EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis — Supplementary —

Mehdi S. M. Sajjadi Bernhard Schölkopf Michael Hirsch

Max Planck Institute for Intelligent Systems
Spemanstr. 34, 72076 Tübingen, Germany

{msajjadi,bs,mhirsch}@tue.mpg.de

Abstract

In this supplemental, we present some further details on our models and their training procedure, provide additional insights about the influence of the different loss functions to the super-resolution reconstruction, discuss applications and limitations of our approach and show further results and comparisons with other methods. The sections in the supplementary are numbered to match the corresponding sections in the main paper.

4 Additional details on the method

4.2.3 Patch size of texture matching loss

We compute the texture loss $\mathcal{L}_T$ patch-wise to enforce locally similar textures between I_{est} and I_{HR}. We found a patch size of 16×16 pixels to result in the best balance between faithful texture generation and the overall perceptual quality of the images. Figure 1 shows ENet-PAT when trained using patches of size 4×4 pixels for the texture matching loss (ENet-PAT-4) and when it is calculated on larger patches of 128×128 pixels (ENet-PAT-128). Using smaller patches leads to artifacts in textured regions while calculating the texture matching loss on too large patches during training leads to artifacts throughout the entire image since the network is trained with texture statistics that are averaged over regions of varying textures, leading to unpleasant results.

4.2.4 Architecture of the adversarial network

Table 1 shows the architecture of our discriminative adversarial network used for the loss term $\mathcal{L}_A$. We follow common design patterns [13] and exclusively use convolutional layers with filters of size 3×3 pixels with varying stride lengths to reduce the spatial dimension of the input down to a size of 4×4 pixels where we append two fully connected layers along with a sigmoid activation at the output to produce a classification label between 0 and 1.

Output size	Layer
$128 \times 128 \times 3$	Input I_{est} or I_{HR}
$128 \times 128 \times 32$	Conv, lReLU
$64 \times 64 \times 32$	Conv stride 2, lReLU
$64 \times 64 \times 64$	Conv, lReLU
$32 \times 32 \times 64$	Conv stride 2, lReLU
$32 \times 32 \times 128$	Conv, lReLU
$16 \times 16 \times 128$	Conv stride 2, lReLU
$16 \times 16 \times 256$	Conv, lReLU
$8 \times 8 \times 256$	Conv stride 2, lReLU
$8 \times 8 \times 512$	Conv, lReLU
$4 \times 4 \times 512$	Conv stride 2, lReLU
8192	Flatten
1024	Fc, lReLU
1	Fc, sigmoid
1	Estimated label

Table 1. The network architecture of our adversarial discriminative network at 4x super-resolution. As in the generative network, we exclusively use 3×3 convolution kernels. The network design draws inspiration from VGG [17] but uses leaky ReLU activations [11] and strided convolutions instead of pooling layers [13].

5 Further evaluation of results

Our models only learn the residual image between the bicubic upsampled input image and the high resolution output which renders training more stable. Figure 3 displays examples for residual images that our models estimate. ENet-E has learned to significantly increase the sharpness of the image and to remove aliasing effects in the bicubic interpolation (as seen in the aliasing effects in the residual image that cancel out with the aliasing in the bicubic interpolation). ENet-PAT additionally generates fine high-

frequency textures in regions that should be textured while leaving smooth areas such as the sky and the red front areas of the house untouched.

5.1 Additional combinations of losses

In general, we found training models with the adversarial and texture matching loss in conjunction with the Euclidean loss (in place of the perceptual loss) to be significantly less stable and the perceptual quality of the results oscillated heavily during training, *i.e.*, ENet-EA and ENet-EAT are harder to train than ENet-PA and ENet-PAT. This is because the adversarial and texture losses encourage the synthesis of high frequency information in the results, increasing the Euclidean distance to the ground truth images during training which leads to loss functions that counteract each other. The perceptual loss on the other hand is more tolerant to small-scale deviations due to pooling. The results of ENet-EA and ENet-EAT are shown in Fig. 2. We note that the texture matching loss in ENet-EAT leads to a more stable training than ENet-EA and slightly better results, though worse than ENet-PAT. This means that the texture matching loss not only helps create more realistic textures, but it also stabilizes the adversarial training to an extent.

5.2 Comparison with further methods

Figure 5 shows a comparison of our method with Bruna *et al*. [2]. Our model does not suffer from jagged edges and is much sharper.

