Face Super-Resolution with Better Semantics and More Efficient Guidance

Jin Chen[1,2], Jun Chen[1,2(✉)], Zheng Wang[1,2], Chao Liang[1,2], Zhen Han[1,2], and Chia-Wen Lin[3]

[1] National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University, Wuhan, China
chenj.whu@gmail.com
[2] Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan, China
[3] Department of Electrical Engineering, National Tsinghua University, Hsinchu 30013, Taiwan

Abstract. Recently, facial priors have been widely used to improve the quality of super-resolution (SR) facial images, but it is underutilized in existing methods. On the one hand, facial priors such as semantic maps may be inaccurately estimated on low-resolution (LR) images or low-scale feature maps with L_1 loss. On the other hand, it is inefficient to guide SR features with constant prior knowledge via concatenation at only one intermediate layer of the guidance network. In this paper, we focus on face super-resolution (FSR) based on semantic maps guidance and propose two simple and efficient designs to address the above two limitations respectively. In particular, to address the first limitation, we propose a novel one-hot supervision strategy to pursue accurate semantic maps, which focuses more on penalizing misclassified pixels by relaxing the regression constraint. In addition, a semantic progressive guidance network (SPGN) is proposed that uses semantic maps to learn modulation parameters in normalization layers to efficiently guide SR features layer by layer. Extensive experiments on two benchmark datasets show that the proposed method improves the state-of-the-art in both quantitative and qualitative results at $\times 8$ scale.

Keywords: Face super-resolution · One-hot supervision strategy · Semantic progressive guidance

1 Introduction

Face super-resolution aims to generate high-resolution (HR) facial images from low-resolution observations, which is a challenging problem since it is highly ill-posed due to the ambiguity of the super-resolved pixels. It is a fundamental problem in face analysis and can make a significant contribution to face-related work [1–9].

In contrast to Single Image Super-Resolution (SISR), FSR only focuses on the recovery of facial images. Since different faces share the same components, these

N. Magnenat-Thalmann et al. (Eds.): CGI 2022, LNCS 13443, pp. 53–64, 2022.
https://doi.org/10.1007/978-3-031-23473-6_5

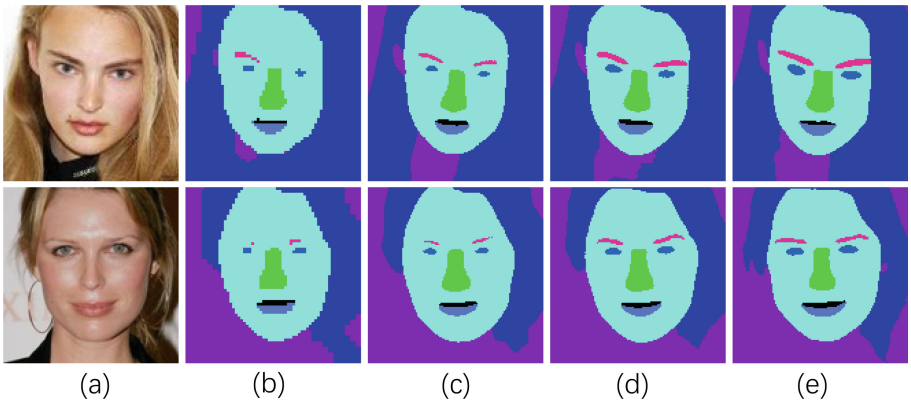

Fig. 1. (a) shows ground truth image, (b) and (c) are parsing maps with resolution of 64 and 128 pixel predicted by L1 loss supervision, (d) shows the predicted parsing map of the SPN via the proposed supervision strategy, (e) shows a ground truth parsing map.

specific facial configurations are a strong prior knowledge that is very useful for FSR. Many FSR methods based on facial priors have been proposed and achieved impressive performance [10–20]. CBN [10] estimates dense correspondence fields as structure prior to guide FSR. Facial component heatmaps are predicted in [11] to provide structure prior to improve the SR quality. SuperFAN [12] uses facial component heatmaps as structure prior to supervise SR network training. In addition, PFSR [13] improves the quality of SR images via a progressive training strategy and multi-scale heatmaps supervision. FSRNet [14] simultaneously estimates landmark heatmaps and semantic maps to improve the details of SR images. JASRNet [17] makes the two mutually reinforcing by jointly learning the SR task and the face alignment task, and using a shared encoder to extract complementary features.

