Supplementary Material

YU YIN, Northeastern University, USA
JOSEPH P. ROBINSON, Vicarious Surgical, USA
SONGYAO JIANG, Northeastern University, USA
YUE BAI, Northeastern University, USA
CAN QIN, Northeastern University, USA
YUN FU, Northeastern University, USA

ACM Reference Format:
Yu Yin, Joseph P. Robinson, Songyao Jiang, Yue Bai, Can Qin, and Yun Fu. 2021. Supplementary Material. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3474085.3475300

We provide additional results on both constrained (*i.e.*, MultiPIE [3] and CAS-PEAL-R1 [1]) and unconstrained (*i.e.*, CelebA-HQ [9, 11] and LFW [6, 8]) datasets (Section 1). We demonstrate the effectiveness of the proposed SuperFront GAN (SF-GAN) when deployed in realistic settings. From there, we then show its robustness across different views and illumination (Section 2) in a controlled environment (*i.e.*, MultiPIE). Moreover, we gain insight by comparing the qualitative results of SF-GAN and its variants in Section 3. Finally, we describe the network architectures used in the experimental section (Section 4).

1 ADDITIONAL RESULTS

1.1 Constrained faces

Detailed results for MultiPIE [3] are now assessed. Table 1 compares face recognition performance with existing state-of-the-art on setting 1 of Multi-PIE across different poses.

To verify the improved results of SF-GAN across multiple yaws and pitches, we also compare with CAS-PEAL-R1 with its large pose variations. Synthesis results of the state-of-the-art methods are shown in Fig. 2. We show that our method generates the most realistic faces (*i.e.*, finer details in appearances and texture), while preserving identity.

We further show the synthesized high-resolution (HR) frontal results of both SI and MI SF-GAN with poses of 15°, 30°, 45°, 60°. Notice, photo-realistic faces are synthesized from one-to-many LR inputs of arbitrary views. The results for MI SF-GAN were from two low-resolution (LR) inputs: the one used for the SI, and the other the inverted counterpart (*i.e.*, $\pm15^\circ$, $\pm30^\circ$, $\pm45^\circ$, $\pm60^\circ$). Note that all synthesized results of SI SF-GAN are consistent with the ground-truth (GT) faces, showing clear superiority across the different pose and lighting variations. Moreover, MI SF-GAN further improves the image quality of the synthesized images, while preserving the identity even better than SI SF-GAN.

Table 1. **Multi-PIE Setting 1.** Rank-1 (%) for views (α).

	α	±90°	±75°	±60°	±45°	±30°	±15°	Avg
TP-GAN [7]	**LR**	41.07	55.42	62.85	73.03	75.93	76.67	72.12
	SR	52.68	65.98	71.47	77.93	82.53	85.61	79.39
CR-GAN [13]	**LR**	38.94	48.32	54.82	60.25	62.35	67.25	61.17
	SR	49.97	57.18	65.42	69.25	72.72	72.65	70.01
M^{2}FPA [10]	**LR**	46.72	61.03	72.85	82.93	88.01	90.10	83.47
	SR	56.11	76.46	84.90	89.55	94.09	97.25	91.45
single-image (SI) SF-GAN	**LR**	54.10	76.31	89.79	94.24	96.54	98.51	94.77
multi-image (MI) SF-GAN	**LR**	**72.19**	**85.95**	**93.63**	**96.17**	**98.80**	**99.62**	**97.06**

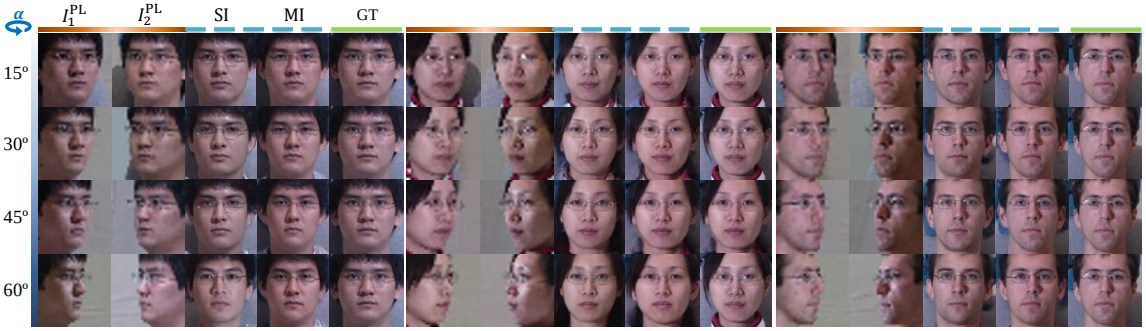

Fig. 1. **SI and MI SF-GAN synthesis results.** SI SSF-GAN recovers better frontal faces than existing methods for different yaws. However, MI SF-GAN further improves the image quality and identity preserving ability.

