Supplementary material for: "Super-FAN: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with GANs"

Adrian Bulat and Georgios Tzimiropoulos
Computer Vision Laboratory, The University of Nottingham, United Kingdom
{adrian.bulat, yorgos.tzimiropoulos}@nottingham.ac.uk

Figure 1: A few examples of visual results produced by our system on real-world low resolution faces from WiderFace.

1. Ablation studies

This section describes a series of experiments, further analyzing the importance of particular components on the overall performance. It also provides additional qualitative results.

1.1. On the pixel loss

In this section, we compare the effect of replacing the L2 loss of Eq. 1 with the L1 loss. While the L1 loss is known to be more robust in the presence of outliers, we found no improvement of using it over the L2 loss. The results are shown in Table 1.

$$l_{pixel} = \frac{1}{r^2WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}(I^{HR}_{x,y} - G_{\theta_G}(I^{LR})_{x,y})^2, \quad (1)$$

where W and H denote the size of I^{LR} and r is the upsampling factor (set to 4 in our case).

1.2. On the heatmap loss

Similarly to the above experiment, we also replaced the L2 heatmap loss of Eq. 2 with the L1 loss. The results are shown in Table 3, showing descent improvement for large poses.

$$l_{heatmap} = \frac{1}{N}\sum_{n=1}^{N}\sum_{ij}(\widetilde{M}^n_{i,j} - \widehat{M}^n_{i,j})^2, \quad (2)$$

where $\widetilde{M}^n_{i,j}$ is the heatmap corresponding to the $n-$th landmark at pixel (i,j) produced by running the FAN integrated into our super-resolution network on the super-resolved image $\hat{I}_{HR}$ and $\widehat{M}^n_{i,j}$ is the heatmap obtained by running another FAN on the original image I_{HR}.

1.3. On the importance of the skip connection

Herein, we analyzed the impact of the long-skip connections on the overall performance of the generator. The results, shown in Table 2, show no improvement.

1.4. On network speed

Besides accuracy, another important aspect of network performance is speed. Compared with SR-GAN [3], our generator is only 10% slower, being able to process 1,000 images in 4.6s (vs. 4.3s required by SR-GAN) on an NVIDIA Titan-X GPU.

1.5. Additional qualitative results

Fig. 4 shows the results produced by Super-FAN on all of the 200 randomly selected low-resolution images from

Method	PSNR			SSIM		
	30	60	90	30	60	90
Ours-pixel (L2)	21.55	22.45	23.05	0.8001	0.8127	0.8240
Ours-pixel (L1)	21.47	22.40	23.00	0.7988	0.8120	0.8229

Table 1: PSNR and SSIM when training our generator with L2 and L1 pixel-losses.

Method	PSNR			SSIM		
	30	60	90	30	60	90
Ours-pixel (no-skip)	21.55	22.45	23.05	0.8001	0.8127	0.8240
Ours-pixel (with skip)	21.56	22.45	23.04	0.8021	0.8132	0.8241

Table 2: PSNR and SSIM for "no-skip" and "with skip" versions. The "no-skip" version indicates the absence of the long skip connection, while the "with skip" version adds two new long skip connections, similarly to [2].

WiderFace. Fig. 3 shows the face size distribution of the selected images. Notice that our method copes well with pose variation and challenging illumination conditions. There were a few failure cases, but in most of these cases, it is impossible to tell whether the low-resolution image was actually a face.

Fig. 5 shows a few fitting results produced by Super-FAN on the LS3D-W Balanced dataset. The predictions were plotted on top of the low resolution input images. We observe that our method is capable of producing accurate results even for faces found in arbitrary poses exhibiting various facial expressions.

We also tested our system on images from the Surveillance Cameras Face dataset (SCface) [1]. The dataset contains 4,160 images of 130 unique subjects taken with different cameras from different distances. Fig. 2 shows a few qualitative results from this dataset.

Figure 2: Qualitative results on the SCface dataset [1].

