Component Attention Guided Face Super-Resolution Network: CAGFace

Ratheesh Kalarot
The University of Auckland
rkal018@aucklanduni.ac.nz

Tao Li
Purdue University
taoli@purdue.edu

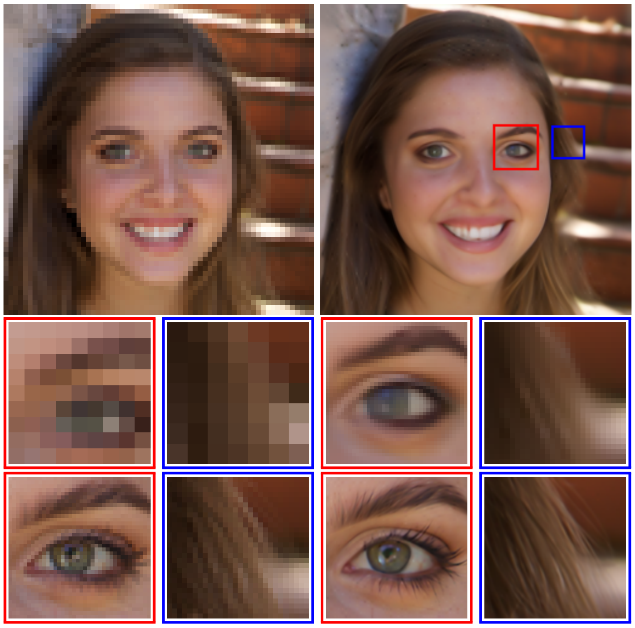
Fatih Porikli
The Australian National University
fatih.porikli@anu.edu.au

Abstract

To make the best use of the underlying structure of faces, the collective information through face datasets and the intermediate estimates during the upsampling process, here we introduce a fully convolutional multi-stage neural network for 4× super-resolution for face images. We implicitly impose facial component-wise attention maps using a segmentation network to allow our network to focus on face-inherent patterns. Each stage of our network is composed of a stem layer, a residual backbone, and spatial upsampling layers. We recurrently apply stages to reconstruct an intermediate image, and then reuse its space-to-depth converted versions to bootstrap and enhance image quality progressively. Our experiments show that our face super-resolution method achieves quantitatively superior and perceptually pleasing results in comparison to state of the art.

1. Introduction

Our brains are wonderfully attuned to perceiving faces. In addition to the visual cortex in the occipital lobe, the entire region of the brain called the fusiform gyrus is dedicated to interpreting and forming a mental representation of faces [36]. From early childhood, even very shortly after birth, human brains possess facial inference capacities and display more interest in face images than any other pattern [38]. As a species, we almost obsessively monitor and pay close attention to subtle details in paces that can allow gleaning into the origin, emotional state, internal thought process, level of engagement, and health qualities of others around us. Most of us pay more attention to faces than we do to anything other object categories. Supporting this, many gaze tracking studies show that the profile picture or avatar is the first place the eye is drawn to on social media profiles [50]. Pictures with human faces are with a large margin more likely to receive likes than the ones with no faces. It is not surprising that almost one-third of social media images are selfies and more than half are tagged with a label relates to face.



(a) Input LR image. (b) Our SR results.

Figure 1: Our method can 4× super-resolve face images of any size. Top row: 64×64 LR input and our result. Middle row: enlarged areas from the above images. Bottom row: enlarged areas when the input LR image is 256×256 (see supplementary for the whole image and its super-resolved counterpart). Please **zoom in** for the best view.

The resolution of the faces is an essential factor, and certain features appear to be found more attractive in higher resolution across individuals and cultures [29, 34]. Especially the eye and mouth regions are critical for face perception, as well as for neural responses to faces [44, 9]. Moreover, the interpretation of facial features is determined by the relative arrangement of parts within a face context [37]. Attention selection and guidance, thus, are important elements of high-resolution stimuli in the modeling of the processes in visual processing.

High-resolution face images provide crucial clues not only for human observation but also for computer analysis [12, 73]. The performance of common facial analysis techniques, such as face alignment [3] and identification [49], degrade when the resolution of a face is low. To provide a viable way to recover a high-resolution (HR) face image from its low-resolution (LR) counterpart, many face super-resolution methods [74, 67, 68, 66, 75, 5, 8] that rely on deep learning networks are proposed in recent years. Some of these methods explore direct image intensity correspondences between LR and HR faces, albeit being limited to low-resolution, e.g., 16×16, input images where the whole face to be included in the image. They can neither handle large input faces due to computational and memory requirements in training and inference times nor can they resolve fine-grained face-specific patterns. Besides, their dependency on near-frontal faces, which is prevalent in popular datasets [35, 20], restricts their usage for large pose variations causing distorted facial details. A naive idea to remedy this problem is to augment the training data with large pose variations during the training stage. However, this strategy leads to suboptimal results due to the increased variance of face data to be modeled and also potentially erroneous localization of facial landmarks, which is a difficult task in small LR images under large pose variations.

In this paper, in contrast to previous attempts that often demand and apply the whole face image through their neural layers, we adapt a patch-based face super-resolution method that can operate efficiently on large input faces. Our intuition is that, although it is challenging to detect facial landmarks of the face accurately, it is possible to estimate patch-based attention maps of facial components approximately and steer the super-resolution process with these attention maps to facilitate more natural and accurate resolution enhancement.

Our model consists of an off-line trained component network and two super-resolution stages. We first segment facial components using a neural network trained off-line. These components can be hair, skin, eyes, mouth, eyebrows, nose, ears, neck, and similar facial regions. In particular, we use three components; hair, skin, and other parts (eyes, mouth, nose, eyebrows, ears) for simplicity. We apply Gaussian smoothing to decrease the sensitivity of component segmentation errors. We multiply the input image pixel-wise with each component heatmaps to obtain heatmap-weighted components, which allows us to impose components as implicit attention priors. We stack the original image and the attention maps into a block. In the training phase, we randomly sample patches from this face-wise block where each patch includes the cropped original image and the corresponding attention maps. The random sampling generates identically sized patches and their augmented (flipped) versions. In testing, we process the LR image patch-wise and aggregate their HR estimations.

Each super-resolution stage has three main components, as shown in Fig. 2; a stem layer that blends the input patch channels, a residual backbone that applies fully convolutional blocks on low-resolution feature maps, and a spatial upsampling layer that reconstructs the high-resolution image. The residual backbone is made up of fully convolutional residual units. After a series of residual units, we embed a direct skip connection from the first feature layer to the last one to maintain the influence of the original reference image on the feature map of the last layer. Thus, our backbone is conditioned on reconstructing the residual info, which includes the missing high-resolution patterns in visual data. The residual blocks and direct skip connection also allow us to deepen the backbone, which boosts the overall representation capacity of the network and to increase the areas of the receptive fields for the higher level convolutional layers, which enables better contextual feedback. The residual backbone utilizes the low-resolution image and space-to-depth shuffled estimated high-resolution output of the previous stage, which permits transferring the initial model into progressively more complex networks in the following stages. Note that, each state is an independent network. Following the residual backbone, we apply spatial upsampling layers to reconstruct a higher-resolution image from its feature map. These layers use pixel shuffling with learned weights; therefore, we do not require deconvolutions. The residual backbone prepares the best possible feature maps, which have a large number of channels, and the spatial upsampling layers rearrange these feature maps into the high-resolution images using the learned weights of the filters of these layers.

To summarize, the contributions of this paper are:

- We introduce a patch-based, fully convolutional network for single image face super-resolution that processes patches in their original low-resolution throughout its backbone and layers then reconstructs the high-resolution output from rearranged feature maps.
- We recurrently apply the super-resolution stages to leverage on the reconstructed high-resolution outputs from the previous stage to enhance estimated high-resolution details progressively.
- As our experiments demonstrate, our method outperforms existing face super-resolution methods by a large margin without inducing perceptual artifacts.