1.2 Unconstrained faces

We evaluate the frontalization performance of SF-GAN with additional unconstrained datasets. Fig. 3 shows the qualitative results of SF-GAN on the uncontrolled dataset LFW [5] and CelebA-HQ [9]. The result shows that our SF-GAN can properly preserve details, as well as the subject identity of input face.

2 VARIOUS POSE AND ILLUMINATION

To validate the high-level of photo-realism of the images synthesized from arbitrary view and illumination, we show more synthesized frontal results across different poses and illuminations in Fig. 4. Additionally, we vary the illumination from bright-to-dark. Note that all synthesized results of SF-GAN remain consistent with the true faces. This shows the superiority of the proposed across large variations in pose and lighting.

3 QUALITATIVE RESULTS FOR ABLATION STUDY

We gain insights by comparing the quality of outputs of SF-GAN and its variants (Fig. 5). To highlight the importance of *SR side-view, we first compare SF-GAN with and without the SR module. Specifically, we conduct two experiments: (1) remove the SR module (i.e.,* baseline_1); (2) keep the same structure as SF-GAN except without supervision for *SR side-view (i.e.,* baseline_2). Then, we further show the effectiveness of our three-level loss by removing one of the three losses: pixel- (*i.e.*, L_1), patch- (*i.e.*, L_{SSIM}), or global- (*i.e.*, L_{ID}, L_{Adv}). We observe that SF-GAN can synthesize frontal face of higher quality (*e.g.*, finer details and more accurate structures) than all variants. The variants of baseline_1

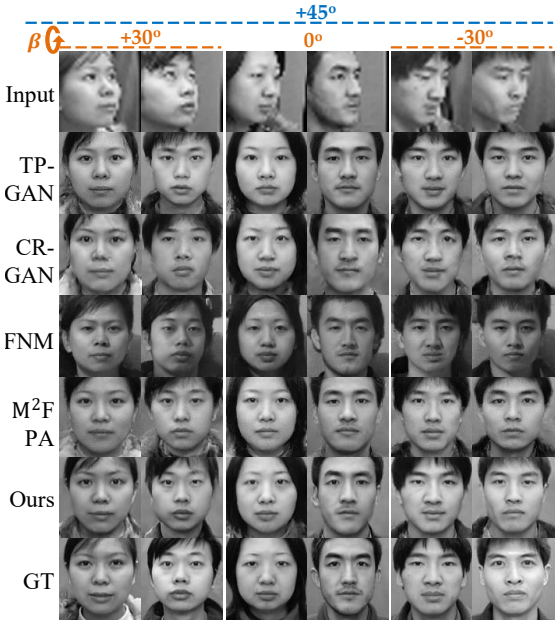

Fig. 2. **Qualitative results.** Comparison with SOTA on constant yaw (*i.e.*, $45°$) with varying pitch ($\beta \in [0°, \pm 30°]$).)

(a) LFW. **(b) CelebA-HQ.**

Fig. 3. **Synthesis of unconstrained data.** Results for LFW (a) and CeleA-HQ (b).

and 2 tend to fail recovering the details of ear and haircut parts. Without L_1 loss: the synthesized images are similar to SF-GAN, but the local textures (*e.g.*, mouth region) of the synthesized results are less like the GT. Without L_{SSIM}

Fig. 4. **Synthesized results across different poses and illuminations.** The first column is input (*i.e.*, LR side-view images). The second column is the side-view super-resolution (SR) face. The third column is the ground-truth (*i.e.*, HR frontal face). The fourth column is the frontal face sythesized with SF-GAN.

loss: the structures of face elements (*e.g.*, mouth, eyebrows, face shape) deform. Without L_{ID} loss: the facial contour becomes distorted. Without L_{Adv} loss: the synthesized images are extremely blurry. In the end, these qualitative results demonstrate the effective of each component of SF-GAN.

4 NETWORK ARCHITECTURES

The generator (G) of our SF-GAN has a deep encoder and a decoder with an SR module integrated. The encoder consists of two 3×3 convolution layers with a stride 1 and 16 residual dense blocks (RDBs) [14]. The RDBs combine the merits of a multi-level residual network and dense connections (Fig. 6). The output sizes of convolution layers and RDBs are $32 \times 32 \times 64$.

Then, features extracted by the deep encoder are reconstructed by the decoder. From the encoder, features split into two branches: a side-view SR-branch that synthesizes a super-resolve side-view image, which is ultimately fed back into the main branch to reconstruct the HR frontal faces.