Figure 6 shows a comparison with RAISR [14] at 2x super-resolution. Since RAISR has been designed for speed rather than state-of-the-art image quality, it reaches a lower performance than previous methods [7, 8, 12] so ENet-E yields visually sharper images even at this low scaling factor. ENet-PAT is the only model to reconstruct sharp details and it is visually much less distinguishable from the ground truth. Despite not being optimized for speed, EnhanceNet is even faster than RAISR at test-time: 9/18ms (EnhanceNet) vs. 17/30ms (RAISR) on average per image at 4x super-resolution on Set5/Set14, though EnhanceNet runs on a GPU while RAISR has been benchmarked on a 6-core CPU.

To demonstrate the performance of our method, we compare the result of ENet-PAT at 4x super-resolution with the current state of the art models at 2x super-resolution in Fig. 4. Although 4x super-resolution is a greatly more demanding task than 2x super-resolution, the results are comparable in quality. Small details that are lost completely in the 4x downsampled image are more accurate in VDSR and DRCN's outputs, but our model produces a plausible image with sharper textures at 4x super-resolution that even outperforms the current state of the art at 2x super-resolution in sharpness, *e.g.*, the area below the eyes is sharper in ENet-PAT's result and looks very similar to the ground truth.

Model	Loss	Weight	VGG layer
ENet-P	$\mathcal{L}_P$	$2 \cdot 10^{-1}$	pool_2
		$2 \cdot 10^{-2}$	pool_5
ENet-PA	$\mathcal{L}_P$	$2 \cdot 10^{-1}$	pool_2
		$2 \cdot 10^{-2}$	pool_5
	$\mathcal{L}_A$	1	–
ENet-PAT	$\mathcal{L}_P$	$2 \cdot 10^{-1}$	pool_2
		$2 \cdot 10^{-2}$	pool_5
	$\mathcal{L}_A$	2	–
	$\mathcal{L}_T$	$3 \cdot 10^{-7}$	$\text{conv}_{1.1}$
		$1 \cdot 10^{-6}$	$\text{conv}_{2.1}$
		$1 \cdot 10^{-6}$	$\text{conv}_{3.1}$

Table 2. Weights for the losses used to train our models.

5.2.1 Quantitative results by PSNR, SSIM and IFC

Tables 3, 4 and 5 show quantitative results measured by PSNR, SSIM and IFC [16] for varying scaling factors. None of these metrics is able to correctly capture the perceptual quality of ENet-PAT's results.

5.2.3 Screenshot of the survey

Figure 7 shows a screenshot of the survey that we used to evaluate the perceptual quality of our results. The subjects were shown the target image on the top and were asked to click the image on the bottom that looks more similar to the target image. Each subject was shown up to 30 images.

5.3 Implementation details and training

The model has been implemented in TensorFlow r0.10 [1]. For all weights, we apply Xavier initialization [5]. For training, we use the Adam optimizer [9] with an initial learning rate of 10^{-4}. We found common convolutional layers stacked with ReLU's to yield comparable results, but training converges faster with the residual architecture. All models were trained only once and used for all results throughout the paper and the supplementary, no fine-tuning was done for any specific dataset or image. Nonetheless, we believe that a choice of specialized training datasets for specific types of images can greatly increase the perceptual quality of the produced textures (*c.f*. Sec. 6).

For the perceptual loss $\mathcal{L}_P$ and the texture loss $\mathcal{L}_T$, we normalized feature activations to have a mean of one [4]. For the texture matching loss, we use a combination of the first convolution in each of the first three groups of layers in VGG, similar to Gatys *et al*. [4]. For the weights, we chose the combination that produced the most realistically looking results. The exact values of the weights for the different losses are given in Table 2.

Figure 1. Comparing different patch sizes for the texture matching loss during training for ENet-PAT on images from ImageNet at 4x super-resolution. Computing the texture matching loss on small patches fails to capture textures properly (ENet-PAT-4) while matching textures on the whole image leads to unpleasant results since different texture statistics are averaged (ENet-PAT-128).

