this
uncertainty in the decoding phase which makes the final result image from the extracted information. The network thus reflects all the possibilities in the single output image, resulting in blurry shapes such as the eyebrows, as shown in Figure 2.

To overcome this limitation, this paper proposes a novel face SR framework that can remove the blur effect caused by the uncertainty from the stochastic attributes. First, the information of the LR image is separated into deterministic and stochastic attributes by using the error encoding framework [6]. The extracted stochastic attributes are encoded into the latent space in the form of a multivariate Gaussian distribution by using a Residual Encoding Network (REN). The stochastic attributes can be selected by sampling from the distribution and later added to a middle feature map of the face SR network. When we get the stochastic attributes in training phase, it is necessary to utilize the HR image in the proposed framework. However, it is not possible to show the HR image in the inference phase. To solve this problem, an Input Conditional Attribute Predictor (ICAP) is proposed, which requires only an LR image as the input to infer the stochastic attributes. This network is forced to imitate the output values of the REN so that the network can predict the partially alive stochastic attributes from the LR image. We conduct experiments on UTKFace and CelebA datasets using the proposed method, and as a result, demonstrate the superiority of the method.

The main contributions of our work are three-fold:

– To the best of our knowledge, this is the first face SR framework that can eliminate the blurry effect caused by the uncertainty from the stochastic attributes.
– The deterministic and stochastic attributes is separated from the LR face images in unsupervised manner.
– Using the proposed ICAP, the partially alive stochastic attributes can be inferred by considering only the LR input, thereby eliminating the dependence of the framework on the ground-truth HR image.

2 Related Works

In the face SR domain, there are many prior works, which tries to make a clear SR output. The recent approaches can be divided into two categories. The first one involves using the the facial prior knowledge in the learning process. In an
existing study, Super-FAN [1] proposed the framework that addresses face SR and alignment simultaneously to improve the SR performance. In the FSRNet [2], the facial boundary was attempted to be clarified by using a geometry prior such as parsing maps and facial landmarks. In the RAAN [21], both the pixel level and semantic level facial priors were utilized. Most of these methods can be interpreted as an effort to reduce the width of the Gaussian distribution illustrated in Figure 1. Such techniques can effectively improve the results; however, the outputs are still blurry because the stochastic attributes still remain in the Gaussian form.

The second approach utilizes adversarial learning framework, inspired by Generative Adversarial Network (GAN) [5]. It is a consensus that GAN-based models can generate photo realistic images. There are a lot of papers such as SRGAN [11] which utilize adversarial loss with pixel-wise loss in general SR problem. Furthermore, in the face SR domain, the approaches involving UR-DGN [22], FSRGAN [2], and RAAN-GAN [21] adopted the adversarial training method, yielding visually plausible results. However, quantitatively, all these approaches suffer from a low peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). This means that the resulting image is considerably different from the correct HR image. This is because GAN-based methods attempt to create the output images that have the same distribution as that of the HR images, and the focus is not on restoring the correct HR corresponding to the LR image. In the general SR domain [17,10,13,24], which yields just visually plausible results is sufficient. However, the finding is not consistent with the goal of the face SR, that a person’s identity should be preserved. Therefore, in this work, we attempt to produce a clear output without using the GAN-based method of changing a person’s identity.

3 Method

The objective here is to reduce the blurry effect caused by stochastic attributes in the training phase. The proposed approach can be divided into two stages. First, the information contained in an LR image is separated into deterministic and stochastic attributes by utilizing the Error Encoding Network [6]. The stochastic attributes from the residual image is encoded to a low-dimensional latent space in the form of a multivariate Gaussian distribution. Finally, we train another network that can predict the partially alive stochastic attributes directly from the LR image without using the ground-truth HR image in the inference.

3.1 Network Architecture

As shown in Figure 3, the proposed networks mainly consist of three parts: SR Encoder-Decoder, Residual Encoding Network, and Input Conditional Attribute Predictor. We now present each part of the network as step by step. For training phase, we use a given training set \( \{x^{(i)}, y^{(i)}\}_{i=1}^{N} \), where \( N \) is the number of training images, \( x^{(i)} \) is the LR image, and \( y^{(i)} \) is the ground-truth HR image corresponding to \( x^{(i)} \).
Fig. 3. Network structure of the proposed method.