2. Related Work

Image super-resolution aims at restoring an HR counterpart of a given an LR image as input. This task has been one of the most fundamental challenges in computer vision, and many approaches have been proposed within the last two decades including kernel interpolations [31], edge statistics [13, 46], patch-based schemes [15, 52, 22, 14, 62, 61, 21, 41], Bayesian methods [47, 26, 43], and supervised

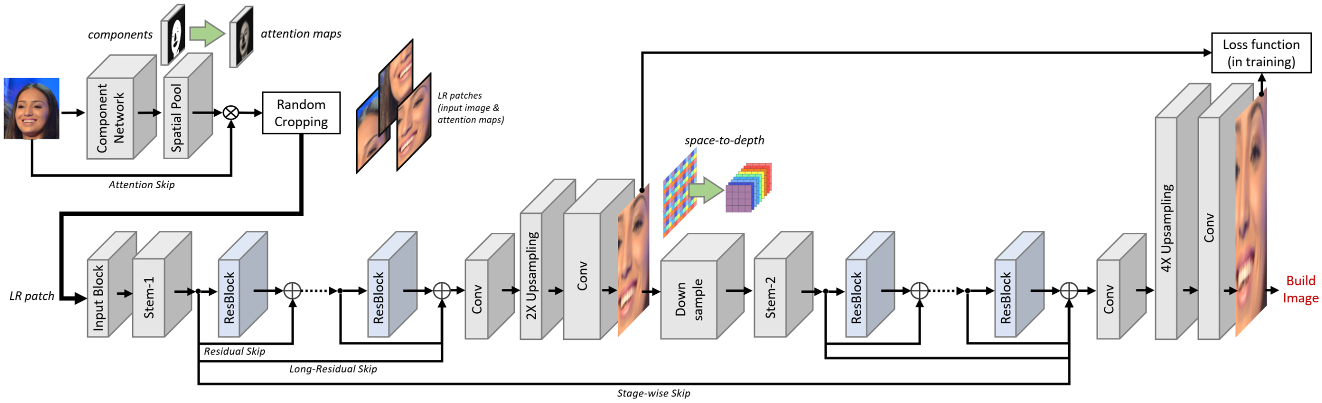


Figure 2: CAGFace architecture. First, facial components are segmented, and component-wise attention maps are generated. For training, random patches are sampled. The super-resolution network has two stages; the first stage estimates a 2× intermediate HR image. The second stage builds on the space-to-depth converted intermediate HR image and uses the original features of the first stem layer through a stage-wise skip-connection while implicitly imposing the component-wise attention.

learning [1, 40, 70]. An in-depth discussion of the available solutions can be found in recent surveys [69, 17, 63].

With the compelling advance of deep learning models, in particular, the generative adversarial networks (GAN) [16], a new wave of convolutional neural network (CNN) based image super-resolution methods have also been proposed. Most notably, SRCNN [10] and SRGAN [30] apply a CNN and a GAN, respectively, to hallucinate HR image pixels. The work in [28] progressively estimates the residual of high-frequency details using a Laplacian pyramid super-resolution network, [54] introduces the SFT-GAN for class-conditioned image super-resolution, and [55] proposes ESRGAN that leveraged a relativistic GAN [24] to estimate the distance between two images. Unlike most super-resolution models that are trained using synthetic LR images, [72] obtains LR-HR image pairs by zooming-in and -out camera lens to characterize the imaging system degradation functions. We refer readers to [64] for a comprehensive overview of deep learning-based and to [60] for canonical super-resolution approaches.

Many super-resolution methods employ facial priors to achieve higher-resolution faces. Earlier methods assume faces are in a controlled environment, and the variations are minor. For instance, the work proposed in [2] uses a spatial gradient distribution as a prior for the frontal faces. In [53], a mapping between LR and HR faces is modeled by an eigen-transform. [27] learns a nonlinear Lagrangian model for HR face images by finding the model parameters that best fit the LR image. The work in [59] incorporates face priors by mapping specific facial components (similar to our method), yet the correspondence between the components is explicitly based on landmark detection, which is difficult to obtain when the upsampling factor is large. The cascaded framework proposed by [75] super-resolves tiny faces by alternatively optimizing for face hallucination and dense correspondence field estimation. The method presented in [45] generates facial parts by CNNs and explicitly synthesizes fine-grained facial structures through part enhancement. FSRNet [6] computes facial landmark heatmaps and alignment parsing maps for end-to-end training.

The imposed image quality measure and the terms of the loss function between the reconstructed and original HR images play a critical role in super-resolution. Peak Signal-to-Noise Ratio (PSNR) is the most common metric to measure the quality [19]; however, a higher PSNR value does not necessarily imply a more visually appealing result [57]. To better simulate the human visual perception, Structural Similarity Index Measure (SSIM) [56] separates the task of similarity measurement into three components: luminance, contrast, and structure. Multi-Scale Structural Similarity Index Measure (MS-SSIM) [58] adapts to the variations of viewing conditions, and Feature Similarity Index (FSIM) [71] extends SSIM to feature space. Inception Score (IS) [42] measures the quality of generated images and their diversity, and Fréchet Inception Distance (FID) [18] extracts features from an intermediate layer of an Inception Network [48].

Accordingly, many loss functions have been proposed to train deep neural networks for super-resolution, such as pixel-wise mean squared error (MSE) [51]. While the MSE results in higher PSNR values, it often causes blur and suppresses sharp textures [57]. To overcome this, perceptual loss [23] imposes feature similarity between the super-resolved and LR images. Perceptual loss is computed over the layer right before the FC layers of VGG19 in [30], or the B1, B2, and B3 blocks of ResNet50 in [4]. A heatmap loss is proposed to preserve structural consistency between LR and HR images further [4]. Leveraging an adversarial loss from a discriminator has been shown to generate convincing results [30, 55, 39] as well.

3. Proposed Method: CAGFace

Our face super-resolution solution is composed of multiple stages. Here, we use two consecutive stages that achieve 4× super-resolution, yet our methodology can be applied recurrently for higher upsampling goals. As aforementioned, we bootstrap the super-resolution process by space-to-depth rearranging the estimated high-resolution image into multiple low-resolution channels, imposing the feature maps of the first stem layer (explained below) via a stage-wise skip connections for additional regularization, and applying a second stage network. Our patch-recurrent approach progressively bootstraps on the estimated high-resolution results, thus provides additional performance improvements.

First, we use a network that segments the facial components. We apply a layer that imposes spatial attention by multiplying the LR input image by the component heatmaps. After a random sampling of patches, we apply two stages of super-resolution networks.

We achieve 4× super-resolution with two consecutive stages of 2× resolution enhancing networks. Notice that, unlike existing methods, our method does not employ a 2× super-resolution, followed by a second 2× super-resolution on the output of the first stage. The spatial size of the input feature map to the second stage is **identical** to the size of the original LR image. We learn the most useful features for 4× super-resolution after the first stage that also reconstructs a 2× image. This mid-stage reconstruction enables us to provide an additional regularization for our loss function.

Each stage contains a separate stem layer, a collection of multiple residual blocks, and an upsampling layer followed by a final mixing layer, as illustrated in Figure 2. Please see Table 1 for the network parameters. In addition to these, the second stage has a depth-to-image conversion layer. These stages have similar kernels, yet their hyperparameters are different. Our network has conventional residual block skip connections and also a stage-wise skip connection that propagates the original features after the first stem layer to just before the final 4× upsampling layers, as well as a skip-connection over the component network that imposes attention priors on the input image. In testing, we process the LR image patch-wise and aggregate their HR estimations.

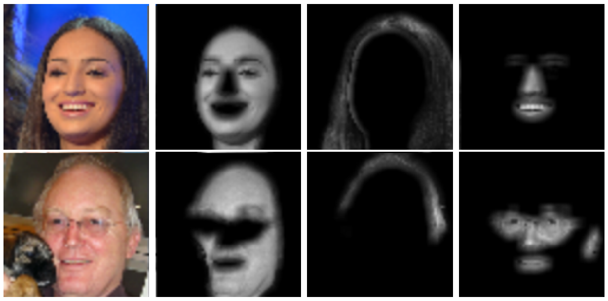

(a) Input (b) Skin (c) Hair (d) Other parts

Figure 3: Sample attention maps from component network.

3.1. Component Network

For segmentation of the facial components, we followed a similar model that to BiSeNet [65] that is developed initially for generic purpose pixel labeling such as segmentation of Cityscapes images. BiSeNet has spatial and context paths that are devised to handle the loss of spatial information and shrinkage of the receptive fields, respectively. The spatial path has three convolution layers to obtain a smaller feature map. Context path appends a global average pooling layer on the tail of the Xception network [7]. We fine-tuned this model on the CelebAMask-HQ dataset, which has 30,000 high-resolution (1024×1024) face images selected from the CelebA-HQ. Each image has a 512×512, manually-annotated, binary segmentation mask and 19 facial attributes such as skin, nose, eyes, eyebrows, ears, mouth, lip, hair, hat, eyeglass, earring, necklace, neck, and cloth.