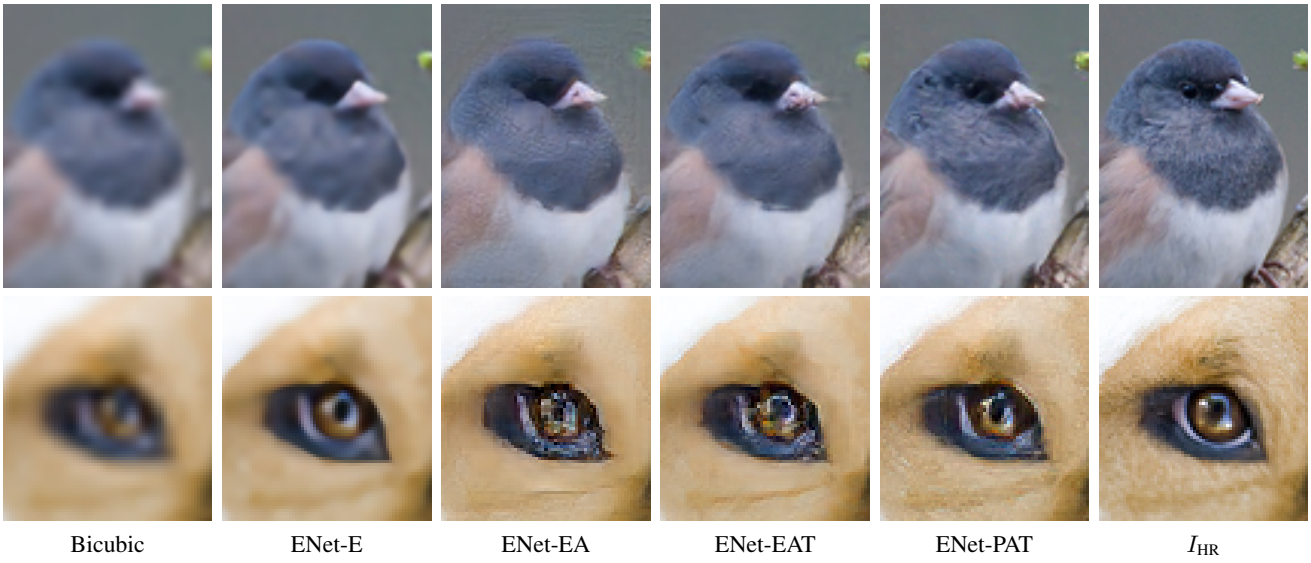
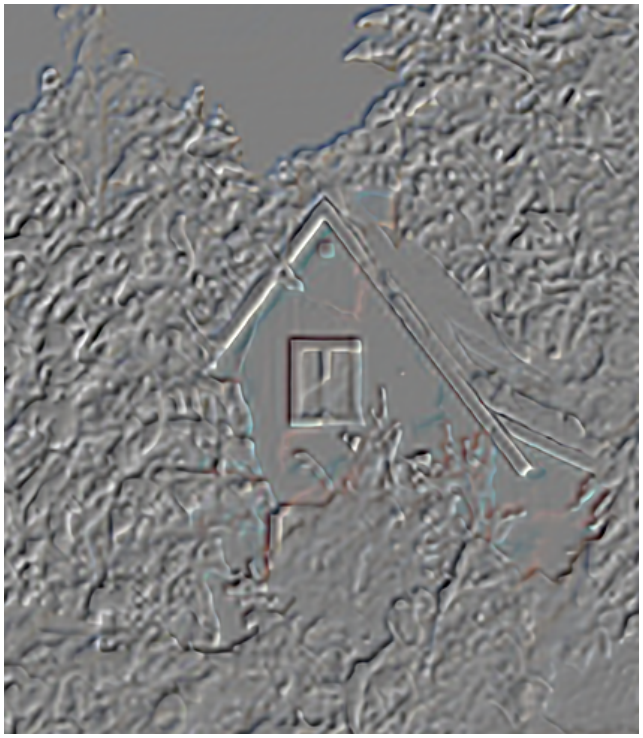


Figure 2. Replacing the perceptual loss in ENet-PA and ENet-PAT with the Euclidean loss results in images with sharp but jagged edges and overly smooth textures (4x super-resolution). Furthermore, these models are significantly harder to train.