**SR Encoder-Decoder** The SR Encoder-Decoder, which is denoted as $f_\theta$ with parameters $\theta$, is the main face SR network in the proposed framework. This makes a $128 \times 128$ SR image $\tilde{y}$ from a given $128 \times 128$ bicubic upsampled LR image $x$. The network is composed of the encoder network $f_{\text{enc}}$ that encodes the information of the LR image, and the decoder network $f_{\text{dec}}$ that creates a final output image from the encoded feature map. Additionally, there is a deconvolution layer to accept the external information, which comes from a latent vector $z$, and adds this to the output feature map of the encoder. The deconvolutional layer allows the SR Encoder-Decoder to create an output with a more detailed facial representation.

The SR Encoder-Decoder can produce outputs of different quality according to the sampled $z$ vectors. If a latent vector $z$ is set as a zero vector $z_0$, the network generates a blurry output as a predicted HR output $\tilde{y}$ without the external information. In this case, the network infers the HR target $y$ as much as possible without using the external information.

Consequently, the information mainly used in an LR image is related to the deterministic attributes, which can be clearly predicted the HR’s attributes from an LR image. This output image known as $\tilde{y}_d$ is obtained using the following equation:

$$\tilde{y}_d = f_\theta(x, z = z_0). \quad (1)$$

If a latent vector $z$ is set as $z_{\text{res}}$, which includes the external information from the Residual Encoding Network, the network generates a high-quality output as the predicted HR output $y$ by using the information obtained from both LR input and $z_{\text{res}}$. In this case, the network generates the output by using the latent vector $z_{\text{res}}$, which includes the stochastic attributes. The output image in this case, $\tilde{y}_s$, is obtained using the following equation:

$$\tilde{y}_s = f_\theta(x, z = z_{\text{res}}). \quad (2)$$
We call each process of generating $\hat{y}_d$ and $\hat{y}_s$ as a deterministic path and stochastic path, respectively. First, a low-quality image $\hat{y}_d$ is obtained through the deterministic path without reference to any external information. Next, a high-quality image $\hat{y}_s$ is created from an input LR image, referring to $z_{res}$ obtained from the residual between $y$ and $\hat{y}_d$.

**Residual Encoding Network** The Residual Encoding Network (REN), which is denoted as $g_\phi$ with parameters $\phi$, is used to encode the stochastic attributes into the low-dimensional latent space. The network uses the residual image $r$, which is obtained using the following equation:

$$r = y - \hat{y}_d.$$  

The residual of the blurry image $\hat{y}_d$, which is a result of the failure to perform the restoration in the deterministic path, is considered as the stochastic attributes. When the network encodes the stochastic attributes into latent space, the encoding result is in the form of a multivariate Gaussian distribution parameterized by the mean $\mu_{res}$ and standard deviation $\sigma_{res}$. We sample from the posterior $z_{res} \sim g_\phi(\cdot|r)$ using $z_{res} = \mu_{res} + \sigma_{res} \odot \epsilon$, where $\epsilon$ is drawn from $\mathcal{N}(0, I)$ and $\odot$ denotes the element-wise multiplication.

Consequently, the closest attribute to that of the HR image is the mean vector of the distribution, and the other attributes that are different from those of the HR image have a lower probability. Furthermore, by selecting only one of the possible attributes through sampling, the blurry effect in the output images, which arises because of the uncertainty caused by multiple possibilities, can be eliminated.