We apply a spatial pooling layer that employs fixed Gaussian spatial kernels to suppress the segmentation errors by providing smoothing. This layer also allows higher values in attention maps to be assigned to more confidently segmented pixels. Finally, we multiply the input image with three spatially pooled components to obtain three gray-level attention maps. Sample attention maps are shown in Figure 3. We stack the original LR image and the attention maps into a block, which is our approach to administer attention to the input image. We steer the remaining of our super-resolution network using these maps as attention priors. In the training phase, we randomly sample patches from this block. Each patch, as a result, has the cropped original image and the corresponding attention maps. The random sampling generates identically sized patches and their augmented (flipped in 6 ways) versions.

3.2. Stem Layer

The layer takes a patch block as the input tensor and applies convolutional filters on it. Each depth-wise channel is a color channel of the LR image and corresponding heatmap-weighted components, which are normalized to [-

1,1] for efficient backpropagation. The stem layer in the first stage arranges the patch block in a 6-channel tensor. It then applies 256 filters, 3×3×6 each. For the following stage, we have additional channels. After the first stage, we estimated a 2× super-resolved HR image; thus, we first rearrange (space-to-depth) the pixels of an estimated HR image into 4 LR images. We then combine these LR images into a 12-channel tensor. Notice that we do not impose the heatmaps explicitly again.

Our network uses the original LR resolution frames in all its layers and stages. Since we use the same image size for all layers (except the upsampling layers), the learning becomes more efficient. Multiple references provide spatially vibrant local patterns.

3.3. Residual Backbone

The residual backbone applies fully convolutional blocks on low-resolution feature maps generated by the stem layers. It is made up of 16 fully convolutional residual units. Each residual unit has a front convolutional layer followed by a ReLU and a second convolutional layer with a skip connection from the first one. Similarly, the residual backbone has also a direct skip connection from the input to the last residual block. This skip connection allows our network to learn the missing high-resolution details by reconstructing the residual info. The structure of the residual backbone of each stage is identical. The residual blocks and the direct skip connection also permits deepening the residual backbone for each stage. This boosts the overall capacity and increases the receptive field sizes. Thus, residual backbone feature maps have better access to contextual information.

3.4. Spatial Upsampling

We apply spatial upsampling layers to reconstruct a higher-resolution image from the feature map of the residual backbone. Since we shuffle pixels and we apply a set of convolutional filters, our upsampling does not require deconvolution operations. We rearrange the comparably large number of feature map channels per pixel into a high-resolution image using the learned weights of the filters of the upsampling layers. We set the number of layers for the first stage and the second stage to 4 and 5 as the second stage feature map has to generate pixels. Each stage provides 2× super-resolution, yet it is possible to set the upsampling factor to larger ratios since there the feature maps are sufficiently deep.

Subnetwork	Kernel shape	Kernel params (bias)
Stem 1	3x3x6x256	13824 (256)
Backbone 1: 16× ResBlocks (2 layer)	3x3x256x256	9437184 (4096)
	3x3x256x256	9437184 (4096)
Spatial Upsampling 1	3x3x256x256	589824 (256)
	3x3x256x1024	2359296 (1024)
	3x3x256x3	6912 (3)
	3x3x12x256	27648 (256)
Stem 2	3x3x256x256	587520 (256)
Backbone 2: 16× ResBlocks (2 layer)	3x3x256x256	9437184 (4096)
	3x3x256x256	9437184 (4096)
Spatial Upsampling 2	3x3x256x256	589824 (256)
	3x3x512x2048	9437184 (2048)
	3x3x512x2048	9437184 (2048)
	3x3x512x3	13824 (3)
	3x3x3x3	81 (3)
Total trainable parameters in Stage 1		21881859
Total trainable parameters in Stage 2		38952791

Table 1: CAGFace network parameters.

For the goal of higher PSNR results, MSE would be the ideal loss function. However, MSE heavily penalizes the outliers. Recently, the work in [33] empirically demonstrated that the mean absolute error (MAE) works better than the MSE. In our experiments, we also made a similar observation. In particular, at the initial stages of the training, using the MSE based loss functions caused instability. However, MAE-based loss at the later epochs converges slowly. Therefore, we opted to impose the Huber loss function, which is differentiable and combines the benefits of the MAE and MSE. It is defined as

$$L_\delta(d) = \begin{cases} \frac{1}{2}d^2 & \text{for } |d| \leq \delta, \\ \delta|d| - \frac{\delta^2}{2} & \text{otherwise} \end{cases} \tag{1}$$

where

$$d = I_{HR}(x,y) - \hat{I}_{HR}(x,y) \tag{2}$$

is the pixel-wise difference between the target (ground-truth) HR image I_{HR} and the estimated HR image $\hat{I}_{HR}$. Above, we set $\delta = 1$, which is the point where the Huber loss function changes from quadratic to linear.

We trained the first stage and then the second stage by using the learned first stage parameters for initialization.