The specifications of architecture are listed in Table 2. Notice there are 2 up-sampling blocks (*i.e.*, *sr*_1, *sr*_2) in the SR-branch. This recovers the side-view HR image. The up-sampling blocks include a 3×3 convolution layer and a pixel-shuffle [12] layer - it up-scales the feature maps (2×), leading to a final up-scale factor of 4 (*i.e.*, 128×128).

The main branch contains two parts: (1) a fully connected layer (*i.e.*, fc) followed by a maxout [2] and a shallow deconvolution structure (*i.e.*, *dec*_0, *dec*_1, *dec*_2, *dec*_3) to upscale the feature map of fc; (2) the stacked convolution layers (*i.e.*, $conv_{32}$, $conv_{64}$, $conv_{128}$), followed by up-sampling layers (*i.e.*, *upsample*_1, *upsample*_2) for reconstruction.

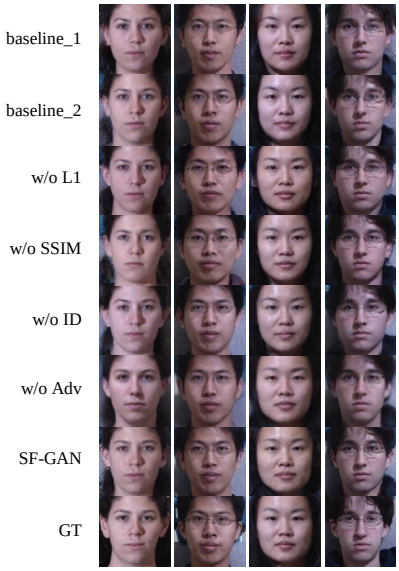

Fig. 5. **Model comparisons.** Synthesis results of the proposed SF-GAN and its variants.

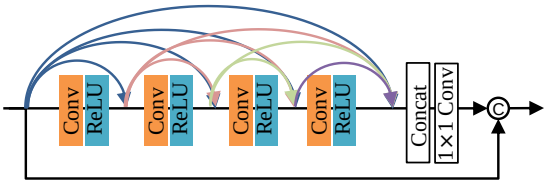

Fig. 6. **Details of RDB.** RDB [14] employed in our encoder for deep feature extraction.

In the second part, convolution layers (*i.e.*, $conv_{32}$, $conv_{64}$, $conv_{128}$) precede two residual blocks [4]. The up-sampling layers are the same as in SR-branch.

Inspired by [10], we employ two discriminators (D) at training (*i.e.*, one for frontal faces D_f and another parsing-guided D_p). The detailed structures of D_f and D_p are summarized in Table 3 and 4, respectively. Each f_convk ($k \in [1, 7]$) in D_f contains a 3×3 convolution layer, batch normalization, leaky ReLU, and a residual block. The p_convk ($k \in [0, 4]$) in D_p is structured like f_convk minus the last res-block.

Table 2. Structure of the generator (G). Upper part shows the structure of encoder, middle is for SR-branch of decoder, and lower is for main branch of decoder.

Layer	Input	Filter size	Output size
$encoder$	I^{LP}	$3 \times 3/1$	$32 \times 32 \times 64$
sr_1	$encoder$	$3 \times 3/1$	$64 \times 64 \times 64$
sr_2	sr_1	$3 \times 3/1$	$128 \times 128 \times 64$
$conv0$	sr_2	$3 \times 3/1$	$128 \times 128 \times 3$
$flatten, fc$	$encoder$	–	512
$maxpool$	fc	–	256
dec_0	$maxpool$	$8 \times 8/1$	$8 \times 8 \times 64$
dec_1	dec_0	$3 \times 3/4$	$32 \times 32 \times 32$
dec_2	$dec0\_1$	$3 \times 3/2$	$64 \times 64 \times 16$
dec_3	$dec0\_2$	$3 \times 3/2$	$128 \times 128 \times 8$
$conv\_32$	$encoder,\ dec\_1,\ sr\_2_{32}$	$3 \times 3/1$	$32 \times 32 \times 64$
$upsample\_1$	$conv\_32$	–	$64 \times 64 \times 64$
$conv\_64$	$upsample\_1,\ dec\_2,\ sr\_2_{64}$	$3 \times 3/1$	$64 \times 64 \times 64$
$upsample\_2$	$conv\_64$	–	$128 \times 128 \times 64$
$conv\_128$	$upsample\_2,\ dec\_3,\ sr\_2_{128}$	$3 \times 3/1$	$128 \times 128 \times 64$
$conv1$	$conv\_128$	$3 \times 3/1$	$128 \times 128 \times 64$
$conv2$	$conv1$	$3 \times 3/1$	$128 \times 128 \times 3$

Table 3. Structure of the frontal face discriminator D_f.