We train the networks described so far as follows: First, the SR encoder-decoder is trained to predict the HR as much as possible from the LR with the latent vector $z$ set to zero to ensure that the deterministic information can be fully used without any additional information. The SR Encoder-Decoder is trained using the pixel-wise Mean Squared Error (MSE) function as follows:

$$L_d(\theta) = \frac{1}{N} \sum_{i=1}^{N} \|y^{(i)} - f_\theta(x^{(i)}, z = z_0)\|_2^2.$$  

(4)

To ensure that the REN successfully encodes the stochastic information and the SR Encoder-Decoder reflects the encoded vector into the predicted output image, the pixel-wise L1 loss function was adopted:

$$L_s(\theta, \phi) = \frac{1}{N} \sum_{i=1}^{N} \|y^{(i)} - f_\theta(x^{(i)}, z = z_{res})\|_1.$$  

(5)

In this case, we adopt the L1 loss to better preserve the stochastic attributes by imposing stronger constraints. Furthermore, the parameters of the SR Encoder-Decoder for the deterministic path and stochastic path is shared and trained jointly using the following loss function:

$$L_{\text{phase1}}(\theta, \phi) = L_d(\theta) + \lambda \cdot L_s(\theta, \phi)$$  

(6)
We constrain the values of $\mu_{\text{pred}}$ and $\sigma_{\text{pred}}^2$ to the specific range instead of making the distribution of $z_{\text{res}}$ similar to prior distribution due to the posterior collapse.

**Input Conditional Attribute Predictor** We divided the information contained in an LR image into deterministic and stochastic attributes and successfully reflected the sampled latent vector in an SR output. However, the HR images must be used to obtain the stochastic attributes in the current framework. This is not consistent with our goal to obtain a clear predicted output image using only a single LR image. Therefore, we propose the Input Conditional Attribute Predictor (ICAP), which is denoted as $q_\omega$ with parameters $\omega$. This predictor can be used to immediately predict the partially alive stochastic attributes from the LR image in the inference phase. The ICAP network outputs the stochastic attributes $z_{\text{pred}}$ reparameterized by $\mu_{\text{pred}}$ and $\sigma_{\text{pred}}$ similar to the REN. However, because this network receives only an LR image as the input, the network is trained using the loss function, as follows:

$$L_{\text{phase2}}(\omega) = D_{\text{KL}}(q_\omega(z_{\text{pred}}|x) \| g_\phi(z_{\text{res}}|r)).$$  \hspace{1cm} (7)

Consequently, the network can infer the characteristics of HR from the LR image by making the output distributions of the ICAP and REN closer.

### 3.2 Training Procedure

This section describes the whole training procedure of the proposed method. As shown in Algorithm 1, the training procedure can be divided into two phases: phase1 and phase2. In phase1, both the SR Encoder-Decoder and REN are trained by directly referring to the ground-truth HR images when generating the predicted HR images. In this process, we update the parameters $\theta$ and $\phi$ with the loss function described in Equation 6 until the networks are sufficiently converged. In phase2, the ICAP is trained using the ground-truth $\mu_{\text{res}}$ and $\sigma_{\text{res}}$ obtained from the phase1. Subsequently, the parameters $\omega$ of the ICAP is updated using the KL divergence loss function described in Equation 7 to ensure that the output distribution of the ICAP is similar to the ground-truth distribution. While training the ICAP, the parameters $\theta$ and $\phi$ of the SR Encoder-Decoder and REN are frozen.

### 4 Experiments

Extensive experiments were conducted to verify the effectiveness of the proposed method. In this section, first, we explain the implementation details of the conducted experiments. To demonstrate the quantitative and qualitative superiority of the proposed method compared to the state-of-the-art approaches, two kinds of experiments were performed, and the results are described herein. Furthermore, ablation studies were performed to validate the network performance in terms of the stochastic attribute modeling.
Algorithm 1 Training Procedure of the Proposed Method

**Data:** Given training set \( \{x^{(i)}, y^{(i)}\}_{i=1}^{N} \), where \( N \) is the number of training images, \( x^{(i)} \) is the LR image, and \( y^{(i)} \) is the ground-truth HR image corresponding to \( x^{(i)} \).

