DeepSEE: Deep Disentangled Semantic Explorative Extreme Super-Resolution

Marcel C. Bühler[0000−0001−8104−9313], Andrés Romero[0000−0002−7118−5175], and Radu Timofte[0000−0002−1478−0402]

Computer Vision Lab, ETH Zürich, Switzerland
{buehlmar,roandres,timofter}@ethz.ch

Abstract. Super-resolution (SR) is by definition ill-posed. There are infinitely many plausible high-resolution variants for a given low-resolution natural image. Most of the current literature aims at a single deterministic solution of either high reconstruction fidelity or photo-realistic perceptual quality. In this work, we propose an explorative facial super-resolution framework, DeepSEE, for Deep disentangled Semantic Explorative Extreme super-resolution. To the best of our knowledge, DeepSEE is the first method to leverage semantic maps for explorative super-resolution. In particular, it provides control of the semantic regions, their disentangled appearance and it allows a broad range of image manipulations. We validate DeepSEE on faces, for up to 32× magnification and exploration of the space of super-resolution. Our code and models are available at: https://mcbuehler.github.io/DeepSEE/.

Keywords: explorative super-resolution, face hallucination, stochastic super-resolution, extreme super-resolution, disentanglement, perceptual super-resolution, generative modeling, disentanglement

1 Introduction

In super-resolution (SR), we learn a mapping G_Θ from a low-resolution (LR) image x_{lr} to a higher-resolution (HR) image $\hat{x}_{hr}$:

$$\hat{x}_{hr} = G_\Theta(x_{lr}). \tag{1}$$

Simple methods, like bilinear, bicubic or nearest-neighbour, do not restore high-frequency content or details—their output looks unrealistic. Most modern super-resolution methods rely on neural networks to learn a more complex mapping. Typical upscaling factors are 4× to 8× [1–11]; generating 4^2, respectively 8^2 pixels for one input pixel. Very recent works upscale up to 16× [12–14] and 64× [15].

The mapping between the low- and the high-resolution domain is not well defined. There exist multiple (similar) high-resolution images that would downscale to the same low-resolution image. This is why super-resolution is an *ill-posed inverse problem*. Yet, most modern methods assume a ground truth; and learn to generate a single result for a given input [16, 17, 4, 5, 7–11].

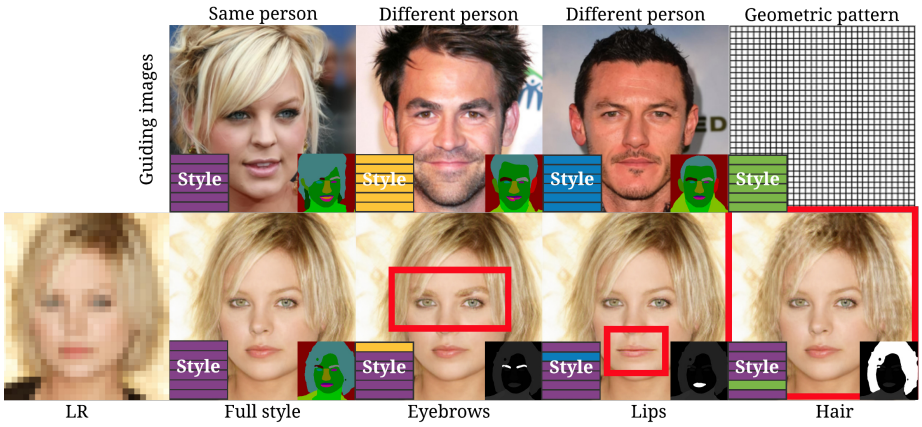

Fig. 1. Upscaling and Manipulations with Disentangled Style Injection. The bottom row shows the low-resolution input and four high-resolution variants; the top row displays the guiding images. Our model, *DeepSEE*, can apply the *full style* matrix from an image of the same person, and alter it with styles from guiding images. We learn 19 semantic regions, such as *eyebrows*, *lips*, *hair*, etc. *DeepSEE* also allows style extraction from *geometric patterns*, sampling in the solution space (Fig. 2), style interpolation (Fig. 6), semantic manipulations (Fig. 7), and upscaling to extreme magnification factors (Fig. 2 and Fig. 8).

To add more guidance, some methods leverage additional information to tackle the super-resolution problem. This can include image attributes [5, 8, 11, 6], reference images [18, 19] and/or a guidance by facial landmark [9, 7, 10]. Still, neither of those approaches allows to produce more than very few variants for a given input. Ideally, we could generate an infinite number of potentially valid solutions an pick the one that suits our purpose best.

Our proposed method, *DeepSEE*, is capable of generating a large number of high-resolution candidates for a low-resolution face image. The outputs differ in both appearance and shape, but they are overall consistent with the low-resolution input. Our method learns a one-to-many mapping from a low-frequency input to a disentangled manifold of potential solutions with the same low-frequencies, but diverse high-frequencies. For inference, a user can specifically tweak the shape and appearance of individual semantic regions to achieve the desired result. *DeepSEE* allows to sample randomly varied solutions (Fig. 2), interpolate between solution variants (Fig. 6), control high-frequency details via a guiding image (Fig. 1), and manipulate pre-defined semantic regions (Fig. 7). In addition, we go beyond common upscaling factors and magnify up to 32×.