To fully utilize semantic prior knowledge to assist FSR, there exist two key challenges: how to extract accurate semantic prior knowledge and how to effectively use semantic prior knowledge to guide FSR. However, most existing FSR methods do not fully address these issues. On the one hand, previous approaches [14,15] used L_1 loss to supervise the estimation of semantic maps on LR images or low-scale feature maps. However, the semantic maps are difficult to be estimated accurately at the LR level using L_1 loss directly. As shown in Fig. 1(b), when the resolution of the output of semantic maps is 64, the semantic prior network pays more attention to large semantic components (e.g., skin, hair) while small semantic components (e.g., eyes, eyebrows) are easily ignored in the estimation due to the averaging effect of L_1 loss. Therefore, the estimated semantic maps are not accurate enough. On the other hand, in terms of guidance, the semantic maps are simply concatenated with SR features at an intermediate layer of the guidance network and then followed by a convolution layer in [14]. Due to the domain gap effect between SR features and semantic maps, it is

inefficient to capture collaborative knowledge using single-layer guidance. Therefore, it weakens the role of semantic maps in guiding SR features.

In this paper, we propose two simple and efficient designs to address the above two limitations respectively. First, we propose a novel one-hot supervision strategy to pursue accurate semantic maps. Unlike previous approaches that directly use L_1 loss to accurately regress all dimensions of semantic label, the proposed supervision strategy relaxes this constraint by only guaranteeing that the corresponding dimension of the true semantic label has the maximum prediction output. With this relaxation, the semantic prior network can focus more on penalizing misclassified pixels to improve the accuracy of the predicted semantic maps. Second, considering the inadequacy of single-layer guidance, we design a progressive guidance strategy. Also, since different feature layers in the guidance network have different characteristics, different layers should be guided by layer-adaptive semantic prior. Based on these two considerations, we design a semantic progressive guidance network that makes full use of the semantic maps to guide SR features adaptively layer by layer.

In summary, the main contributions of the proposed method are as follows: (i) We propose a novel one-hot supervision strategy to pursue accurate semantic maps by relaxing the regression constraint and focusing more on penalizing misclassified pixels; (ii) We design a semantic progressive guidance network to guide SR features by adaptively learning the modulation parameters of different layers of SR features with semantic maps.

2 Method

2.1 Overview of the Proposed Framework

As shown in Fig. 2, the proposed framework consists of three parts: the Coarse SR Network (CSN), the Semantic Prior Network (SPN), and the Semantic Progressive Guidance Network (SPGN). Given an LR input I_{lr}, we first use the CSN to produce a rough SR facial image I_c to recover the facial structure. Then, the I_c is sent to the SPN to extract semantic maps $\hat{M}$. Finally, both $\hat{M}$ and I_c are sent to the SPGN to progressively guide SR features and recover the final SR facial image I_{sr}.

2.2 Better Semantic Prior

It is crucial to extract accurate semantic maps to guide SR features in the following process. To extract semantic maps, the L_1 loss is usually used to supervise the learning of SPN [14,15]. Due to the averaging effect of L_1 loss, the large semantic components with more pixels seem to dominate the training resulting in small semantic components that are easily ignored in the estimation. As a result, the extracted semantic maps are inaccurate, especially for small semantic components shown in Fig. 1(b)–(c).