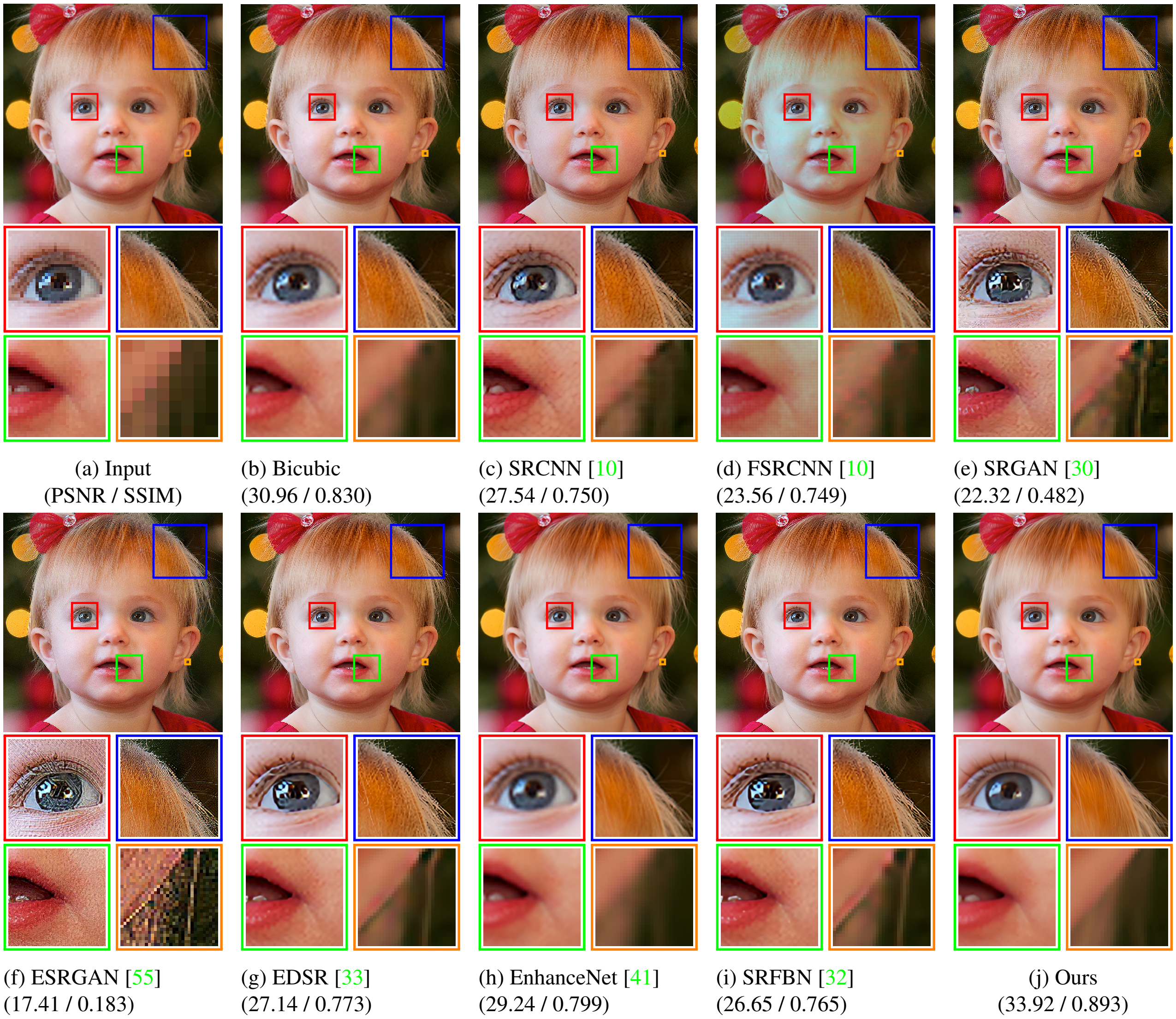
4. Experiments

4.1. Dataset

We use 1024×1024 face images from the Flickr-Faces-HQ Dataset (FFHQ) [25], which consists of 70,000 high-quality PNG images with considerable variation in terms of facial attributes such as age and ethnicity as well as image background. It also provides sufficient coverage of accessories such as eyeglasses, sunglasses, and hats. The images were crawled from Flickr. We then randomly split the FFHQ dataset into non-overlapping training, testing, and validation subsets of ratio 80%, 15%, and 5%, respectively.

4.2. Evaluation Metrics

To quantitatively measure the performance and provide comprehensive comparisons with state-of-the-art methods,



(a) Input
(PSNR / SSIM)

(b) Bicubic
(30.96 / 0.830)

(c) SRCNN [10]
(27.54 / 0.750)

(d) FSRCNN [10]
(23.56 / 0.749)

(e) SRGAN [30]
(22.32 / 0.482)

(f) ESRGAN [55]
(17.41 / 0.183)

(g) EDSR [33]
(27.14 / 0.773)

(h) EnhanceNet [41]
(29.24 / 0.799)

(i) SRFBN [32]
(26.65 / 0.765)

(j) Ours
(33.92 / 0.893)

Figure 4: Comparison with state-of-the-art methods for the **patch-based** version (output HR image is 1024×1024). As visible, our method can super-resolve without artifacts and noise-like patterns. Reconstructed images are visually pleasing and resemble the ground-truth better than the existing methods (for a better view, see in color on digital display).

we used four quality assessment metrics including PSNR, SSIM, FID, and MS-SSIM. FID [18] is defined as:

$$\text{FID} = ||\mu_r - \mu_g||^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}), \quad (3)$$

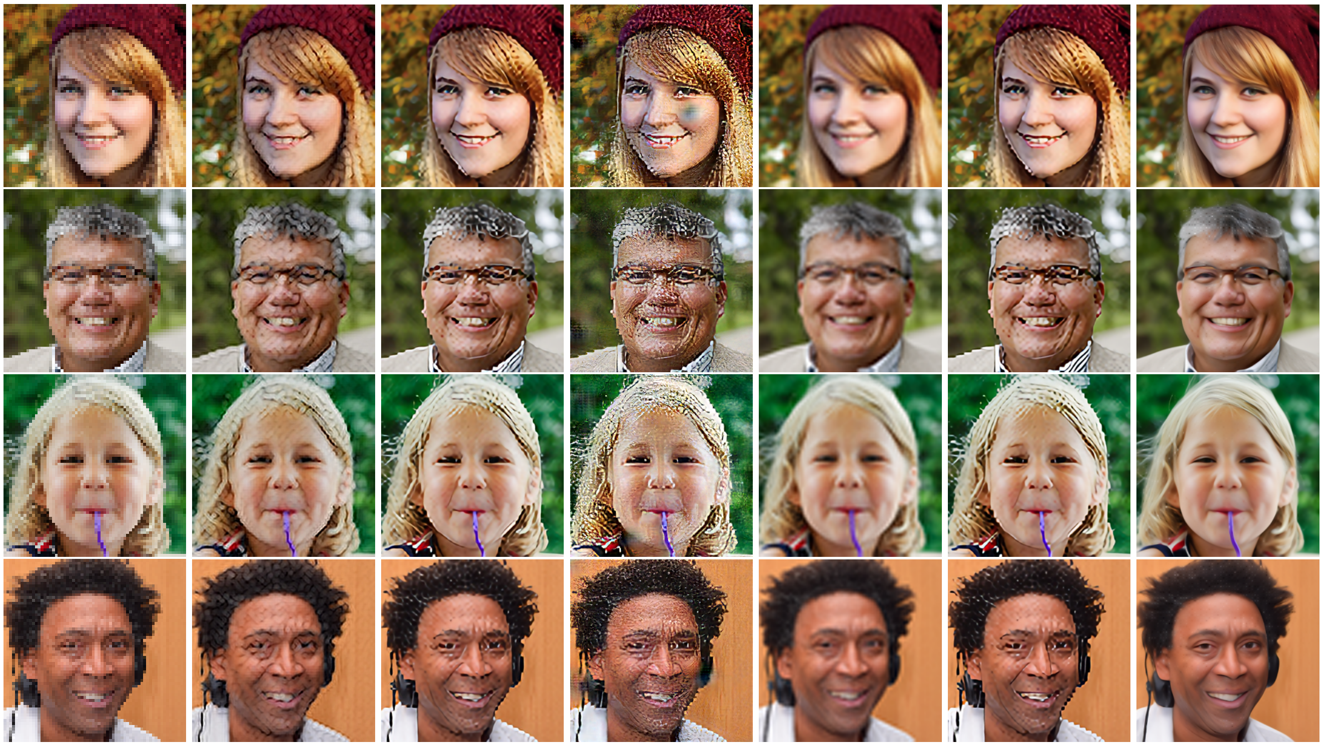
where $X_r \sim \mathcal{N}(\mu_r, \Sigma_r)$ and $X_g \sim \mathcal{N}(\mu_g, \Sigma_g)$ are activations of Inception-v3 pool3 layer for real and generated samples, respectively. Lower FIDs mean the generated results are closer to the original image, measured by the Fréchet distance between two distributions.

4.3. Comparisons

We quantitatively compare our method with seven state-of-the-art super-resolution approaches as well as with the bicubic upsampling.

In the inference (test) time, we can process the given image either by taking it as a whole (whole-face) or by patch-by-patch (patch-based). The memory limitations of GPUs set an upper bound on the input image size, in particular for the training phase. For example, memory limitations of the single GPU we used prohibited training with 1024×1024 input images. Thus, to train with and infer from such relatively large images (e.g., 1024×1024), we employ the patch-based version. For these two alternative versions, we trained separate models:

- The whole-face version uses the entire face image as an input. One can argue that using the face as a whole would provide better semantics. In our experiments,



(a) Input (PSNR / SSIM)
(b) SRCNN [10] (22.82 / 0.668)
(c) EDSR [33] (21.78 / 0.689)
(d) SRGAN [30] (17.48 / 0.420)
(e) E-Net [41] (23.08 / 0.679)
(f) SRFBN [32] (21.12 / 0.673)
(g) Ours (26.79 / 0.800)

Figure 5: Comparisons with the state-of-the-art for the **whole-face** version, i.e. training with 64×64 face images as input to generate 4× HR outputs of size 256×256. As visible, our method generates superior results for whole-face training as well.

we set the input size to 256×256 face images.

- The patch-based version uses the same network as above and identical size overlapping patches. We set the patch size to 256×256 for 1024×1024 HR outputs. We also tested 128×128 patch size. The patch-based version allows generating very large output faces without being restricted by the GPU memory.

The only difference between the above versions is the training data. For training of the patch-based version, we sampled randomly sampled around 2 million patches (48 per image) from 56,000 HR training images. For both versions, we used conventional 4× bicubic downsampling to obtain the LR input image. We augmented the training data by applying one of these six geometric transformations; rotating 90°, 180°, 270°, flipping vertically, and flipping horizontally. Even for the whole-face version, the local receptive fields do not derive semantics from the entire face. Supporting the assumption that a network would be as best as its local constituent kernels, we observed that the patch-based version generates competitive models while accessing finer granularity textures.