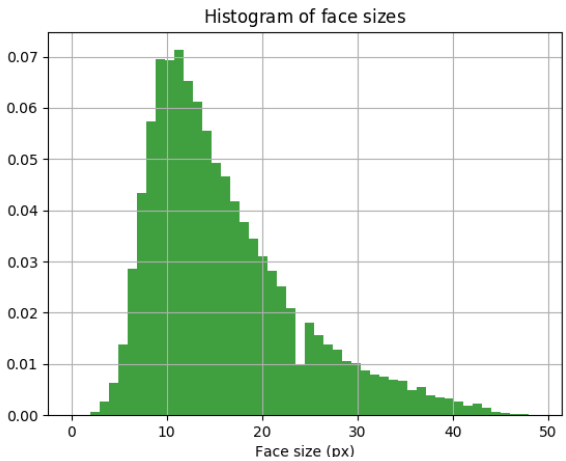


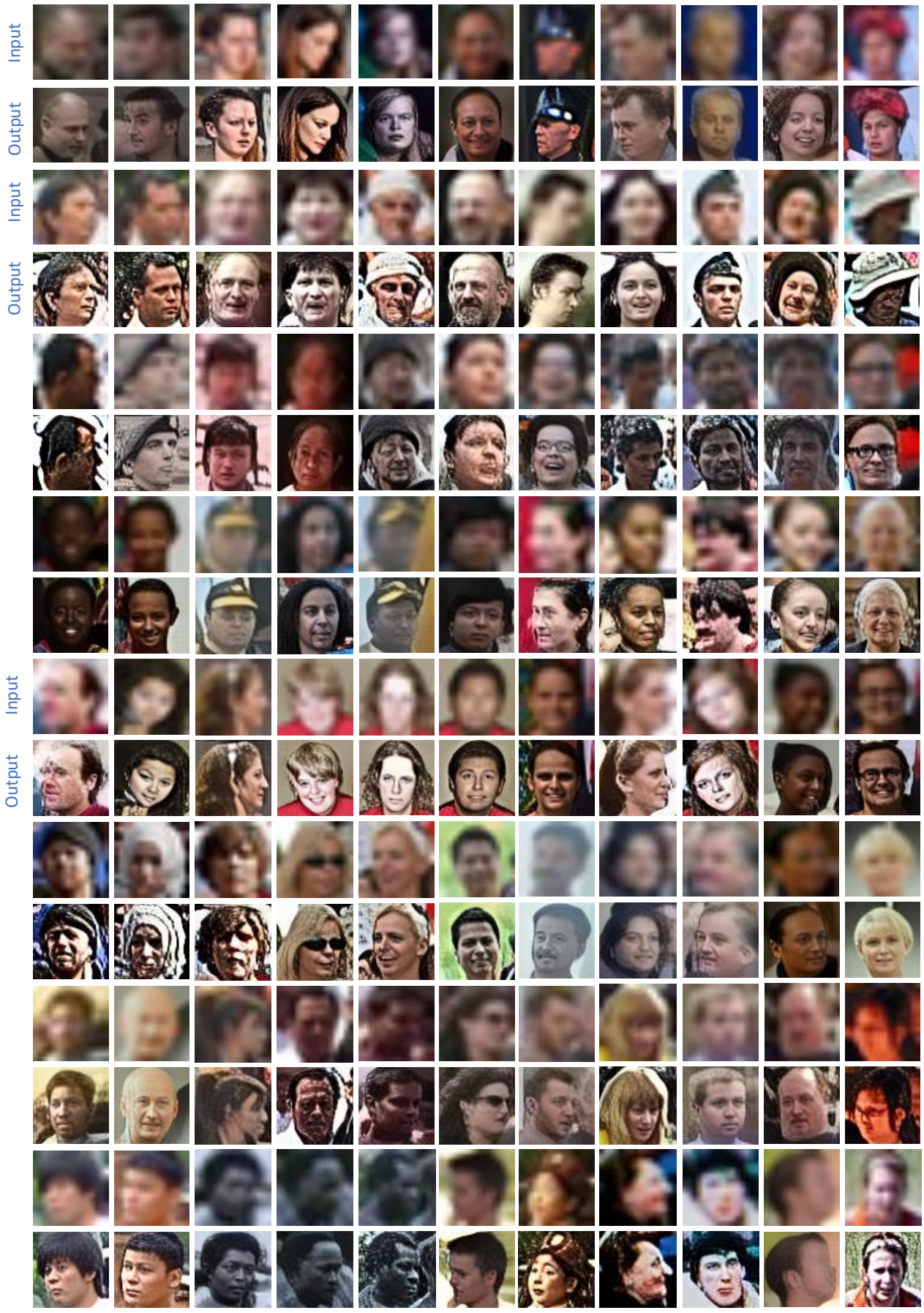
Figure 3: Face size (defined as $\max(width, height)$) distribution of the selected images from WiderFace.