**Initialize:** All the settings include the learning rates, optimization method, training iteration, etc.

**Output:** Proposed Network

\[
\text{while not converged do}
\]

\[
\text{Update the parameter sets } \theta \text{ and } \phi \leftarrow \text{minimize loss } L_{\text{phase1}}
\]

\[
\text{end while}
\]

\[
\text{while not converged do}
\]

\[
\text{Update the parameter set } \omega \leftarrow \text{minimize loss } L_{\text{phase2}}
\]

\[
\text{end while}
\]

Target(T) Input(Bicubic) Deterministic(D) Residual(T-D) Stochastic(S) Residual(T-S)

![Visual result of residual encoding.](image)

**4.1 Implementation Details**

Experiments were conducted on two datasets: UTKFace [26] and, CelebA [14]. For training, we use the randomly chosen 23,408 images for UTKFace dataset and 28,970 images for CelebA dataset without any augmentation methods. For evaluation, 300 images are randomly chosen from each dataset. The proposed model is trained using the Adam optimizer with an initial learning rate of \( 2.5 \times 10^{-4} \) for phase1 and 0.005 for phase2. We constrain the values of \( \mu_{\text{pred}} \) and log of \( \sigma_{\text{pred}}^2 \) to the range \([-1, 1]\] to make a compressed distribution in the latent space for the training in phase1. All the input LR images are enlarged by bicubic upsampling before using because the proposed model has a fixed input size 128 \( \times \) 128, same as that of the output.

**4.2 Residual Encoding Results**

In this section, before validating the final results of generating the HR image using only an LR image, we ensure that the network successfully encodes the stochastic attributes and reflects the value in the final results. To this end, we generate the SR results using both the LR and HR images by employing both the SR Encoder-Decoder and REN. Figure 4 shows the visual result of the residual encoding. Target(T) and Input(Bicubic) denote the ground-truth HR image and degraded LR image that the network uses to make the SR output, respectively. Deterministic(D) is the result of using only the LR image with the SR Encoder-Decoder with \( z = 0 \). In this case, the network generates the result
Table 1. Quantitative comparisons with PSNR/SSIM. The red/blue colors indicate the highest/second-highest performance.

| Scale factor | Bicubic | SRCNN | VDSR | SRResNet | URDGN | Wavelet-SRNet | Ours  |
|--------------|---------|-------|------|----------|-------|---------------|-------|
| 4            | 31.69/0.91 | 33.44/0.99 | 31.48/0.96 | 30.79/0.96 | 31.88/0.91 | 33.52/0.94 |
| 8            | 25.21/0.76  | 26.73/0.81 | 26.51/0.80 | 25.52/0.77 | 26.66/0.82 | 27.86/0.85 |
| 16           | 21.57/0.65  | 22.73/0.76 | 22.82/0.66 | 22.53/0.66 | 23.44/0.74 |

using only the deterministic information. The Residual(T-D) can be obtained using the difference between T and D determined previously, which contains the information that is difficult to predict from Input(Bicubic) under the condition of z = 0. The REN encodes this Residual(T-D) using a multivariate Gaussian distribution. Stochastic(S) is the result of reflecting the specific vector sampled from the encoded distribution. As shown in Residual(T-S), which represents the difference between T and S, we can see that S is about the same as T. These findings confirm that the proposed network can separate the stochastic information from the LR image and encode it, and this information is successfully reflected into the final results.

4.3 Comparison with State-of-the-Art Methods

In this section, we compare our final framework, which makes a predicted HR output from only a single LR image using both the SR Encoder-Decoder and the ICAP with state-of-the-art SR methods which do not use facial priors, such as the SRCNN [3], VDSR [9], SRResNet [11], URDGN [22], and Wavelet-SRNet [8].

First, a quantitative analysis is performed. Table 1 presents the quantitative results with scale factor of 4, 8, and 16 for datasets. It can be noted that the proposed method significantly outperforms the state-of-the-art methods on both the datasets and all the scale factors in terms of both the PSNR and SSIM. Additionally, Figure 5 shows that a lower resolution of the input image corresponds to a higher performance. This means that the proposed method can better restore the image as the proportion of the stochastic information increases.