Since the L_1 loss aims at regressing all dimensions of semantic label accurately. Given a pixel with semantic label $M_{gt} \in R^N$, and the predicted semantic

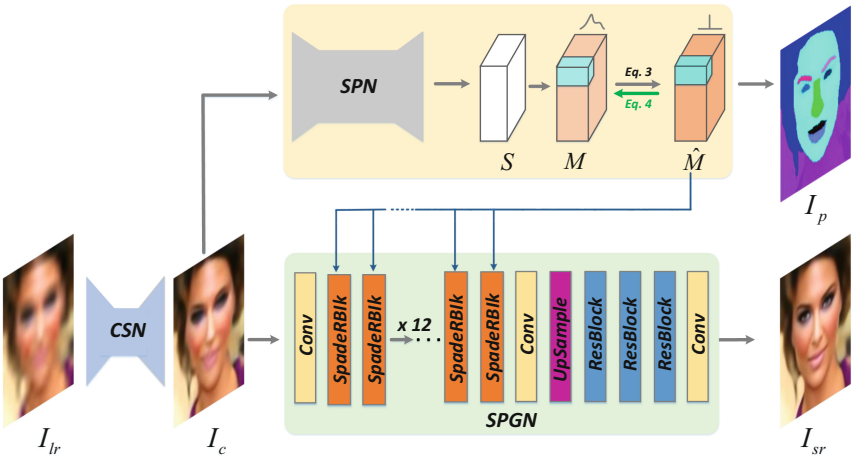

Fig. 2. Overview of the proposed framework. The framework consists of three parts: the CSN is used to recover the coarse SR facial image I_c. The SPN aims at pursuing accurate semantic maps $\hat{M}$. The SPGN focuses on guiding SR features progressively with semantic maps and recovering the final SR facial image I_{sr}.

label is $M_p \in R^N$, and N is the number of semantic class. The L_1 loss of the pixel can be calculated as:

$$L_1 = \sum_{i=1}^{N} |M_{i,p} - M_{i,gt}|, \tag{1}$$

If the true semantic class dimension of a pixel is the j-th dimension, the L_1 loss can be further decomposed into two parts: the loss of semantic-related dimension and other semantic-uncorrelated dimensions. Then the L_1 loss can be further represented as:

$$L_1 = \underbrace{|M_{j,p} - M_{j,gt}|_1}_{semantic-related} + \underbrace{\sum_{i=1,i\neq j}^{N} |M_{i,p} - M_{i,gt}|_1}_{semantic-unrelated}, \tag{2}$$

If we directly use L_1 loss to supervise SPN, there are two drawbacks. On the one hand, it increases the difficulty of network optimization. The result is that when the predicted semantic label has sufficient semantic information, but semantic-unrelated part still causes a loss that cannot be ignored. For example, the true semantic label of a pixel is [0, 0, 1, 0], and when the predicted semantic label is [0.1, 0.1, 0.8, 0.0], we can easily achieve the correct semantic label via *argmax* operation. However, the loss of semantic-unrelated part still brings about 0.2 cost (0.2 = 0.1 + 0.1). We argue that the loss of semantic-unrelated part is unnecessary when the true semantic class dimension has the maximum predicted value. On the other hand, due to the exclusiveness of semantic class definitions, if we try to regress all semantic dimensions accurately, we will ignore

the correlation between different semantic classes. For example, the texture of the skin is closer to the texture of the nose than to the texture of the hair. If we treat the semantic class of hair and nose's equally, it also confuses the network training to some extent.

To alleviate the learning difficulty of SPN and achieve accurate semantic maps, we propose a novel one-hot supervision strategy that relaxes the constraint of regressing all dimensions of semantic label accurately and only guarantees that the corresponding dimension of the true semantic label has the maximum prediction output. An intuitive solution is to transform the predicted semantic maps into one-hot semantic maps using the *argmax* operation before computing the loss. Let's look at the above example again. When the true semantic class dimension of the predicted label has the maximum output, the predicted label [0.1, 0.1, 0.8, 0.0] is first transformed into a one-hot label [0, 0, 1, 0] via *argmax* operation, and since the transformed one-hot label is the same as the true label, there is no loss in updating the network, thus reducing the focus of the SPN on pixels with correctly predicted semantic class. When the predicted label [0.1, 0.5, 0.4, 0.0] has no maximum output for the true semantic dimension. After it is converted to a one-hot label [0, 1, 0, 0], the loss of semantic-related part increases from 0.6 to 1 cost, and it forces the SPN to focus more on penalizing misclassified pixels (*e.g.*, small semantic components), resulting in a more accurate semantic maps.