Fig. 2. Multiple Potential Solutions for a Single Input. We upscale with factor 32× to different high-resolution variants. *Which one would be the correct solution?*

1.1 Contributions

i) We introduce *DeepSEE*, a novel face hallucination framework for **Deep** disentangled **S**emantic **E**xplorative **E**xtreme super-resolution.
ii) We tackle the ill-posed super-resolution problem in an **explorative** approach based on **semantic maps**. *DeepSEE* is able to sample and manipulate the solution space to produce an infinite number of high-resolution faces for a single low-resolution input.
iii) *DeepSEE* gives control over both shape and appearance. A user can tweak the **disentangled semantic** regions individually.
iv) We super-resolve to the **extreme**, with upscaling factors up to 32×.

2 Related Work

2.1 Fidelity vs. Perceptual Quality in Super-resolution

Single image super-resolution assumes the availability of a low-resolution image carrying low-frequency information—basic colors and shapes—and aims to restore the high-frequencies—sharp contrasts and details. The output is a high-resolution image that is consistent with the low-frequency input image.

Traditional super-resolution methods focused on ***fidelity***: low distortion to a high-resolution ground truth image. These methods based on edge [20, 21] and image statistics [22, 23] and relied on traditional supervised machine learning algorithms: support-vector regression [24], graphical models [25], Gaussian process regression [26], sparse coding [27] or piece-wise linear regression [28].

With the advent of deep learning, the focus shifted to ***perceptual*** quality: photo-realism as perceived by humans. Their results are less blurry and more realistic [2], defining more and more the current main stream research [16, 17, 29, 30, 1, 31, 3, 14].

Evaluation. Traditional evaluation metrics in super-resolution are *Peak Signal-to-Noise Ratio* (PSNR) or *Structural Similarity Index* [32] (SSIM). However, these *fidelity* metrics are simple functions that measure the distortion to reference images and correlate poorly with the human visual response of the output [16, 2, 17, 4]. A high PSNR or SSIM does not guarantee a perceptually good looking output [33]. Alternative metrics evaluate perceptual quality, namely the

Learned Perceptual Image Patch Similarity (LPIPS) [34] and the *Fréchet Inception Distance* (FID) [35]. In this work, we emphasize our validation on high visual quality as in [16, 17, 36, 4], exploration of the solution space [37] and extreme super-resolution [12–14].

2.2 Perceptual Super-resolution

Super-resolution methods with focus on high fidelity tend to generate blurry images [2]. In contrast, perceptual super-resolution targets photo-realism. Training perceptual models typically includes perceptual losses [38, 39], or Generative Adversarial Networks (GAN) [16, 2, 17, 40, 4].

Generative Adversarial Networks [41] (GAN) have become increasingly popular in image generation [42–46]. The underlying technique is to alternately train two neural networks—a generator and a discriminator—with contrary objectives, playing a MiniMax game. While the discriminator aims to correctly classify images as real or fake, the generator learns to produce photo-realistic images fooling the discriminator.

A seminal GAN-based work for perceptual super-resolution, SRGAN [16], employed a residual network [47] for the generator and relied on a discriminator [41] for realism. A combination of additional losses encourage reconstruction/fidelity and texture/content. ESRGAN [17] further improved upon SRGAN by tweaking its architecture and loss functions.

In this work, we propose a *GAN-based perceptual* super-resolution method.

2.3 Explorative Super-resolution

One severe shortcoming of existing approaches is that they consider super-resolution as a *1:1* problem: A low-resolution image maps to a single high-resolution output [4, 5, 7–11, 16, 17]. In reality, however, an infinite number of consistent solutions would exist for a given low-frequency input. Super-resolution is by definition an *ill-posed inverse problem*. Downscaling many (similar) high-resolution variants would yield the same low-resolution image [3, 48–52, 15]. In our work, we regard super-resolution as a *1:n* problem: A low-resolution image maps to many consistent high-resolution variants.

In a concurrent work, Bahat *et al.* [37] suggest an editing tool with which a user can manually manipulate the super-resolution output. Their manipulations include adjusting the variance or periodicity for textures, reducing brightness, or brightening eyes in faces. Two recent works leverage normalizing flows [53, 54] for non-deterministic super-resolution [51, 55]. In our work, we allow to freely walk a latent style space and manipulate semantic masks to explore even more solutions.

To the best of our knowledge, [37] and ours are the first works targeting *semantically controllable* explorative super-resolution; and *DeepSEE* is the first method that achieves explorative super-resolution using semantically-guided style imposition.

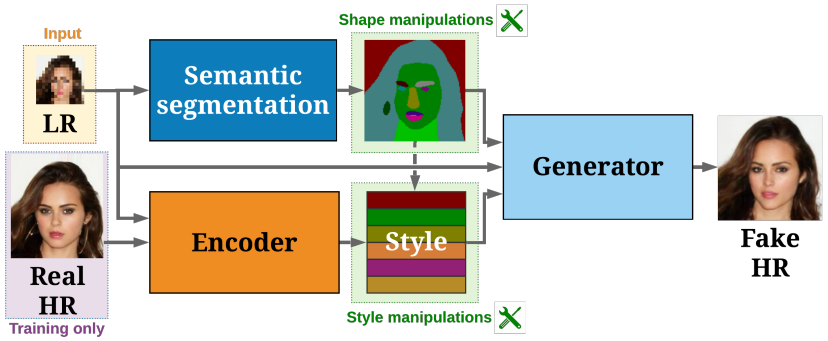

Fig. 3. Overview of Components and Information Flow. *DeepSEE* guides the upscaling with a semantic map extracted from the low-resolution input, and a latent style matrix encoded from a high- or low-resolution image. During inference, a user can tweak the output by manipulating shapes and style codes.