Figure 6 shows the qualitative comparison of the proposed method for a scale factor of 8. It can be noted that the proposed method generates relatively sharper shapes and edges for the important attributes in a person’s face such as the eyebrow, lips, and nose. This considerably helps restoration of a person’s impression, and therefore, the proposed method can retain the identity of the individual. More qualitative results pertaining to the scale factors 4 and 8 are shown in Figures 9 and 10, respectively.
Fig. 5. **Left**: The difference of SSIM between the proposed method and two baselines on UTKFace. The baseline1 is VDSR [9], and the baseline2 is Wavelet-SRNet [8]. **Middle**: The difference of SSIM between the proposed method and two baselines on CelebA. **Right**: The difference of SSIM between the proposed method and our deterministic baseline.

Fig. 6. Qualitative comparisons for a scale factor of 8. The images shown in the top four rows are derived from CelebA, and the other images are from the UTKFace dataset.
4.4 Ablation Study

The objective in this study is to reduce the inherent uncertainty in the existing face SR networks by extracting the stochastic attributes and modeling them as a latent space in the form of a probability distribution. Although we have demonstrated the excellent performance of the proposed method, as described in the previous subsection, this alone is not sufficient to confirm that the encoded result is stochastically meaningful. Therefore, in this part, we described the evaluation of our method regarding the encoding of the stochastic information into meaningful latent spaces through various ablation studies.

Multiple Sampling Results We encode the stochastic attributes into a multivariate Gaussian distribution, assuming that the attributes closest to the HR image should be a mean vector having the highest probability in the distribution. To demonstrate the effectiveness of the proposed method, when sampling the latent vector from the distribution parameterized by $\mu_{\text{pred}}$ and $\sigma_{\text{pred}}$, the mean vector should yield a result closest to that for the HR image such that it can be inferred from the LR image. To this end, the following experiment is performed.

We predict $\mu_{\text{pred}}$ and $\sigma_{\text{pred}}$ from an LR image by using the ICAP. Next, we use $\mu_{\text{pred}}$ as the closest attribute to the ground-truth and obtain the three other latent vectors by sampling from the distribution. Finally, the final SR results, named mean, Sample 1, Sample 2, and Sample 3, are generated by adding the latent vectors to the middle feature map of SR Encoder-Decoder. The results are shown in Figure 7. It can be observed that the eyebrow shape of mean is closer to the HR image than that for the other samples far from the center of the probability distribution. Therefore, the proposed method can be considered to be effective.

Furthermore, we determine the best PSNR/SSIM performance using only the best results after sampling multiple result images from the predicted distribution. The procedure of obtaining the samples is the same as that in the previous experiments. The results of the experiment are presented in Table 2. w/o sampling in the table corresponds to the case of in which $\mu_{\text{pred}}$ is used as the latent
Table 2. Highest PSNR/SSIM performances of multiple sampling results with $\mu_{\text{pred}}$ and $\sigma_{\text{pred}}$.

| Dataset   | Scale factor | w/o sampling | w/ sampling  |
|-----------|--------------|--------------|--------------|
|           |              | mean         | n=10         | n=100        | n=1000       |
| UTKFace   | 4            | 33.5196/0.9362 | 33.5262/0.9362 | 33.5291/0.9363 | 33.5314/0.9363 |
| CelebA    | 4            | 29.2391/0.8816 | 29.2744/0.8822 | 29.2949/0.8826 | 29.3100/0.8828 |
| UTKFace   | 8            | 27.8597/0.8457 | 27.9778/0.8478 | 28.1203/0.8507 | 28.2345/0.8528 |
| CelebA    | 8            | 24.9094/0.7574 | 25.0080/0.7600 | 25.0005/0.7627 | 25.1572/0.7646 |
| UTKFace   | 16           | 23.4414/0.7384 | 23.5632/0.7433 | 23.7190/0.7486 | 23.8462/0.7530 |
| CelebA    | 16           | 21.3295/0.6339 | 21.4404/0.6394 | 21.5698/0.6454 | 21.6533/0.6496 |

vector, and w/ sampling is the case in which the distribution parameterized by $\mu_{\text{pred}}$ and $\sigma_{\text{pred}}$ is sampled. $n$ refers to the number of samplings. The results show that a higher number of samples correspond to a higher PSNR/SSIM for each setting of the dataset and scale factor.