Figures 4 and 5 provide qualitative comparisons with state-of-the-art super-resolution techniques for 1024×1024 images (patch-based) and 256×256 images (whole-face), respectively. We used the best models available for the state-of-the-art methods provided by their authors. Our results showcase the superior quality of the proposed method. In particular, our patch-based version achieves the most pleasing HR reconstructions without any artifacts. In comparison, the GAN-based methods introduce perceptually unignorable fragmentations, remnant noise-like patterns, and broken textures. We also provide quantitative results in Tables 2 and 3, where our model outperforms the compared state-of-the-art methods with a remarkable margin under various metrics, including PSNR, SSIM, and FID. Notice that the compared methods either do not take advantage of facial semantic information or impose incorrect semantic bias. They do not use facial components to guide the super-resolution process. Most use lower resolution images in their training, which may be further limiting their representation capacity. These explain why bicubic upsampling, a deterministic approach without any semantic bias, performs better than its data-driven counterparts.

4.4. Ablation Study

We evaluated variants of our patch-wise model with different hyper-parameterizations, *i.e.*, F, the number of fea-

	PSNR	SSIM	MS-SSIM	FID
Bicubic	31.87	0.872	0.956	**10.65**
SRCNN [10]	27.40	0.801	0.924	31.84
FSRCNN [11]	24.71	0.804	0.951	23.97
EDSR [33]	28.34	0.827	0.933	15.54
SRGAN [30]	21.49	0.515	0.807	60.67
ESRGAN [55]	19.84	0.353	0.782	72.73
EnhanceNet [41]	29.42	0.832	0.934	19.07
SRFBN [32]	27.90	0.822	0.931	17.14
Ours	**34.10**	**0.906**	**0.971**	12.40

Table 2: Comparison results for 1024×1024 outputs. Our method is trained with **patches**.

	PSNR	SSIM	MS-SSIM	FID
Bicubic	25.57	0.766	0.935	135.51
SRCNN [10]	23.12	0.688	0.900	147.21
FSRCNN [11]	22.45	0.709	0.930	139.78
EDSR [33]	22.47	0.706	0.901	129.14
SRGAN [30]	17.57	0.415	0.757	156.07
ESRGAN [55]	15.43	0.267	0.747	166.36
EnhanceNet [41]	23.64	0.701	0.897	116.38
SRFBN [32]	21.96	0.693	0.895	132.59
Ours	**27.42**	**0.816**	**0.958**	**74.43**

Table 3: Comparison results for 256×256 outputs. Our method is trained with **whole-faces**.

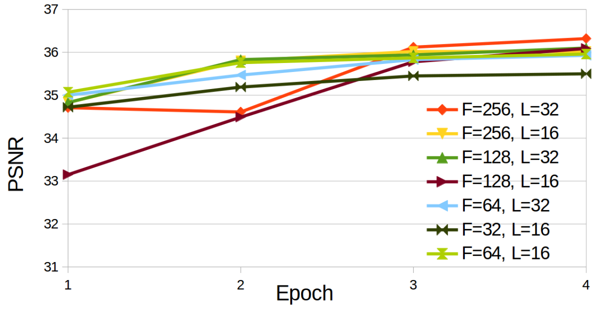


Figure 6: Effect of different network parameters on accuracy on the test dataset at initial epochs during the training phase. F is the number of features in each layer and L is the number of layers. PSNR is in dB. As visible, most versions converge to the higher PSNR scores quickly in a few epochs. This shows that our network is robust to the different hyper-parameterizations.

tures per layer, and L, the number of Resblock layers. Figure 6 demonstrates the training performance in terms of the attained PSNR scores after the initial training epochs of different configurations. As expected, with the increasing number of features and Resblock layers the performance gets better. Most of the progress towards convergence were achieved in the initial epochs when we trained the final model using an NVIDIA DGX-1. As shown in Figure 6, the network trained with different hyperparameters converges to a similar level after a few epochs, indicating that the proposed network is generally applicable regardless of the settings of F and L, thus can adapt in accordance with real world scenarios such as GPU memory limit.

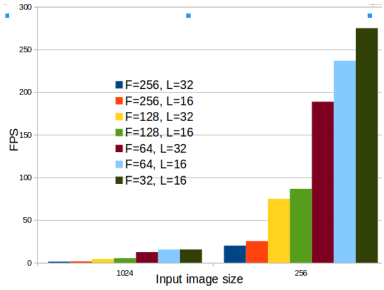


Figure 7: Effect of different network parameters on speed.

We also analyzed the performance of using a single stage 4× super-resolution instead of two-stage network. This version, even though attained better scores than the compared state-of-the-art methods, could not reach our two-stage PSNR performance: 33.71 dB (single stage) vs. 34.10 dB (two stage) for the patch-based version, and 26.46 dB (single stage) vs. 27.42 dB (two stage) for the whole-face version.

Figure 7 compares the inference speed of our model achieved for both 1024×1024 (patch-based) and 256×256 (whole-face) image resolutions on a GTX 2080Ti GPU at the 4× super-resolution setting. We can attain 270 fps on the whole 256×256 image size and 15 fps of 1024×1024 in typical settings. The whole-face processing sets an upper bound on the overall latency of the model for the parallel-processing platforms while significantly exceeding real-time speed. Our patch-based solution can reach 15 fps and substantially improves the PSNR scores. It is possible to attain real-time performance for the patch-based version by sampling the input patches from the input image with lesser degrees of overlaps.

5. Conclusion

We show that imposing attention maps implicitly and regularizing the super-resolution process using loss functions from intermediate and final upscaling stages significantly improves the performance as demonstrated in our superior results. Our patch-based method has the advantage of processing any input size image. As future work, we plan to train the entire network, including the component segmentation part in an end-to-end fashion.

References

[1] H. A. Aly and E. Dubois. Image up-sampling using total-variation regularization with a new observation model. *IEEE Transactions on Image Processing*, 14(10):1647–1659, 2005. 3

[2] S. Baker and T. Kanade. Hallucinating faces. *fg*, 2000:83–88, 2000. 3

[3] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision (ICCV)*, 2017. 2

[4] A. Bulat and G. Tzimiropoulos. 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*, pages 109–117, 2018. 4

[5] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li. Attention-aware face hallucination via deep reinforcement learning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 690–698, 2017. 2

[6] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedigns of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[7] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedigns of IEEE Conference on Computer Vision and Pattern Recognitio (CVPR)*, 2017. 4

[8] R. Dahl, M. Norouzi, and J. Shlens. Pixel recursive super resolution. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5439–5448, 2017. 2

[9] B. deHaas, D. Schwarzkopf, I. Alvarez, R. Lawson, L. Henriksson, N. Kriegeskorte, and G. Rees. Perception and processing of faces in the human brain is tuned to typical feature locations. In *The Journal of Neuroscience*, volume 36, pages 9289–9302, 2016. 1

[10] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 3, 6, 7, 8

[11] C. Dong, C. C. Loy, and X. Tang. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, pages 391–407. Springer, 2016. 8

[12] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003. 1

[13] R. Fattal. Image upsampling via imposed edge statistics. *ACM transactions on graphics (TOG)*, 26(3):95, 2007. 2

[14] G. Freedman and R. Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics (TOG)*, 30(2):12, 2011. 2

[15] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, (2):56–65, 2002. 2

[16] 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. 3

[17] K. Hayat. Multimedia super-resolution via deep learning: A survey. *Digital Signal Processing*, 81:198 – 217, 2018. 3

[18] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 3, 6

[19] A. Hore and D. Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369. IEEE, 2010. 3

[20] 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, 2007. 2

[21] J.-B. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2015. 2

[22] D. G. S. B. M. Irani. Super-resolution from a single image. In *Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan*, pages 349–356, 2009. 2

[23] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 3

[24] A. Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018. 3

[25] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 5

[26] K. I. Kim and Y. Kwon. Single-image super-resolution using sparse regression and natural image prior. *IEEE transactions on pattern analysis and machine intelligence*, 32(6):1127–1133, 2010. 2

[27] S. Kolouri and G. K. Rohde. Transport-based single frame super resolution of very low resolution face images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4876–4884, 2015. 3

[28] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017. 3

[29] J. Langlois, L. Kalakanis, A. Rubenstein, A. Larson, M. Hallamm, and M. Smoot. Maxims or myths of beauty? a meta-analytic and theoretical review. In *Psychological Bulletin*, volume 126, pages 390–423, 2000. 1