2.4 Domain-specific Super-resolution

Typical domain-specific applications include super-resolution of faces [5, 7–11], outdoor scenes [4] or depth maps [56–59]. Applying super-resolution in a constraint domain allows to leverage prior knowledge and additional guidance, like enforcing characteristics via attributes or identity annotations [8, 5, 11, 6], facial landmarks [9, 7, 10], guiding images [18, 19], or semantic maps [60, 4].

In this work, we focus on super-resolution for faces, namely *face hallucination*. Despite the important roles of facial keypoints, attributes and identities, they are a high-level supervision that does not allow fine-grained manipulation of the output—oftentimes a desired property. In contrast to previous works, we use a predicted discrete semantic prior for each region of the face.

2.5 Extreme Super-resolution

Recent extreme super-resolution train on the DIV8K dataset [61] and target 16× upscaling [12–14]. A concurrent work [15] searches the latent space of a pre-trained face generation model [42] to find high-resolution images that match a low-resolution image, when downscaled 64×.

3 DeepSEE

3.1 Problem Formulation

A low-resolution input ($x_{lr} \in \mathbb{R}^{H_{lr} \times W_{lr} \times 3}$) image acts as a starting point that carries the low-frequency information. A generator (G_Θ) upscales this image and hallucinates the high-frequencies yielding the high-resolution image $\hat{x}_{hr} \in \mathbb{R}^{H_{hr} \times W_{hr} \times 3}$. As a guidance, G_Θ leverages both a high resolution semantic map

($M \in \mathbb{R}^{H_{hr} \times W_{hr} \times N}$, where N is the number of the semantic regions) and independent styles per region ($S \in \mathbb{R}^{N \times d}$, where d is the style dimensionality). The upscaled image should thus retain the low-frequency information from the low-resolution image. In addition, it should be consistent in terms of the semantic regions and have specific, yet independent styles per region. We formally define our problem as

$$\hat{x}_{hr} = G_\Theta(x_{lr},\, M,\, S). \tag{2}$$

Remarkably, thanks to the flexible semantic layout, a user is able to control the *appearance* and *shape* of each semantic region through the generation process. This allows to tweak an output until the desired solution has been found.

3.2 Architecture

Following the GAN framework [41], our method consists of a generator and a discriminator network. In addition, we employ a segmentation network and an encoder for style. Concretely, the segmentation network predicts the semantic mask from a low-resolution image and the encoder produces a disentangled style. Fig. 3 illustrates our model at a high level and Fig. 4 provides a more detailed view. In the following, we describe each component in more detail.

Style Encoder. The style encoder E extracts N style vectors of size d from an input image and combines them to a style matrix $S \in \mathbb{R}^{N \times d}$. Remarkably, it can extract the style from either a low-resolution image x_{lr} or a high-resolution image x_{hr} and maps the encoded style to the same latent space $\mathcal{S}$. The encoder disentangles the regional styles via the semantic layout M. The resulting style matrix serves as guidance for the generator. During inference, a user can sample from the latent style space $\mathcal{S}$ to produce diverse outputs. Please note that the encoder never combines high- and low-resolution inputs; the input is *either* a high-resolution image *or* a low-resolution image.

The style encoder consists of a convolutional neural network E_{lr} for the low-resolution and a similar convolutional neural network E_{hr} for the high resolution input. Their output is mapped to the same latent style space via a shared layer E_{Shared}. Fig. 4 illustrates the flow from the inputs to the style matrix. The architecture for the high-resolution input E_{hr} consists of four convolution layers. The input is downsized twice in the intermediate two layers and upsampled again after a bottleneck. Similarly, the low-resolution encoder E_{lr} consists of four convolution layers. It upsamples the feature map once before the shared layer. The resulting feature map is then passed through the shared convolution layer E_{Shared} and mapped to the range $[-1, 1]$.

Inspired by Zhu *et al.* [62], as a final step, we collapse the output of the shared style encoder for each semantic mask using regional average pooling. This is an important step to disentangle style codes across semantic regions. We describe the regional average pooling in detail in the supplementary material.

Generator. Our generator learns a mapping $G_\Theta(x_{lr}|M, S)$, where the model conditions on both a semantic layout M and a style matrix S. This allows to influence the *appearance*, as well as the *size* and *shape* of each region in the semantic layout.

The semantic layout $M \in \{0,1\}^{H_{hr} \times W_{hr} \times N}$ consists of one binary mask for each semantic region $\{M_0, \cdots, M_{N-1}\}$. For style, we assume a uniform distribution $S \in [-1,1]^{N \times d}$, where each row in S represents a style vector of size d for one semantic region.

At a high level, the generator is a series of residual blocks with upsampling layers in between. Starting from the low-resolution image, it repeatedly doubles the resolution using nearest neighbor interpolation and processes the result in residual blocks. In the residual blocks, we inject semantic and style information through multiple normalization layers.