Due to the *argmax* operation is not differentiable, to achieve one-hot maps while enabling the SPN to be optimized end-to-end, we introduce the Gumbel Softmax trick [31] for this purpose. As shown at the top of Fig. 2, **in the forward process**, the SPN first extracts semantic maps $M \in R^{N\times H\times W}$ with an input I_c, then the M is transformed to one-hot semantic maps $\hat{M} \in R^{N\times H\times W}$, N is the number of semantic class. The one-hot semantic maps $\hat{M}$ can be computed as:

$$\hat{M} = one\_hot(\underset{n}{argmax}\, M_n), \tag{3}$$

Then, we use one-hot semantic maps $\hat{M}$ to compute L_1 loss. **In the backward process**, we compute the gradient of $\hat{M}$ by the following formula:

$$\hat{M} = \frac{exp((M+g)/\tau)}{\sum_{n=1}^{N} exp((M_n+g_n)/\tau)}, \tag{4}$$

where g is drawn from $Gumbel(0,1)$, τ is a temperature value and set as 1 in our experiments. And the optimization process of the one-hot semantic maps $\hat{M}$ in the training stage can be summarized as follows:

$$\hat{M} = \begin{cases} (3),\ forward, \\ (4),\ backward. \end{cases} \tag{5}$$

2.3 More Efficient Guidance

After extracting accurate semantic maps, the key here is how to exploit the semantic maps to guide SR features and further improve the quality of SR

images. To make full use of semantic maps for SR guidance, we design an SPGN. Different from single-layer guidance in previous works [14,15], the SPGN guides SR features at each intermediate layer in the guidance network. Also, since different layers in the guidance network have different characteristics, so SR features of different intermediate layers should be guided by layer-adaptive semantic prior. Motivated by the success of SPADE [25] in the semantic image synthesis task by learning to adaptively modulate the normalization layer, we employ the SpadeRBIk as a unit module to adaptively guide SR features.

As shown at the bottom of Fig. 2, the SPGN starts with a convolution layer (Conv) of stride 2, and followed by twelve SpadeRBIk blocks to progressively guide SR features with semantic maps adaptively, then a Conv, an Upsample block of factor 2, three ResBlock blocks and a Conv to reconstruct the final facial image. A SpadeRBIk [25] block stacks two SPADE blocks and two Conv together and ends with a skip connection. In SPADE, given an SR features $f^i \in R^{C^i \times H^i \times W^i}$, the activation value at site $(c \in C^i, y \in H^i, x \in W^i)$ is given by,

$$\hat{f}^i_{c,y,x} = \gamma^i_{c,y,x} \frac{f^i_{c,y,x} - \mu^i_c}{\sigma^i_c} + \beta^i_{c,y,x}, \tag{6}$$

where $f^i_{c,y,x}$ and $\hat{f}^i_{c,y,x}$ are the input and modulated activation at site (c,y,x), respectively. μ^i_c and σ^i_c are the mean and standard deviation of $f^i_{c,y,x}$ in channel c. The variables $\gamma^i_{c,y,x}$ and $\beta^i_{c,y,x}$ are the learned modulation parameters of the normalization layer using a two-layer convolutional network with semantic maps as input. All convolutions in the SPGN are 3×3 kernel size with 64 channels. The UpSample block and the ResBlock block are described in [14].

2.4 Loss Functions

To make the recovered facial images are of similar visual quality as the origin HR versions, we use L_1 loss as the content loss. The total loss for training can be defined as:

$$L_{total} = \|I_{hr} - I_{sr}\|_1 + \alpha \|I_{hr} - I_c\|_1 + \lambda \left\| M_{gt} - \hat{M} \right\|_1, \tag{7}$$

where I_c, I_{sr}, I_{hr}, $\hat{M}$, M_{gt} are the coarse SR facial image, the final SR facial image, the origin HR image, the predicted semantic maps and the ground truth semantic maps respectively. α and λ are weights of individual loss terms and we set $\alpha = 0.5$, $\lambda = 1.0$ in our experiments empirically.

3 Experiments

3.1 Implementation Details

The proposed method is based on the framework of FSRNet [14], but our Coarse SR Network contains more ResBlock blocks. And the Semantic Prior Network can be replaced by a state-of-the-art parsing network.