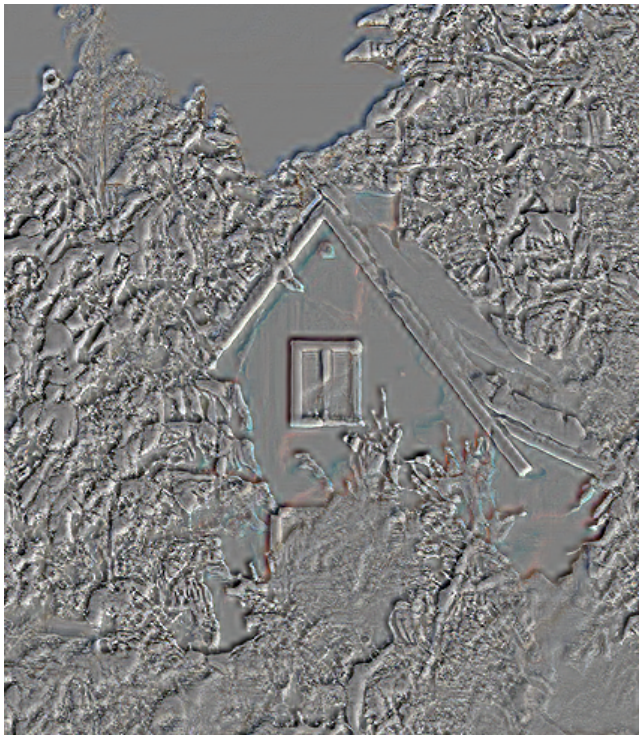
6 Specialized training datasets

Figure 8 shows an example for an image where the majority of subjects in our survey preferred ENet-E's result over the image produced by ENet-PAT. In general, ENet-PAT trained on MSCOCO struggles to reproduce realistically looking faces at high scaling factors and while the overall image is significantly sharper than the result of ENet-E, the human perception is highly sensitive to small changes in the appearance of human faces which is why many subjects preferred the blurry result of ENet-E in those cases. To demonstrate that this is not a limitation of our model, we train ENet-PAT with identical hyperparameters on the CelebA dataset [10] (ENet-PAT-F) and compare the results with ENet-PAT trained on MSCOCO as before. The results are shown in Fig. 9. When trained on CelebA, ENet-PAT-F has significantly better performance.



ENet-E residual

ENet-E result



ENet-PAT residual

ENet-PAT result

Figure 3. A visualization of the residual image that the network produces at 4x super-resolution. While ENet-E significantly sharpens edges and is able to remove aliasing from the bicubic interpolation, ENet-PAT produces additional textures yielding a sharp, realistic result. Image taken from the SunHays80 dataset [18].