For the semantic layout, we use *spatially adaptive normalization* (SPADE) [44]. SPADE learns a spatial modulation of the feature maps from a semantic map. For the style, we utilize *semantic region adaptive normalization* in a similar fashion as [62]. *Semantic region adaptive normalization* is an extension to SPADE, which includes style. Like SPADE, it computes spatial modulation parameters, but also takes into consideration a style matrix computed from a reference image. In our case, we extract the style S from an input image through our style encoder as described in Section 3.2. For more details, please check the supplementary material.

Discriminator. We use an ensemble of two similar discriminator networks. One operates on the full image, and the another one on the half-scale of the generated image. Each network takes the concatenation of an image with its corresponding semantic layout and predicts the realism of overlapping image patches. The discriminator architecture follows [44]. Please refer to the supplementary material for a more detailed description.

Segmentation Network. Our training scheme assumes high-resolution segmentation maps, which in most cases are not available during inference. Therefore, we predict a segmentation map from the low-resolution input image x_{lr}. Particularly, we train a segmentation network to learn the mapping $M = Seg(x_{lr})$, where $M \in \{0,1\}^{H_{hr} \times W_{hr} \times N}$ is a high-resolution semantic map.

3.3 *DeepSEE* Model Variants

We suggest two slightly different variants of our proposed method. The *guided* model learns to super-resolve an image with the help of a high-resolution (HR) reference image. The *independent* model does not require any additional guidance and infers a reference style from the low-resolution image.

The *guided* model is able to apply characteristics from a reference image. When fed a guiding image from the same person, it extracts the original characteristics (if visible). Alternatively, when feeding an image from a different person,

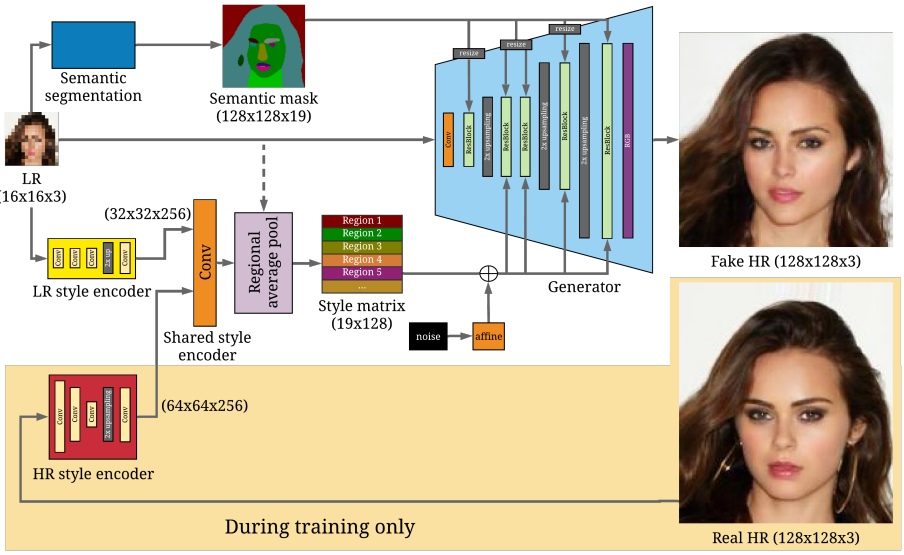

Fig. 4. *DeepSEE* Architecture. Our Generator upscales a low-resolution image (LR) in a series of residual blocks. A predicted semantic mask guides the geometric layout and a style matrix controls the appearance of semantic regions. The noise added to the style matrix increases the robustness of the model. We describe the style encoding, generator and semantic segmentation in Section 3.2.

it integrates those aspects (as long as it is consistent with the low-resolution input). Fig. 1 shows an example, where we first generate an image with the style from the same person and then alter particular regions with styles from other images. The second (*independent*) model applies to the case where no reference image is available.

The *independent* and the *guided* differ in the way the style matrix S is computed. For the *independent* model, we extract the style from the low-resolution input image: $S = E(x_{lr})$. In contrast, the *guided* model uses a high-resolution reference image x_{hr}^{ref} to compute the style $S = E(x_{hr}^{ref})$. It is worth to mention that for training, paired supervision is not necessary as we only require *one* high resolution picture of a person.

3.4 Training

The semantic segmentation network is trained independently from the other networks.

We train the generator, encoder and discriminator end-to-end in an adversarial setting, similar to [44, 63]. As a difference, we inject noise at multiple stages of the generator. We list hyper-parameters and training details in the supplementary material. In the following, we describe the loss function and explain the noise injection.

Loss Function. Our loss function is identical to [44]. Our discriminator computes an adversarial loss $\mathcal{L}_{adv}$ with feature matching $\mathcal{L}_{feat}$: the L1 distance between the discriminator features for the real and the fake image. In addition, we employ a perceptual loss $\mathcal{L}_{vgg}$ from a VGG-19 network [39]. We define our full loss function in Equation 3:

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_{feat}\mathcal{L}_{feat} + \lambda_{vgg}\mathcal{L}_{vgg} \tag{3}$$

We set the loss weights to $\lambda_{feat} = \lambda_{vgg} = 10$; please refer to the supplementary material for more details.

Injection of Noise. After encoding the style to a style matrix S, we add uniformly distributed noise. We define the noisy style matrix S' as $S' = S + U$, where $U_{ij} \sim Uniform(-\delta, +\delta)$. We empirically choose δ based on the model variant.