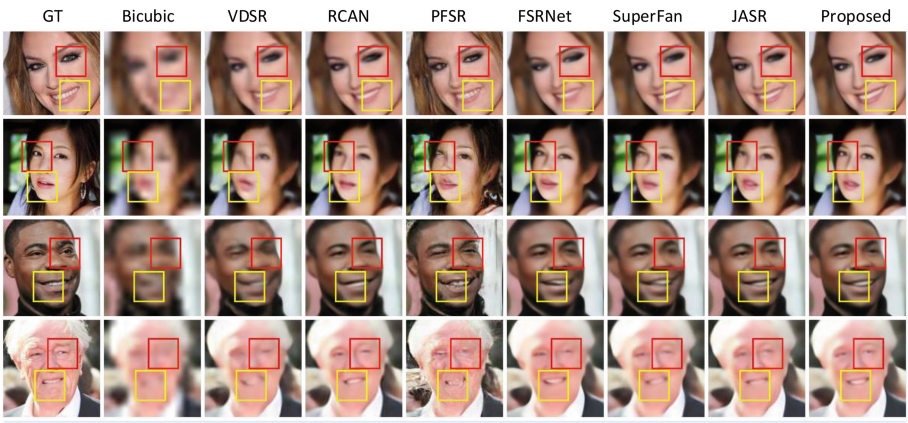

Fig. 3. Visual comparison with state-of-the-art methods. The resolution of input is 16 × 16 and the upscale factor is 8. Other SISR or FSR methods may either produce structural distortions on key facial parts or present undesirable artifacts. The qualitative comparison indicated the proposed method outperforms other SISR or FSR methods.

We conduct experiments on both CelebA [26] and Helen [27]. For both datasets, we use the face parsing model based on EHANet [28] to parse semantic labels as ground truth. For a fair comparison with FSRNet, we merge 19 classes of semantic labels into 11 classes to be consistent with the setup in [14]. Following the experimental setup in [19], we use about 169k images for training and 1k images for testing on the CelebA dataset. For the Helen dataset, we use about 2k images for training and 50 images for testing. All face images in the training and testing stages are resized to 128 × 128 pixels as HR ground truth. The LR faces are obtained by downsampling the HR images to 16 × 16 pixels by bicubic interpolation. To avoid over-fitting, we perform data augmentation on training images with random rotation (90°, 180°, 270°), horizontal flipping and image rescaling in [0.7, 1.3]. PSNR and SSIM [29] are used to quantitatively evaluate SR results. They are computed on the Y channel of transformed YCbCr space. For the quantitative evaluation of semantic maps, the mean of class-wise intersection over union (mIoU) is applied to investigate the accuracy.

For optimization, we set the batch size to 16 and use Adam [30] to optimize the network with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We first train the SPN separately with the learning rate of 1e–4 on HR facial images for 80 epochs, then we jointly train other networks and fine-tune the SPN with fixed learning rate of 1e–5. For CelebA, we train the whole network for 30 epochs with the initial learning rate of 2e–4, divided by 2 at the epoch of [5, 15, 20, 25]. For Helen, we train the whole network for 250 epochs with the learning rate of 2e–4, divided by 2 at the epoch of [40, 120, 200, 220].

3.2 Comparisons with the State-of-the-Arts

We compare our method with state-of-the-art methods general image SR methods, including SRResNet [21], VSDR [22] and RCAN [23], and FSR methods,

Table 1. Comparison of PSNR and SSIM performance with state-of-the-art general SR methods and FSR methods on CelebA and Helen.