[30] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 3, 4, 6, 7, 8

[31] T. M. Lehmann, C. Gonner, and K. Spitzer. Survey: Interpolation methods in medical image processing. *IEEE transactions on medical imaging*, 18(11):1049–1075, 1999. 2

[32] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu. Feedback network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3867–3876, 2019. 6, 7, 8

[33] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017. 5, 6, 7, 8

[34] A. Little, B. Jones, and L. DeBruine. Facial attractiveness: Evolutionary based research. In *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, pages 1638–1659, 2011. 1

[35] 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. 2

[36] S. Martinez-Conde. The fascinating science behind why we see faces in objects. In *Mentalfloss*, 2018. 1

[37] D. Maurer, R. Grand, and C. Mondloch. The many faces of configural processing. In *Trends in Cognitive Science*, volume 6, pages 255–260, 2002. 1

[38] J. Morton and M. Johnson. Conspec and conlern: A two-process theory of infant face recognition. In *Psychological Review*, pages 164–181, 1991. 1

[39] S. Nah, R. Timofte, S. Baik, S. Hong, G. Moon, S. Son, and K. Mu Lee. Ntire 2019 challenge on video deblurring: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 4

[40] K. S. Ni and T. Q. Nguyen. Image superresolution using support vector regression. *IEEE Transactions on Image Processing*, 16(6):1596–1610, 2007. 3

[41] M. S. Sajjadi, B. Scholkopf, and M. Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017. 2, 6, 7, 8

[42] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016. 3

[43] Q. Shan, Z. Li, J. Jia, and C.-K. Tang. Fast image/video upsampling. In *ACM Transactions on Graphics (TOG)*, volume 27, page 153. ACM, 2008. 2

[44] M. Smith, P. Fries, F. Gosselin, R. Goebel, and P. Schyns. Inverse mapping the neuronal substrates of face categorizations. In *The Journal of Cerebral Cortex*, 2009. 1

[45] Y. Song, J. Zhang, S. He, L. Bao, and Q. Yang. Learning to hallucinate face images via component generation and enhancement. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4537–4543. AAAI Press, 2017. 3

[46] J. Sun, Z. Xu, and H.-Y. Shum. Image super-resolution using gradient profile prior. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[47] J. Sun, N.-N. Zheng, H. Tao, and H.-Y. Shum. Image hallucination with primal sketch priors. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–729. IEEE, 2003. 2

[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3

[49] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1708, 2014. 2

[50] A. Todorov and J. Porter. Misleading first impressions: Different for different facial images of the same person. In *Psychological Science*, volume 25(7), pages 1404–1417, 2014. 1

[51] J. Van Ouwerkerk. Image super-resolution survey. *Image and vision Computing*, 24(10):1039–1052, 2006. 3

[52] Q. Wang, X. Tang, and H. Shum. Patch based blind image super resolution. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 709–716. IEEE, 2005. 2

[53] 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. 3

[54] X. Wang, K. Yu, C. Dong, and C. Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 606–615, 2018. 3

[55] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 3, 4, 6, 8

[56] Z. Wang and A. C. Bovik. A universal image quality index. *IEEE signal processing letters*, 9(3):81–84, 2002. 3

[57] Z. Wang and A. C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009. 3

[58] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 3

[59] 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. 3

[60] C.-Y. Yang, C. Ma, and M.-H. Yang. Single-image super-resolution: A benchmark. In *European Conference on Computer Vision*, pages 372–386. Springer, 2014. 3

[61] J. Yang, Z. Lin, and S. Cohen. Fast image super-resolution based on in-place example regression. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, pages 1059–1066, 2013. 2

[62] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang. Coupled dictionary training for image super-resolution. *IEEE transactions on image processing*, 21(8):3467–3478, 2012. 2

[63] W. Yang, X. Zhang, Y. Tian, W. Wang, J. Xue, and Q. Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 2019. 3

[64] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 2019. 3

[65] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedigns of European Conference on Computer Vision (ECCV)*, 2018. 4

[66] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[67] X. Yu and F. Porikli. Ultra-resolving face images by discriminative generative networks. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 318–333, 2016. 2

[68] 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 (CVPR)*, pages 3760–3768, 2017. 2

[69] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang. Image super-resolution: The techniques, applications, and future. *Signal Processing*, 128:389 – 408, 2016. 3

[70] H. Zhang, J. Yang, Y. Zhang, and T. S. Huang. Non-local kernel regression for image and video restoration. In *European Conference on Computer Vision*, pages 566–579. Springer, 2010. 3

[71] 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. 3

[72] X. Zhang, Q. Chen, R. Ng, and V. Koltun. Zoom to learn, learn to zoom. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2019. 3

[73] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003. 1

[74] E. Zhou and H. Fan. Learning Face Hallucination in the Wild. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 3871–3877, 2015. 2