4 Experimental Framework

4.1 Datasets

We train and evaluate our method on face images from CelebAMask-HQ [64, 65] and CelebA [11]. We use the official training splits for developing and training and test on the provided test splits. All low-resolution images (serving as inputs) are computed via bicubic downsampling. The supplementary material shows qualitative results on the Flickr-Faces-HQ Dataset [42] and on outdoor scenes from ADE20K [66, 4].

4.2 Semantic Segmentation

We train a segmentation network [67, 68] on images from CelebAMask-HQ [65, 64]. The network learns to predict a high-resolution segmentation map with 19 semantic regions from a low-resolution image. As a model, we choose DeepLab V3+ [67–69] with DRN [70, 71] as the backbone.

4.3 Baseline and Evaluation Metrics

We establish a baseline via bicubic interpolation; we first downsample an image to a low-resolution and then upsample it back to the high resolution.

We compute the traditional super-resolution metrics *peak signal-to-noise ratio* (PSNR), *structural similarity index* (SSIM) [32] and the perceptual metrics *Fréchet Inception Distance* (FID) [35] and *Learned Perceptual Image Patch Similarity* (LPIPS) [34]. Our method focuses on generating results of high perceptual quality, measured by LPIPS and FID. PSNR and SSIM are frequently used, however, they are known not to correlate very well with perceptual quality [34]. However, we still list SSIM scores for completion and report PSNR in the supplementary material.

Table 1. We compare with related work for 8× upscaling on high-resolution images of size 128×128 (CelebA [11]) and 256×256 (CelebAMask-HQ [65, 7]). We compute all metrics on the official test sets and list quantitative metrics for related work where checkpoints are available. Both our *DeepSEE* variants outperform the other methods on the perceptual metrics (LPIPS [34] and FID [35]). For qualitative results, please look at Fig. 5 and the supplementary material.

	a) 128×128			b) 256×256		
Method	SSIM ↑	LPIPS ↓	FID ↓	SSIM ↑	LPIPS ↓	FID ↓
Bicubic	0.5917	0.5625	159.60	0.6635	0.5443	125.15
FSRNet (MSE) [9]	0.5647	0.2885	54.48	-	-	-
FSRNet (GAN) [9]	0.5403	0.2304	55.62	-	-	-
Kim *et al.* [7]	**0.6634**	0.1175	11.41	-	-	-
GFRNet [18]	-	-	-	0.6726	0.3472	55.22
GWAInet [19]	-	-	-	0.6834	0.1832	28.79
ours (indep.)	0.6631	**0.1063**	13.84	0.6770	0.1691	22.97
ours (guided)	0.6628	0.1071	**11.25**	**0.6887**	**0.1519**	**22.02**

5 Discussion

We validate our method on two different setups. First, we compare with state-of-the-art methods in face hallucination and provide both quantitative and qualitative results. Second, we show results for extreme and explorative super-resolution by applying numerous manipulations for 32× upscaling.

5.1 Comparison to Face Hallucination Methods

To the best of our knowledge, our method is the first face hallucination model based on discrete semantic masks. We compare with *(i)* models that use reference images [18, 19] and *(ii)* models guided by facial landmarks [9, 7].

For *(i)*, we compare with GFRNet [18] and GWAInet [19], both of which leverage an image from the same person to guide the upscaling. Our method achieves the best scores for all metrics in Table 1 (b). For perceptual metrics, LPIPS [34] and FID [35], *DeepSEE* outperforms the other methods by a considerable margin. As we depict in Fig. 5, our proposed method also produces more convincing results, in particular for difficult regions, such as eyes, hair and teeth. We provide more examples in the supplementary material.

For models based on facial landmarks *(ii)*, Table 1 (a) compares *DeepSEE* with FSRNet [9] and Kim *et al.* [7].[1] *DeepSEE* achieves the highest scores for LPIPS and FID. The supplementary material contains a visual comparison.

It is important to note that given the same inputs (*e.g.* a low-resolution and a guiding image), all related face hallucination models output a single solution;

[1] The models from [9, 7] were trained to generate images of size 128×128, so we can evaluate in their setting on CelebA. [18, 19] generate larger images (256×256, whereas CelebA images have size 218×178), hence we evaluate on CelebAMask-HQ [64, 65].

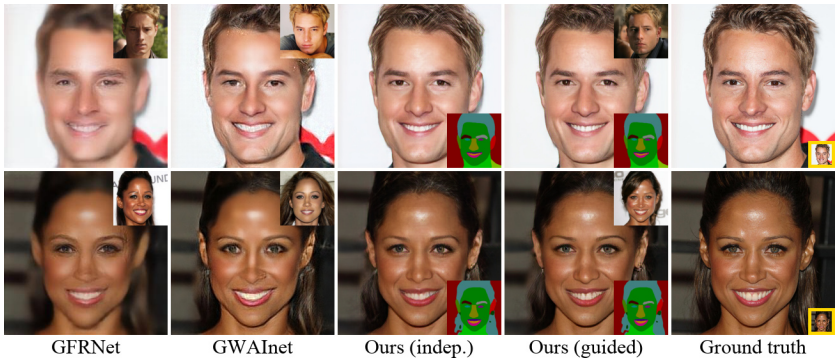

Fig. 5. Comparison to Related Work on 8× Upscaling. We compare with our default solutions for the *independent* and *guided* model. The randomly sampled guiding images are on the top right of each image; the bottom right corner shows the predicted semantic mask (if applicable). Our results look less blurry than GFRNet [18]. Comparing to GWAInet [19], we observe differences in visual quality for difficult regions, like hair, eyes or teeth. With the additional semantic input, our method can produce more realistic textures. Please zoom in for better viewing.