Layer	Input	Filter Size	Output Size
f_conv0	I^{SF}/I^{HF}	$3 \times 3/1$	$128 \times 128 \times 64$
f_conv1	f_conv0	$3 \times 3/2$	$64 \times 64 \times 64$
f_conv2	f_conv1	$3 \times 3/1$	$64 \times 64 \times 128$
f_conv3	f_conv2	$3 \times 3/2$	$32 \times 32 \times 128$
f_conv4	f_conv3	$3 \times 3/1$	$32 \times 32 \times 256$
f_conv5	f_conv4	$3 \times 3/2$	$16 \times 16 \times 256$
f_conv6	f_conv4	$3 \times 3/1$	$16 \times 16 \times 512$
f_conv7	f_conv4	$3 \times 3/2$	$8 \times 8 \times 512$
$fc1$	f_conv6	–	1024
$fc2$	$fc1$	–	1

REFERENCES

[1] Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao. 2007. The CAS-PEAL large-scale Chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 38, 1 (2007), 149–161.

[2] Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. 2013. Maxout networks. In *International conference on machine learning*. PMLR, Atlanta, Georgia, USA, 1319–1327.

[3] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. 2010. Multi-pie. *Image and vision computing* 28, 5 (2010), 807–813.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. IEEE, Las Vegas, USA, 770–778.

[5] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2008. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*. IEEE, Marseille, France.

[6] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Technical Report. University of Massachusetts, Amherst.

[7] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. 2017. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving front view synthesis. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, Venice, Italy, 2439–2448.

Table 4. Structure of the parsing-guided discriminator D_P.

Layer	Input	Filter Size	Output Size
$p1\_conv0$	I^{SF}/I^{HF}	$3 \times 3/2$	$64 \times 64 \times 64$
$p1\_conv1$	$p1\_conv0$	$3 \times 3/2$	$32 \times 32 \times 128$
$p1\_conv2$	$p1\_conv1$	$3 \times 3/2$	$16 \times 16 \times 256$
$p1\_conv3$	$p1\_conv2$	$3 \times 3/2$	$8 \times 8 \times 512$
$p1\_conv4$	$p1\_conv3$	$3 \times 3/2$	$4 \times 4 \times 512$
$p2\_conv0$	I^{SF}/I^{HF}	$3 \times 3/2$	$64 \times 64 \times 64$
$p2\_conv1$	$p2\_conv0$	$3 \times 3/2$	$32 \times 32 \times 128$
$p2\_conv2$	$p2\_conv1$	$3 \times 3/2$	$16 \times 16 \times 256$
$p2\_conv3$	$p2\_conv2$	$3 \times 3/2$	$8 \times 8 \times 512$
$p2\_conv4$	$p2\_conv3$	$3 \times 3/2$	$4 \times 4 \times 512$
$p3\_conv0$	I^{SF}/I^{HF}	$3 \times 3/2$	$64 \times 64 \times 64$
$p3\_conv1$	$p3\_conv0$	$3 \times 3/2$	$32 \times 32 \times 128$
$p3\_conv2$	$p3\_conv1$	$3 \times 3/2$	$16 \times 16 \times 256$
$p3\_conv3$	$p3\_conv2$	$3 \times 3/2$	$8 \times 8 \times 512$
$p3\_conv4$	$p3\_conv3$	$3 \times 3/2$	$4 \times 4 \times 512$
$conv5$	$p1, p2, p3\_conv4$	$3 \times 3/1$	$4 \times 4 \times 512$
$conv6$	$conv5$	$3 \times 3/2$	$2 \times 2 \times 512$
$fc1$	$conv6$	–	1024
$fc2$	$fc1$	–	1

[8] Gary B. Huang Erik Learned-Miller. 2014. *Labeled Faces in the Wild: Updates and New Reporting Procedures*. Technical Report UM-CS-2014-003. University of Massachusetts, Amherst.

[9] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*. IEEE, Virtual, 5549–5558.

[10] Peipei Li, Xiang Wu, Yibo Hu, Ran He, and Zhenan Sun. 2019. M2FPA: A Multi-Yaw Multi-Pitch High-Quality Database and Benchmark for Facial Pose Analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, 3451–3459.

[11] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *ICCV*. IEEE, Las Condes, Chile, 3730–3738.

[12] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, USA, 1874–1883.

[13] Yu Tian, Xi Peng, Long Zhao, Shaoting Zhang, and Dimitris N Metaxas. 2018. CR-GAN: learning complete representations for multi-view generation. In *International Joint Conferences on AI (IJCAI)*. MIT Press, Stockholm, Sweden, 942–948.

[14] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. 2018. Residual dense network for image super-resolution. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Salt Lake City, USA, 2472–2481.