Method	CelebA ×8		Helen ×8		Params(M)
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	
Bicubic	23.58	0.6285	23.89	0.6751	–
VDSR [22]	25.68	0.7219	25.24	0.7253	0.67
SRResNet [21]	25.82	0.7369	25.30	0.7297	0.88
RCAN [23]	26.90	0.7779	26.10	0.7599	15.74
URDGN [24]	24.63	0.6851	24.22	0.6909	1.05
SuperFAN [12]	26.69	0.7679	25.61	0.7545	1.49
FSRNet [14]	26.48	0.7718	25.90	0.7759	3.42
FSRGAN [14]	25.06	0.7311	24.99	0.7424	3.42
PFSR [13]	24.43	0.6991	24.73	0.7323	8.97
JASRNet [17]	27.04	0.7833	25.96	0.7565	19.88
Baseline	26.77	0.7740	25.82	0.7555	1.9
Map×64	26.98	0.7836	26.32	0.7779	5.63
Map×128	27.06	0.7860	26.41	0.7815	6.07
Proposed	**27.16**	**0.7904**	**26.49**	**0.7858**	6.07

including URDGN [24], FSRNet [14], SuperFAN [12], PFSR [13] and JASRNet [17]. As shown in Table 1, these general image SR methods improve performance by optimizing the network architecture or introducing attention design, however, those methods which not fully exploit facial prior knowledge result in sub-optimal performance. Compared with FSRNet, our method improves PSNR performance from 26.48 dB to 27.16 dB. Compared with JASRNet, our method has a 0.53 dB improvement is achieved on the small-scale dataset (Helen). And our method achieves the best PSNR and SSIM performance on both datasets. Unlike some previous FSR methods are guided via concatenation, the proposed method uses semantic maps to adaptively learn the modulation parameters of different semantic components in different feature layers of the guidance network to better represent the characteristics of different semantic components and thus better recover the details of semantic components(*e.g.*, eyes, mouth). Figure 3 shows the visual comparison at scale ×8, and we observe that the proposed method recovers the best quality in fine details.

3.3 Ablation Study

To verify the effectiveness of each module in our method, we further implement a series of ablation studies.

Effect of Semantic Maps. We remove the SPN and replace all SpadeRBIk blocks with ResBlock blocks in the SPGN while keeping the CSN unchanged, denoted

by Baseline. As shown in Table 1, compared with Baseline, the proposed method is superior in SR performance on both test sets, which proves that semantic maps are beneficial for FSR.

Table 2. Ablation study of the parsing performance of different parsing strategies on Celeba dataset.

Method	x64	x128	one-hot	mIoU
Map×64	✓			52.42
Map×128		✓		53.85
Proposed		✓	✓	**57.31**

Scale of the Predicted Semantic Maps. We predict semantic maps at 64 pixels and 128 pixels, denoted by Map×64 and Map×128 respectively. And we study the effect of the scale of the predicted semantic maps on parsing performance and SR performance. As shown in Table 2, compared with Map×64, Map×128 achieves a better parsing performance. As shown in Fig. 1(c), Map×128 achieves a more accurate parsing result. It indicates that it is easier to predict accurate semantic maps in HR level. Furthermore, as shown in Table 1, compared with Map×64, Map×128 achieves a better SR performance on both datasets, which suggests that semantic maps are more accurate and more beneficial for FSR.

Efficiency of the One-Hot Supervision Strategy. We evaluate the effect of the *one_hot* supervision strategy on parsing performance and SR performance further. As shown in Table 2, compared with Map×128, the proposed strategy can achieve a 3.46 mIoU performance improvement. As shown in Fig. 1(d), it achieves a more accurate parsing result, especially on small components like eyes and brows. In Table 1, we show that the SR performance is further improved as well. These performance improvements demonstrate the effectiveness of the proposed *one_hot* supervision strategy.

Table 3. Ablation study of different semantic prior on Helen dataset.

Prior	S	$\hat{M}$
PSNR/SSIM	26.24/0.7768	**26.49/0.7858**

Choice of Semantic Prior. We define the output and the last convolution layer of the SPN as two types of semantic prior, denoted by $\hat{M}$ and S, and study the efficiency of different semantic prior. As shown in Table 3, it achieves a better SR performance by guidance with $\hat{M}$, which indicates that the one-hot semantic maps $\hat{M}$ is more suitable for semantic guidance here.

Table 4. Ablation study of the progressive guidance on Helen dataset.