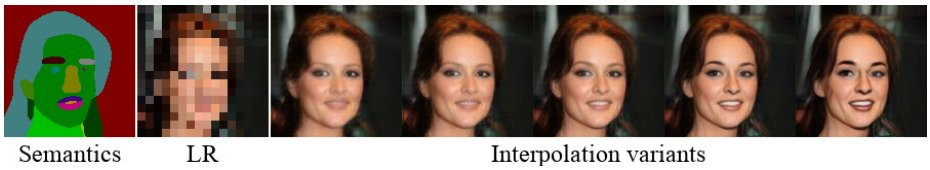

Fig. 6. Interpolation in the Style Latent Space. We linearly interpolate between two style matrices, smoothly increasing contrast. Please refer to the supplementary material for more examples.

despite the fact that there would exist multiple valid results. In contrast, our method can generate an infinite number of solutions, and provides the user with fine-grained control over the output. Fig. 1 and Fig. 2 show several consistent solutions for a low-resolution image. *DeepSEE* can not only extract the overall appearance from a guiding image of the same person, but it can also inject aspects from other people; and even leverage completely different style images, for instance, geometric patterns (Fig. 1). We describe Fig. 1 in more detail in Section 5.2. In addition, our method allows to manipulate semantics, *i.e.* changing the shape, size or content of regions (Fig. 7), *e.g.* eyeglasses, eyebrows, noses, lips, hair, skin, etc. We provide various additional visualizations in the supplementary material.

Fig. 7. Manipulating Semantics for 32× Upscaling. We continuously manipulate the semantic mask and change regional shapes, starting at the default solution (in the first column). In each subsequent column, we highlight the manipulated region and show the resulting image.

5.2 Manipulations

Our proposed approach is an explorative super-resolution model, which allows a user to tune two main *knobs*—the style matrix and the semantic layout—in order to manipulate the model output. Fig. 3 shows these knobs in green boxes.

Style Manipulations. The first way to change the output image is to adapt the disentangled style matrix; for instance by adding random noise (supplementary material), by interpolating between style codes (Fig. 6), or by mixing multiple styles (Fig. 1). Going from one style code to another gradually changes the image output. For example, interpolating between style codes can make contrasts slowly disappear, or on the contrary, become more prominent (Fig. 6).

Semantic Manipulations. The second tuning knob is the semantic layout. The user can change the size and shape of semantic regions, which causes the generator to adapt the output representation accordingly. Fig. 7 shows an example where we close the mouth and make the chin more pointy by manipulating the regions for lips and facial skin. Furthermore, we change the shape of eyebrows, reduce the nose and update the stroke of the eyebrows. It is also possible to create hair on a bold head or add/remove eyeglasses (please check the supplementary material). The manipulations should not be too strong, unrealistic or inconsistent with the low-resolution input. Our model is trained with a strong low-resolution prior and hence, only allows relatively subtle shape manipulations.

5.3 Extreme Super-resolution

While most previous methods apply upscaling factors of 8× [9, 19, 18, 7] or 16× [12–14], *DeepSEE* is capable of going beyond—with upscaling factors of

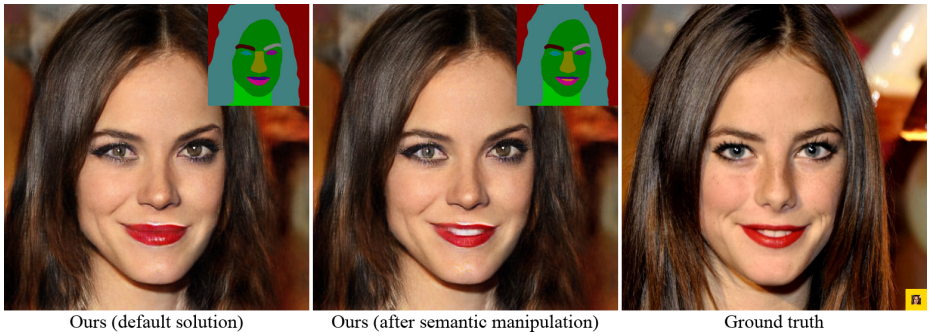

Fig. 8. Extreme Super-resolution. We show how manipulations can align the model output with an expected outcome. Our default solution shows a closed mouth, while given the ground truth, we would expect a smile. After manipulating the semantic mask, *DeepSEE* produces an image that very closely resembles the ground truth.

Table 2. Ablation Study Results. We explore the effect of style and semantics. Semantics have the strongest influence on both fidelity (PSNR, SSIM [32]) and visual quality (LPIPS [34], FID [35]), but the best results require both semantics and style. Finally, using a high-resolution guiding image (*guided*) from the same person provides an additional point of control to the user compared with the *independent* model.