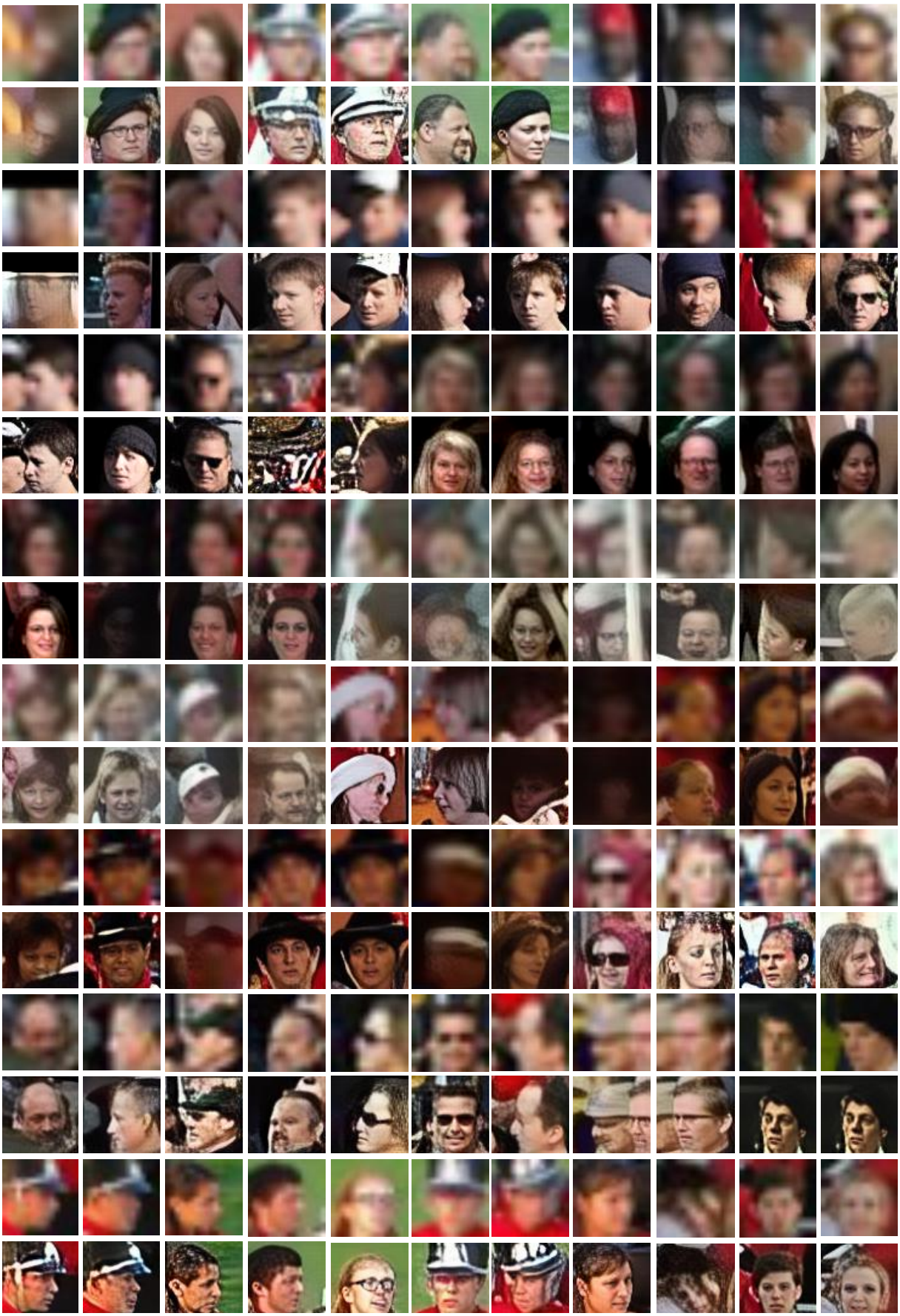
References

[1] M. Grgic, K. Delac, and S. Grgic. Scface–surveillance cameras face database. *Multimedia tools and applications*, 51(3):863–879, 2011. 2

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[3] 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 *CVPR*, 2017. 1


Input
Output
Input
Output
Input
Output



Method	[0-30]	[30-60]	[60-90]
FAN-Ours-pixel-feature-heatmap (L2)	61.0%	55.6%	42.3%
FAN-Ours-pixel-feature-heatmap (L1)	61.1%	55.4%	42.0%

Table 3: AUC across pose (on our LS3D-W balanced test set) for L2 and L1 heatmap losses.