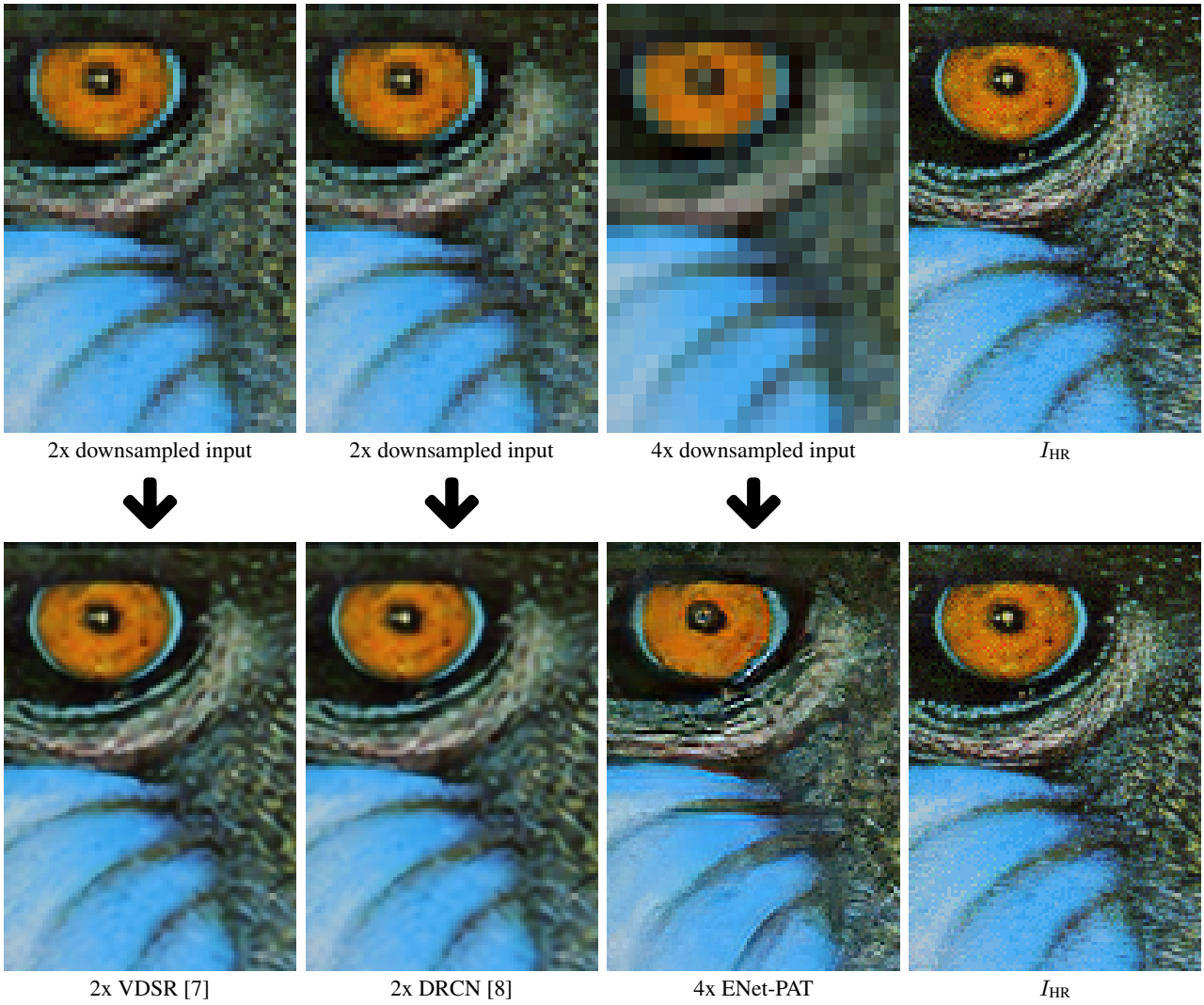


Figure 4. Comparing the previous state of the art by PSNR value at 2x super-resolution (75% of all pixels missing) with our model at 4x super-resolution (93.75% of all pixels missing). The top row shows the input to the models and the bottom row the results. Although our model has significantly less information to work with, it produces a sharper image with realistic textures.

Figure 5. Comparing our model with Bruna *et al*. [2] at 4x super-resolution. ENet-PAT produces images with more contrast and sharper edges that are more faithful to the ground truth.

RAISR [14] ENet-E ENet-PAT I_{HR}

Figure 6. Comparing our model with Romano *et al*. [14] at 2x super-resolution on the butterfly image of Set5. Despite the low scaling factor, image quality gradually increases between RAISR, ENet-E and ENet-PAT, the last of which is not only sharper but also recreates small details better, *e.g*., the vertical white line in the middle of the picture is fully reconstructed only in ENet-PAT's result.

$\alpha = 2$	Bicubic	RFL	A+	SelfEx	SRCNN	PSyCo	DRCN	VDSR	ENet-E	Enet-PAT
Dataset	Baseline	[15]	[19]	[6]	[3]	[12]	[8]	[7]	ours	ours
Set5	33.66	36.54	30.14	36.49	36.66	36.88	**37.63**	37.53	37.32	33.89
Set14	30.24	32.26	27.24	32.22	32.42	32.55	33.04	33.03	**33.25**	30.45
BSD100	29.56	31.16	26.75	31.18	31.36	31.39	31.85	31.90	**31.95**	28.30
Urban100	26.88	29.11	24.19	29.54	29.50	29.64	30.75	30.76	**31.21**	29.00

Table 3. PSNR for different methods at 2x super-resolution. Best performance shown in bold.

$\alpha = 2$	Bicubic	RFL	A+	SelfEx	SRCNN	PSyCo	DRCN	VDSR	ENet-E	Enet-PAT
Dataset	Baseline	[15]	[19]	[6]	[3]	[12]	[8]	[7]	ours	ours
Set5	0.9299	0.9537	0.9544	0.9537	0.9542	0.9559	**0.9588**	0.9587	0.9581	0.9276
Set14	0.8688	0.9040	0.9056	0.9034	0.9063	0.8984	0.9118	0.9124	**0.9148**	0.8617
BSD100	0.8431	0.8840	0.8863	0.8855	0.8879	0.8895	0.8942	0.8960	**0.8981**	0.8729
Urban100	0.8403	0.8706	0.8938	0.8947	0.8946	0.9000	0.9133	0.9140	**0.9194**	0.8303
$\alpha = 4$	Bicubic	RFL	A+	SelfEx	SRCNN	PSyCo	DRCN	VDSR	ENet-E	Enet-PAT
Dataset	Baseline	[15]	[19]	[6]	[3]	[12]	[8]	[7]	ours	ours
Set5	0.8104	0.8548	0.8603	0.8619	0.8628	0.8678	0.8854	0.8838	**0.8869**	0.8082
Set14	0.7027	0.7451	0.7491	0.7518	0.7503	0.7525	0.8670	0.7674	**0.7774**	0.6784
BSD100	0.6675	0.7054	0.7087	0.7106	0.7101	0.7159	0.7233	0.7251	**0.7326**	0.6270
Urban100	0.6577	0.7096	0.7183	0.7374	0.7221	0.7317	0.7510	0.7524	**0.7703**	0.6936