Name	Semantics	LR-Style	HR-Style	SSIM ↑	LPIPS ↓	FID ↓
Prior-only	-	-	-	0.6168	0.1233	25.11
LR-style-only	-	✓	-	0.6485	0.1103	19.29
HR-style-only	-	-	✓	0.6507	0.1108	16.66
Semantics-only	✓	-	-	0.6543	0.1096	12.57
Independent	✓	✓	-	**0.6631**	**0.1063**	13.84
Guided	✓	✓	✓	0.6628	0.1071	**11.25**

up to 32×. Instead of reconstructing a single target, *DeepSEE* can generate multiple variants in a controlled way and hence, a user is more likely to find an expected outcome. Fig. 8 shows an example where the default solution does not perfectly match the ground truth image. A user can now manipulate the semantic mask and create a second version, which is closer to the ground truth image. This shows the power of explorative super-resolution techniques for extreme upscaling factors.

6 Ablation Study

We investigate the influence of *DeepSEE*'s main components—semantics and style injection—in an ablation study. Section 6.1 describes the study setup and we discuss the outcome in Section 6.2.

6.1 Ablation Study Setup

We train four additional models, where we remove the components that inject semantics and/or style. For the first model (*prior-only*), we disable both semantics and style—the model's only conditioning is on the low-resolution input. For the *LR-style-only* and *HR-style-only* models, we do not use the semantic map, but we do condition on the style matrix computed from another low- / high-resolution image of the same person. Lastly, we train a *semantic-only* model that does not inject any style but conditions on semantics.

All models are trained for 7 epochs, which corresponds to 3 days on a single TITAN Xp GPU, with upscaling factor 8× and batch size 4. We use the CelebA [11] dataset. For details, please check the supplementary material.

6.2 Ablation Discussion

All performance scores improve when adding either semantics, style or both (Table 2). Comparing models with either semantics or style (*LR-style-only* and *HR-style-only* vs. *semantics-only*), the perceptual metrics (LPIPS [34] and FID [35]) show better scores when including semantics. Combining both semantic and style yields even better results for both the distortion measures (PSNR and SSIM [32]) and the visual metrics (LPIPS [34] and FID [35]). The performance between our two suggested model variants (the *independent* model and *guided* model) is very similar for fidelity metrics. In terms of perceptual quality, the *guided* image clearly beats the *independent* in FID. However, we empirically find that the *independent* model is more flexible towards random manipulations of the style matrix. Please refer to the supplementary material for visual examples.

7 Conclusion

The super-resolution problem is ill-posed because most high-frequency information is missing and needs to be hallucinated. In this paper, we tackle super-resolution in an explorative approach, *DeepSEE*, based on semantic regions and disentangled style codes. *DeepSEE* allows for fine-grained control of the output, disentangled into region-dependent appearance and shape. Our model goes beyond common upscaling factors and allows to magnify up to 32×. Our validation for faces demonstrate results of high perceptual quality.

Interesting directions for further research could be to identify meaningful directions in the latent style space (*e.g.* age, gender, illumination, contrast, *etc.*), or to apply *DeepSEE* to other domains.

Acknowledgments. We would like to thank the Hasler Foundation. This work was partly supported by the ETH Zürich Fund (OK), by Huawei, Amazon AWS and Nvidia grants.

References

1. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. (2017) 114–125
2. Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., Zelnik-Manor, L.: The 2018 pirm challenge on perceptual image super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018)
3. Cai, J., Gu, S., Timofte, R., Zhang, L.: Ntire 2019 challenge on real image super-resolution: Methods and results. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2019)
4. Wang, X., Yu, K., Dong, C., Change Loy, C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 606–615
5. Yu, X., Fernando, B., Hartley, R., Porikli, F.: Super-resolving very low-resolution face images with supplementary attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 908–917
6. Li, M., Sun, Y., Zhang, Z., Xie, H., Yu, J.: Deep learning face hallucination via attributes transfer and enhancement. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE (2019) 604–609
7. Kim, D., Kim, M., Kwon, G., Kim, D.S.: Progressive face super-resolution via attention to facial landmark. In: Proceedings of the 30th British Machine Vision Conference (BMVC). (2019)
8. Lee, C.H., Zhang, K., Lee, H.C., Cheng, C.W., Hsu, W.: Attribute augmented convolutional neural network for face hallucination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2018) 721–729
9. 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. (2018) 2492–2501
10. Yu, X., Fernando, B., Ghanem, B., Porikli, F., Hartley, R.: Face super-resolution guided by facial component heatmaps. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 217–233
11. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV). (2015)
12. Shang, T., Dai, Q., Zhu, S., Yang, T., Guo, Y.: Perceptual extreme super-resolution network with receptive field block. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. (2020) 440–441
13. Gu, S., Danelljan, M., Timofte, R., Haris, M., Akita, K., Shakhnarovic, G., Ukita, N., Michelini, P.N., Chen, W., Liu, H., et al.: Aim 2019 challenge on image extreme super-resolution: Methods and results. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), IEEE (2019) 3556–3564
14. Zhang, K., Gu, S., Timofte, R.: Ntire 2020 challenge on perceptual extreme super-resolution: Methods and results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. (2020) 492–493
15. Menon, S., Damian, A., Hu, M., Ravi, N., Rudin, C.: Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2020)