[75] 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. 2, 3

A. Supplementary Materials

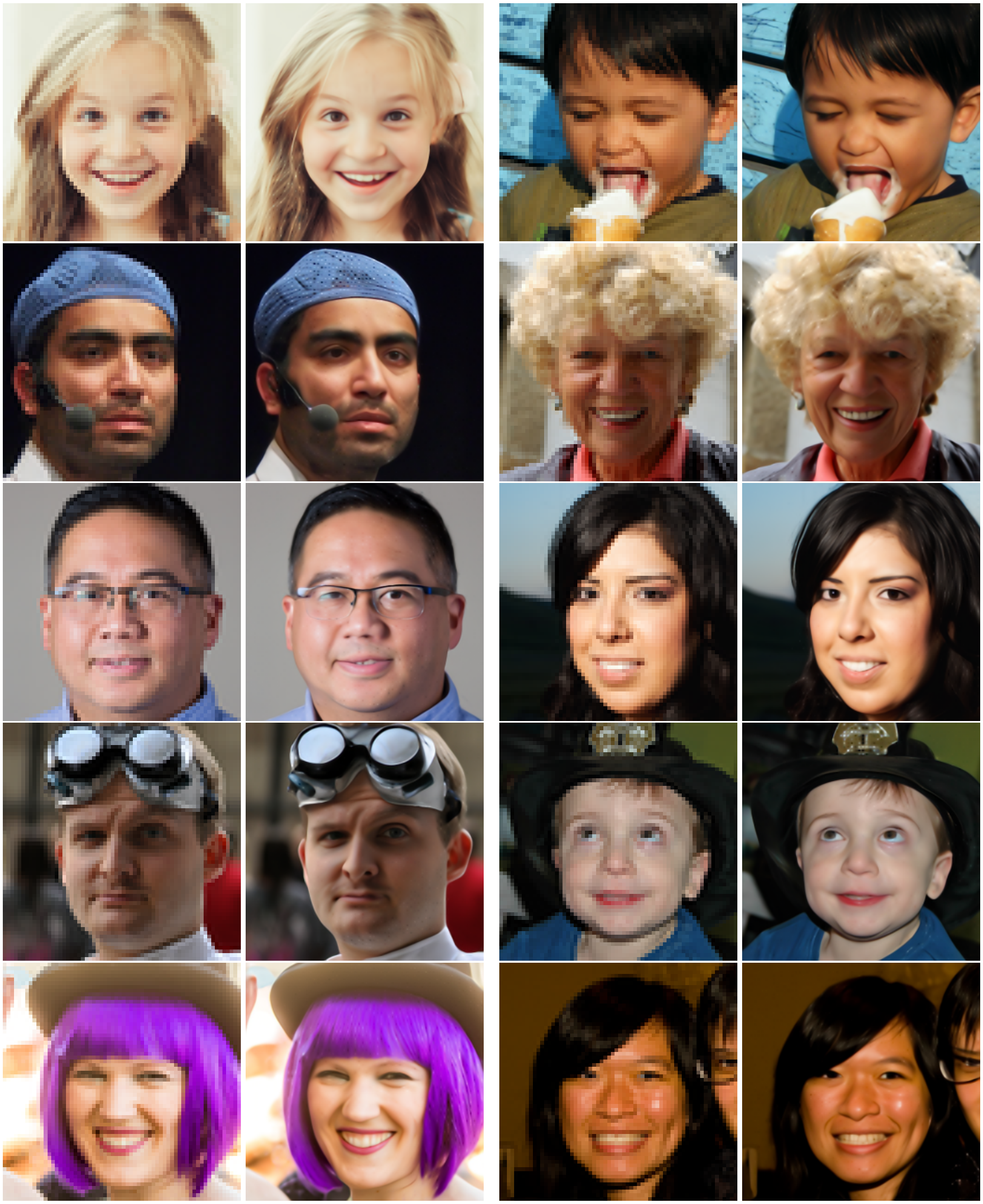
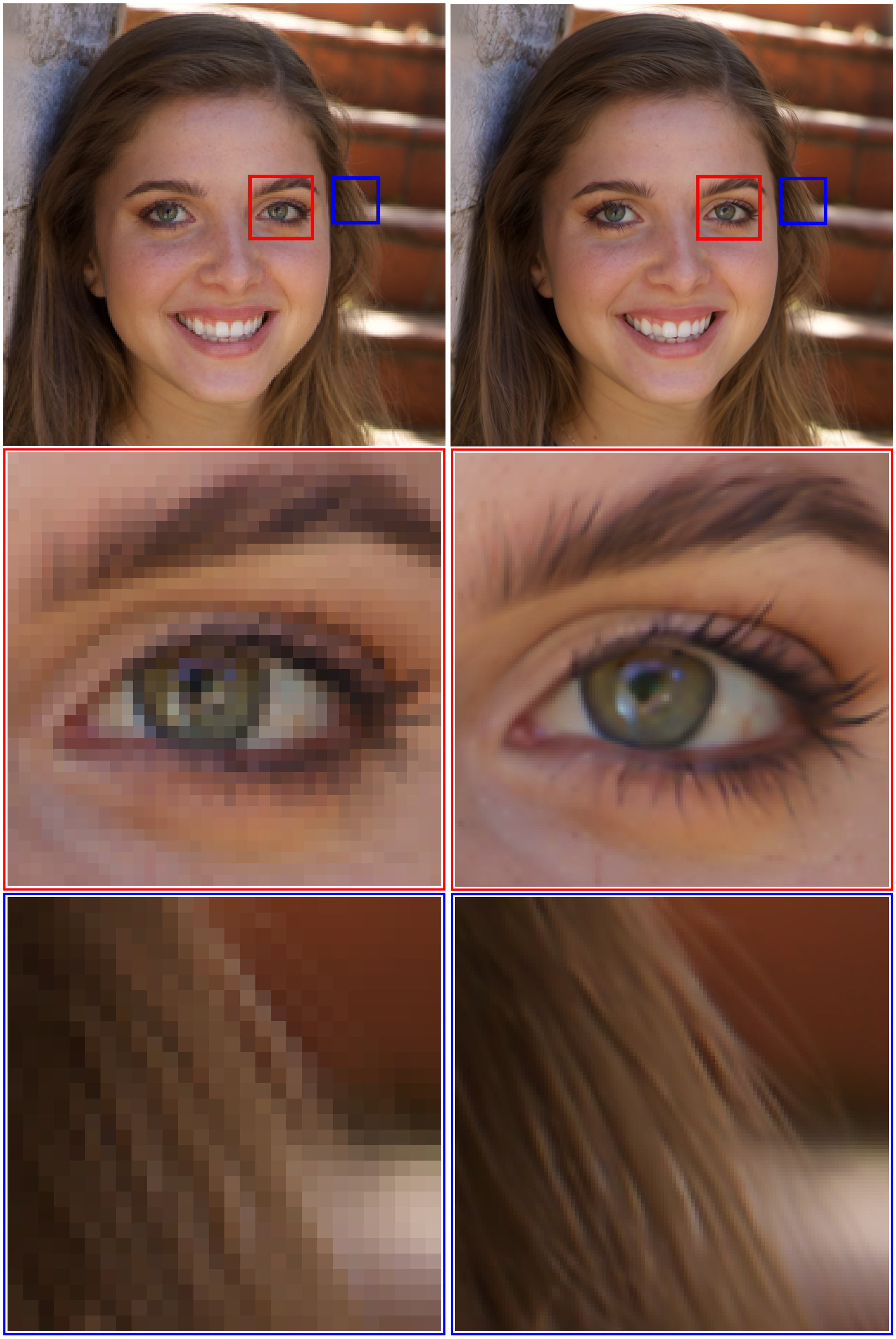


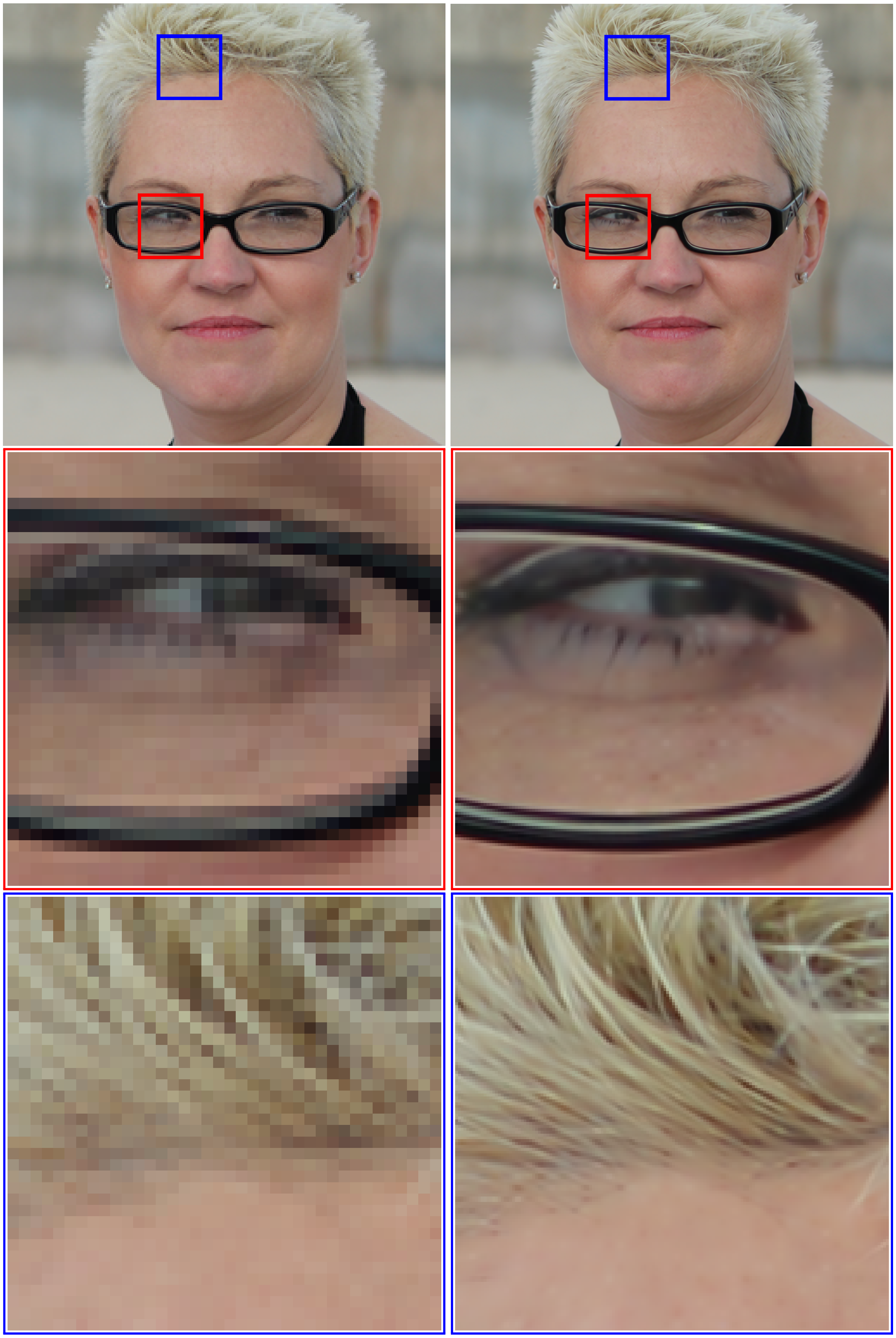
Figure 8: 4× super-resolution with our **whole-face** method. In each pair, input image is 64×64 and output image is 256×256. Please view on digital display for the best view.



(a) Input image and zoomed in regions.

(b) Our result.