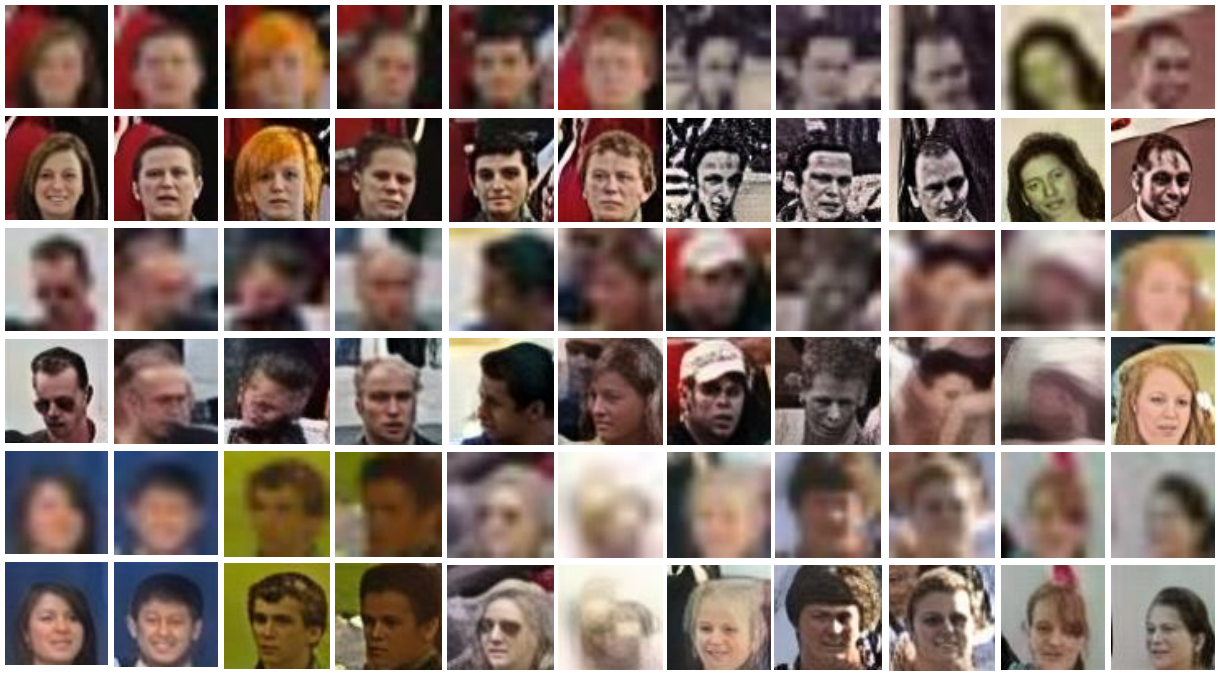


Figure 4: Visual results on a subset of very low resolution images from the WiderFace dataset. The odd rows represent the input, while the even ones the output produced by Super-FAN.

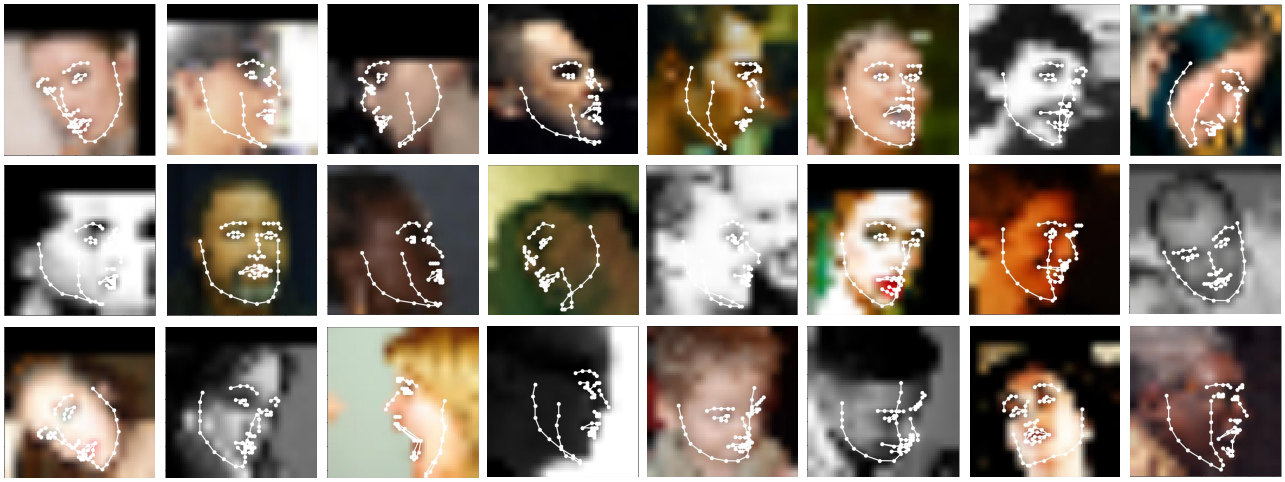


Figure 5: Fitting examples produced by Super-FAN on a few images from LS3D-W. The predictions are plotted over the original low-resolution images. Notice that our method works well for faces found in challenging conditions such as large poses or extreme illumination conditions despite the poor image quality.