Table 4. SSIM for different methods at 2x and 4x super-resolution. Similar to PSNR, ENet-PAT also yields low SSIM values despite the perceptual quality of its results. Best performance shown in bold.

$\alpha = 4$	Bicubic	RFL	A+	SelfEx	SRCNN	PSyCo	DRCN	VDSR	ENet-E	ENet-PAT
Dataset	Baseline	[15]	[19]	[6]	[3]	[12]	[8]	[7]	ours	ours
Set5	2.329	3.191	3.248	3.166	2.991	3.379	**3.554**	3.553	3.413	2.643
Set14	2.237	2.919	2.751	2.893	2.751	3.055	3.112	**3.122**	3.093	2.281
Urban100	2.361	3.110	3.208	3.314	2.963	3.351	3.461	3.459	**3.508**	2.635

Table 5. IFC for different methods at 4x super-resolution. Best performance shown in bold. The IFC scores roughly follow PSNR and do not capture the perceptual quality of ENet-PAT's results.

Image Quality Assessment

30 images to go!

Target Image

Click the image that looks more similar to the target image above.

Figure 7. Example screenshot of our survey for perceptual image quality. Subjects were shown a target image above and were asked to select the image on the bottom that looks more similar to the target image. In 49 survey responses for a total of 843 votes, subjects selected the image produced by ENet-PAT 91.0%, underlining its higher perceptual quality compared to the state of the art by PSNR, ENet-E.

Bicubic ENet-E ENet-PAT I_{HR}

Figure 8. Failure case for ENet-PAT on an image from ImageNet at 4x super-resolution. While producing an overall sharper image than ENet-E, ENet-PAT fails to reproduce a realistically looking face, leading to a perceptually implausible result.

Bicubic ENet-E ENet-PAT (MSCOCO) ENet-PAT-F I_{HR}

Figure 9. Comparing our models on images of faces at 4x super resolution. ENet-PAT produces artifacts since its training dataset did not contain many high-resolution images of faces. When trained specifically on a dataset of faces (ENet-PAT-F), the same network produces realistic very realistic images, though the results look different from the actual ground truth images (similar to the results in Yu and Porikli [20]). Note that we did not fine-tune the parameters of the losses for this specific task so better results may be possible.

References

[1] M. Abadi et. al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] J. Bruna, P. Sprechmann, and Y. LeCun. Super-resolution with deep convolutional sufficient statistics. In *ICLR*, 2016.

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

[4] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016.

[5] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.

[6] J.-B. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015.

[7] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016.

[8] J. Kim, J. Kwon Lee, and K. Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, 2016.

[9] D. Kingma and J. Ba. Adam: A method for stochastic optimization. 2015.

[10] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[11] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013.

[12] E. Perez-Pellitero, J. Salvador, J. Ruiz-Hidalgo, and B. Rosenhahn. PSyCo: Manifold span reduction for super resolution. In *CVPR*, 2016.

[13] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

[14] Y. Romano, J. Isidoro, and P. Milanfar. RAISR: Rapid and accurate image super resolution. *IEEE TCI*, 2016.

[15] S. Schulter, C. Leistner, and H. Bischof. Fast and accurate image upscaling with super-resolution forests. In *CVPR*, 2015.

[16] H. R. Sheikh, A. C. Bovik, and G. De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE TIP*, 2005.

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[18] L. Sun and J. Hays. Super-resolution from internet-scale scene matching. In *ICCP*, 2012.

[19] R. Timofte, V. De Smet, and L. Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *ACCV*, 2014.

[20] X. Yu and F. Porikli. Ultra-resolving face images by discriminative generative networks. In *ECCV*, 2016.