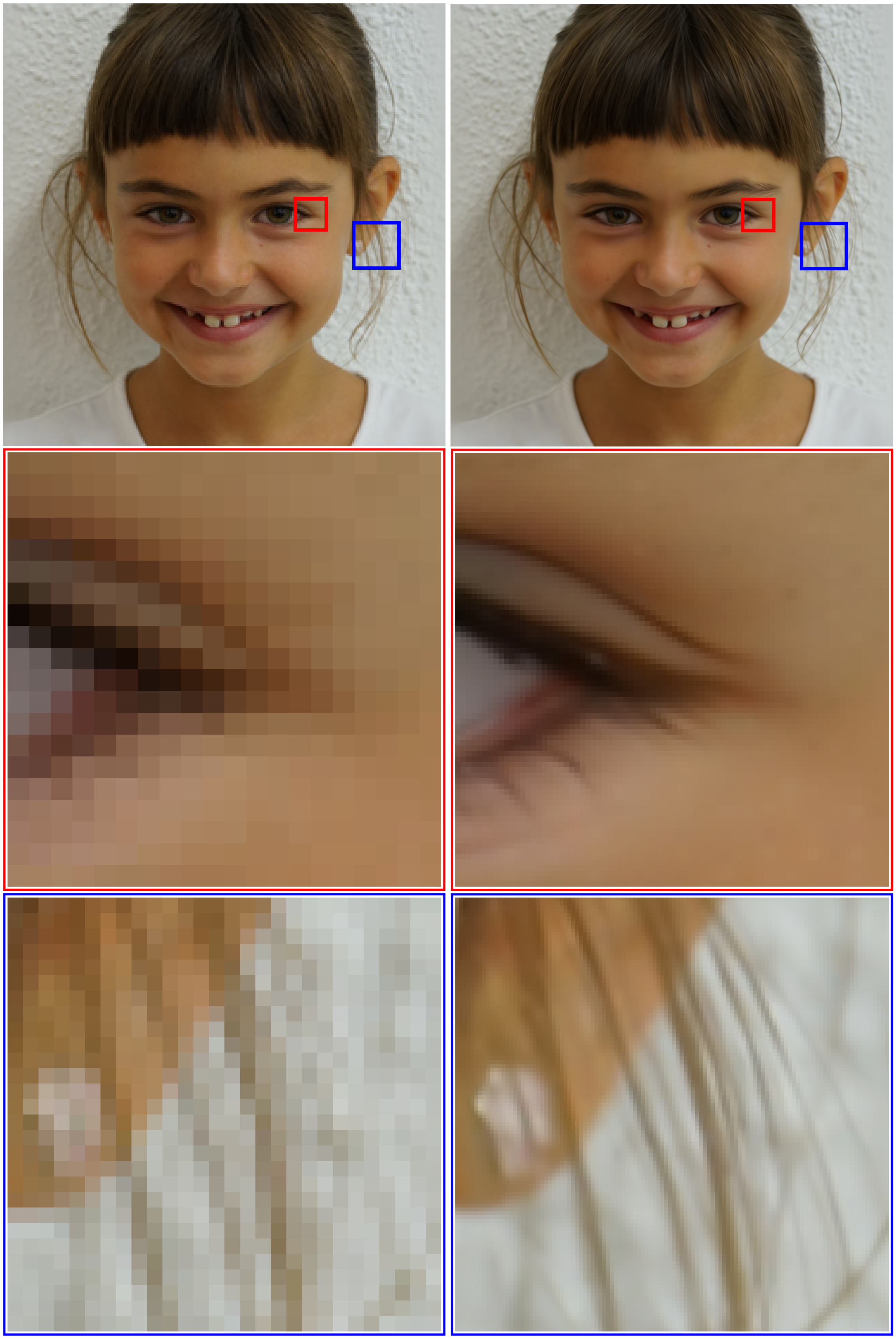
Figure 9: 4× super-resolution with our **patch-based** method. Input image is 256×256, output image is 1024×1024. Please view on digital display for the best view.



(a) Input image and zoomed in regions.

(b) Our result.

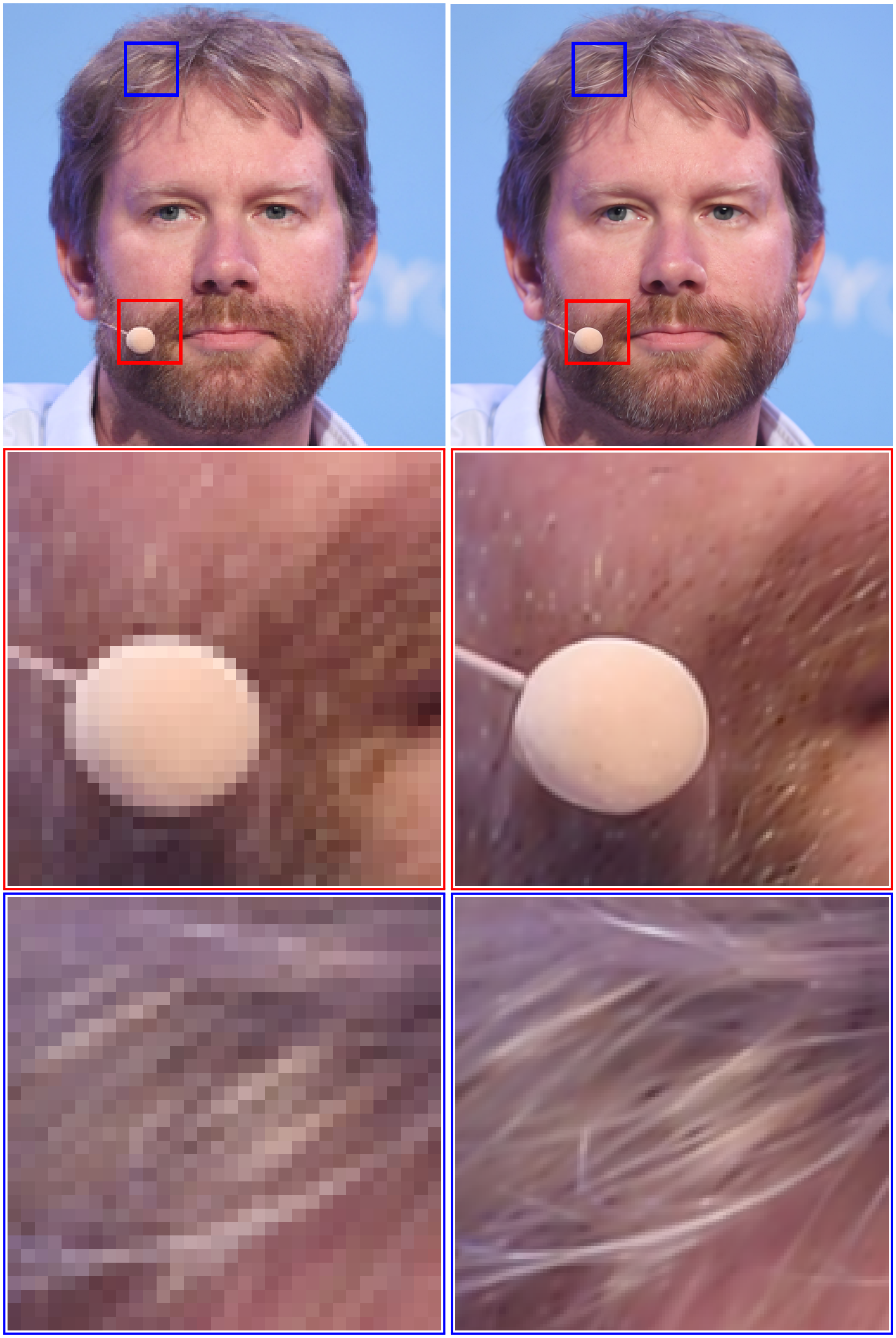
Figure 10: 4× super-resolution with our **patch-based** method. Input image is 256×256, output image is 1024×1024. Please view on digital display for the best view.



(a) Input image and zoomed in regions.

(b) Our result.

Figure 11: 4× super-resolution with our **patch-based** method. Input image is 256×256, output image is 1024×1024. Please view on digital display for the best view.



(a) Input image and zoomed in regions.

(b) Our result.

Figure 12: 4× super-resolution with our **patch-based** method. Input image is 256×256, output image is 1024×1024. Please view on digital display for the best view.