In Section 1, we classified the stochastic information contained in the LR images as partially alive and fully lost attributes. Although the ICAP can predict the stochastic attributes, this approach is only applicable to the partially alive case. The fully lost information can still not be determined. Therefore, the results presented in Table 2 can be interpreted as follows: When multiple sampling is performed, the fully lost attributes are sometimes accidentally matched while maintaining the prediction of the partially alive attributes, which is reflected in the performance increase.

Traversing Latent Space If the encoded attributes exist in a meaningful latent space, significant and continuous changes in the attributes must occur when sliding the latent vector. To confirm this, the latent traversal experiment was performed.

We first obtain $\mu_{\text{pred}}$ and $\sigma_{\text{pred}}$ from an LR image using the ICAP as in previous experiments. Then, two latent vectors are obtained from the distribution by sampling. These two vectors denote the start and end points in the latent space, denoted as $z_{\text{start}}$ and $z_{\text{end}}$, respectively. And then, the other latent vectors denoted as $z_{\text{inter}}$ are obtained by the convex combination of $z_{\text{start}}$ and $z_{\text{end}}$ following the expression:

$$z_{\text{inter}} = z_{\text{start}} + \alpha \cdot (z_{\text{end}} - z_{\text{start}}).$$  (8)

Then, we generate the final SR results using the latent vectors. The results are shown in Figure 8. When the latent vector is changed in a particular direction, the resulting image changes continuously. This aspect confirms that the encoded stochastic attributes exist in a meaningful latent space.
5 Conclusions

This paper proposes a novel method for face SR, which can remove the uncertainty caused by the stochastic attributes in the learning process. The key components of the proposed method are the RENs, which can encode the stochastic attributes into compressed multivariate Gaussian distribution, and the ICAP, which can predict the partially alive stochastic attributes from the LR image. The ablation study, which involves multiple sampling from the predicted attribute distribution and latent space traversal experiments, confirmed that the proposed method can successfully perform stochastic modeling for stochastic attributes. Furthermore, the results of extensive experiments performed on CelebA and UTKFace demonstrates that the proposed method outperforms the state-of-the-art approaches, both qualitatively and quantitatively. In this work, the facial priors such as facial landmarks and parsing maps were not used to focus on eliminating the uncertainty. Therefore, in future work, we aim to combine these priors with the proposed method to achieve better performance.
Fig. 9. Qualitative results for the CelebA dataset for scale factors 4 and 8 - set1.

Fig. 10. Qualitative results for the CelebA dataset for scale factors 4 and 8 - set2.
References

1. Bulat, A., Tzimiropoulos, G.: Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 109–117 (2018)
2. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: Fsrnet: End-to-end learning face super-resolution with facial priors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2492–2501 (2018)
3. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence 38(2), 295–307 (2015)
4. Dong, X., Yan, Y., Ouyang, W., Yang, Y.: Style aggregated network for facial landmark detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 379–388 (2018)
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
6. Henaff, M., Zhao, J., LeCun, Y.: Prediction under uncertainty with error-encoding networks. arXiv preprint arXiv:1711.04994 (2017)
7. Hjelmås, E., Low, B.K.: Face detection: A survey. Computer vision and image understanding 83(3), 236–274 (2001)
8. Huang, H., He, R., Sun, Z., Tan, T.: Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1689–1697 (2017)
9. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016)
10. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Fast and accurate image super-resolution with deep laplacian pyramid networks. IEEE transactions on pattern analysis and machine intelligence 41(11), 2599–2613 (2018)
11. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
12. Li, Y., Liu, S., Yang, J., Yang, M.H.: Generative face completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3911–3919 (2017)
13. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
14. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015)
15. Masi, I., Wu, Y., Hassner, T., Natarajan, P.: Deep face recognition: A survey. In: 2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI). pp. 471–478. IEEE (2018)
16. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
17. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)
18. Smith, B.M., Zhang, L., Brandt, J., Lin, Z., Yang, J.: Exemplar-based face parsing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3484–3491 (2013)
19. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1701–1708 (2014)
20. Tuzel, O., Taguchi, Y., Hershey, J.R.: Global-local face upsampling network. arXiv preprint arXiv:1603.07235 (2016)
21. Xin, J., Wang, N., Gao, X., Li, J.: Residual attribute attention network for face image super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9054–9061 (2019)
22. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: European conference on computer vision. pp. 318–333. Springer (2016)
23. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters 23(10), 1499–1503 (2016)
24. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 286–301 (2018)
25. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: European conference on computer vision. pp. 94–108. Springer (2014)
26. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5810–5818 (2017)
6 Appendix