16. 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. (2017) 4681–4690
17. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018)
18. Li, X., Liu, M., Ye, Y., Zuo, W., Lin, L., Yang, R.: Learning warped guidance for blind face restoration. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 272–289
19. Dogan, B., Gu, S., Timofte, R.: Exemplar guided face image super-resolution without facial landmarks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2019)
20. Fattal, R.: Image upsampling via imposed edge statistics. In: ACM SIGGRAPH 2007 papers. (2007) 95–es
21. Sun, J., Xu, Z., Shum, H.Y.: Image super-resolution using gradient profile prior. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2008) 1–8
22. Aly, H.A., Dubois, E.: Image up-sampling using total-variation regularization with a new observation model. IEEE Transactions on Image Processing **14** (2005) 1647–1659
23. Zhang, H., Yang, J., Zhang, Y., Huang, T.S.: Non-local kernel regression for image and video restoration. In: European Conference on Computer Vision, Springer (2010) 566–579
24. Ni, K.S., Nguyen, T.Q.: Image superresolution using support vector regression. IEEE Transactions on Image Processing **16** (2007) 1596–1610
25. Wang, Q., Tang, X., Shum, H.: Patch based blind image super resolution. In: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. Volume 1., IEEE (2005) 709–716
26. He, H., Siu, W.C.: Single image super-resolution using gaussian process regression. In: CVPR 2011, IEEE (2011) 449–456
27. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE transactions on image processing **19** (2010) 2861–2873
28. Timofte, R., De Smet, V., Van Gool, L.: A+: Adjusted anchored neighborhood regression for fast super-resolution. In: Asian conference on computer vision, Springer (2014) 111–126
29. 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** (2015) 295–307
30. Kim, J., Kwon Lee, J., Mu Lee, K.: Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 1637–1645
31. Timofte, R., Gu, S., Wu, J., Van Gool, L.: Ntire 2018 challenge on single image super-resolution: Methods and results. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. (2018) 852–863
32. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13** (2004) 600–612
33. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 6228–6237

34. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. (2018)
35. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems. (2017) 6626–6637
36. Sajjadi, M.S., Scholkopf, B., Hirsch, M.: Enhancenet: Single image super-resolution through automated texture synthesis. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 4491–4500
37. Bahat, Y., Michaeli, T.: Explorable super resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 2716–2725
38. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision, Springer (2016) 694–711
39. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR). (2015)
40. 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. (2018) 109–117
41. 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. (2014) 2672–2680
42. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 4401–4410
43. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 8110–8119
44. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 2337–2346
45. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 8789–8797
46. Romero, A., Arbeláez, P., Van Gool, L., Timofte, R.: Smit: Stochastic multi-label image-to-image translation. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2019)
47. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
48. Ren, Z., He, C., Zhang, Q.: Fractional order total variation regularization for image super-resolution. Signal Processing **93** (2013) 2408 – 2421
49. Haris, M., Shakhnarovich, G., Ukita, N.: Deep back-projection networks for super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 1664–1673
50. Li, Y., Dong, W., Xie, X., Shi, G., Jinjian, W., li, X.: Image super-resolution with parametric sparse model learning. IEEE Transactions on Image Processing **PP** (2018)

51. Lugmayr, A., Danelljan, M., Van Gool, L., Timofte, R.: Srflow: Learning the super-resolution space with normalizing flow. In: ECCV. (2020)
52. Ravishankar, S., Reddy, C.N., Tripathi, S., Murthy, K.: Image super resolution using sparse image and singular values as priors. In: International Conference on Computer Analysis of Images and Patterns, Springer (2011) 380–388
53. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real nvp. In: International Conference on Learning Representations (ICLR). (2017)
54. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. In: Advances in neural information processing systems. (2018) 10215–10224
55. Xiao, M., Zheng, S., Liu, C., Wang, Y., He, D., Ke, G., Bian, J., Lin, Z., Liu, T.Y.: Invertible image rescaling. In: ECCV. (2020)
56. Wang, Z., Chen, J., Hoi, S.C.: Deep learning for image super-resolution: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
57. Riegler, G., Rüther, M., Bischof, H.: Atgv-net: Accurate depth super-resolution. In: European conference on computer vision, Springer (2016) 268–284
58. Hui, T.W., Loy, C.C., Tang, X.: Depth map super-resolution by deep multi-scale guidance. In: European conference on computer vision, Springer (2016) 353–369
59. Song, X., Dai, Y., Qin, X.: Deep depth super-resolution: Learning depth super-resolution using deep convolutional neural network. In: Asian conference on computer vision, Springer (2016) 360–376
60. Timofte, R., De Smet, V., Van Gool, L.: Semantic super-resolution: When and where is it useful? Computer Vision and Image Understanding **142** (2016) 1–12
61. Gu, S., Lugmayr, A., Danelljan, M., Fritsche, M., Lamour, J., Timofte, R.: Div8k: Diverse 8k resolution image dataset. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), IEEE (2019) 3512–3516
62. Zhu, P., Abdal, R., Qin, Y., Wonka, P.: Sean: Image synthesis with semantic region-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 5104–5113
63. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 8798–8807
64. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: International Conference on Learning Representations (ICLR). (2018)
65. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 5549–5558
66. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017)
67. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40** (2017) 834–848
68. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Rethinking atrous convolution for semantic image segmentation liang-chieh. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
69. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). (2018) 801–818

70. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: International Conference on Learning Representations (ICLR). (2016)
71. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Computer Vision and Pattern Recognition (CVPR). (2017)