This supplementary material provides additional details as the following:

– Final framework of the proposed method for the inference phase.
– The network structure for each components of the proposed method.

6.1 Framework for the Inference Phase

The final framework, which makes a predicted HR output from only a single LR image in the inference phase is described as Figure 11.

The Input Conditional Attribute Predictor predicts the distribution of stochastic attributes from the LR image. Then the mean vector, which has the highest probability value in the distribution is send to the feature maps in the middle of the SR Encoder-Decoder. Finally, the SR Encoder-Decoder generates the HR prediction using both the deterministic attributes and stochastic attributes.

![Fig. 11. Framework of the proposed method for the inference phase](image-url)
6.2 Network Structure

Tables 3, 4 and 5 present the network structure of the SR Encoder-Decoder, Input Conditional Attribute Predictor and Residual Encoding Network, respectively. ‘ConvBlock’ and ‘DeconvBlock’ layers are the same as the ‘Conv-BN-ReLU’ and ‘Deconv-BN-ReLU’ proposed in the FSRNet [2].

| Layer Name       | Output Size | Structure         |
|------------------|-------------|-------------------|
| Input            | 128x128     | -                 |
| f_enc            | 64x64       | 3x3, 64, stride 2 |
| ConvBlock1       | 64x64       | [3x3, 64] x 12    |
| ResBlock1        | 64x64       | [3x3, 64] x 12    |
| Conv1            | 64x64       | 3x3, 64, stride 1 |
| ResBlock2        | 128x128     | [3x3, 64] x 3     |
| Conv1            | 128x128     | 3x3, 3, stride 1  |
| µpred            | 8x8         | 3x3, 64, stride 2 |
| σpred            | 8x8         | 3x3, 64, stride 2 |

Table 3. The network structure of the SR Encoder-Decoder. It is based on the Fine SR Encoder and the Fine SR Decoder of the FSRNet [2].

| Layer Name       | Output Size | Structure         |
|------------------|-------------|-------------------|
| Input            | 128x128     | -                 |
| g                | 64x64       | 3x3, 64, stride 2 |
| ResBlock4        | 16x16       | 3x3, stride 1     |
| Conv5            | 8x8         | 3x3, 64, stride 2 |
| BatchNorm4       | 8x8         | -                 |
| µres, σres       | 8x8         | 3x3, 32, stride 1 |
| σres             | 8x8         | 3x3, 32, stride 1 |

Table 4. The network structure of the Input Conditional Attribute Predictor. For simplicity of training, it makes log of $\sigma_{pred}^2$ instead of $\sigma_{pred}$. 

| Layer Name       | Output Size | Structure         |
|------------------|-------------|-------------------|
| Input            | 128x128     | -                 |
| g                | 8x8         | 3x3, 64, stride 2 |
| Conv6            | 8x8         | 3x3, 32, stride 1 |
| BatchNorm5       | 8x8         | -                 |
| LeakyReLU2       | 8x8         | -                 |
| Conv7            | 8x8         | 3x3, 64, stride 1 |
| BatchNorm6       | 8x8         | -                 |
| Tanh1            | 8x8         | -                 |

Table 5. The network structure of the Residual Encoding Network. For simplicity of training, it makes log of $\sigma_{res}^2$ instead of $\sigma_{res}$. 