Num	0	1	2	4	8	Proposed
PSNR	25.82	26.06	26.28	26.38	26.44	**26.49**
SSIM	0.7555	0.7659	0.7764	0.7787	0.7845	**0.7858**

Progressive Guidance Strategy. To evaluate the effectiveness of progressive guidance strategy in the SPGN, we keep the depth of the SPGN constant, replace SpadeRBIk block with ResBlock block, and gradually increasing the number of SpadeRBIk blocks from the backend. As shown in Table 4, the SR performance gradually improves as the number of SpadeRBIk blocks increases. Compared with single-layer guidance, the proposed progressive guidance strategy achieves a 0.33dB SR performance improvement. It indicates that the progressive guidance strategy is more effective.

Table 5. Ablation study of the adaptive guidance on Helen dataset.

Strategy	SPGN-SHARED	**SPGN**
PSNR/SSIM	26.22/0.7730	**26.49/0.7858**

Adaptive Guidance Strategy. To evaluate the effectiveness of adaptive guidance strategy in each SPADE, we design a new guidance network, denoted by SPGN-SHARED. Different from SPGN which adaptively learns modulation parameters in all SPADE blocks independently, the SPGN-SHARED uses a shared modulation parameters in all SPADE blocks. As shown in Table 5, The SPGN achieves a higher SR performance, which indicates that the adaptive guidance strategy that learns to modulate SR features each layer independently is more effective.

Limitations. Although semantic prior knowledge improves the quality of super-resolution face images, the parameters of the semantic prior network and the semantic guidance network are large. As shown in Table 1, compared with Baseline, the parameters of the proposed method are increased more than three times. So, how to design lightweight semantic prior network and semantic guidance network is the direction of our future research.

4 Conclusion

In this letter, we propose two simple and efficient designs to improve the quality of SR facial images. Specifically, we propose a novel one-hot supervision strategy to pursue accurate semantic maps and design a semantic progressive guidance network to more efficiently guide SR features. Quantitative and qualitative results of FSR on two benchmark datasets demonstrate the effectiveness of the proposed method.

Acknowledgement. This research was supported partially by National Nature Science Foundation of China (U1903214, 62072347, 62071338, 61876135), in part by the Nature Science Foundation of Hubei under Grant (2018CFA024, 2019CFB472), in part by Hubei Province Technological Innovation Major Project (No. 2018AAA062).

References

1. Bai, Y., Zhang, Y., Ding, M., Ghanem, B.: Finding tiny faces in the wild with generative adversarial network. In: Proceedings of the IEEE Computer Vision and Pattern Recognition, pp. 21–30 (2018)
2. Chen, L., Su, H., Ji, Q.: Face alignment with kernel density deep neural network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6992–7002 (2019)
3. Kumar, A., et al.: LUVLi Face alignment: estimating landmarks location, uncertainty, and visibility likelihood. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8236–8246 (2020)
4. Masi, I., Mathai, J., AbdAlmageed, W.: Towards learning structure via consensus for face segmentation and parsing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5508–5518 (2020)
5. Pan, J., Ren, W., Hu, Z., Yang, M.H.: Learning to deblur images with exemplars. IEEE Trans. Patt. Anal. Mach. Intell **41**(6), 1412–1425 (2019)
6. Ge, S., Zhao, S., Li, C., Zhang, Y., Li, J.: Efficient low-resolution face recognition via bridge distillation. IEEE Trans. Image Process. **29**, 6898–6908 (2020)
7. Ge, S., Zhao, S., Gao, X., Li, J.: Fewer-shots and lower-resolutions: towards ultra-fast face recognition in the wild. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 229–237 (2019)
8. Hsu, C.C., Lin, C.W., Su, W.T., Cheung, G.: Sigan: siamese generative adversarial network for identity-preserving face hallucination. IEEE Trans. Image Process. **28**, 6225–6236 (2019)
9. Hong, S., Ryu, J.: Unsupervised face domain transfer for low-resolution face recognition. IEEE Signal Process. Lett. **27**, 156–160 (2019)
10. Zhu, S., Liu, S., Loy, C.C., Tang, X.: Deep cascaded Bi-network for face hallucination. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 614–630. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_37
11. 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, pp. 217–233 (2018)
12. 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, pp. 109–117 (2018)
13. Kim, D., Kim, M., Kwon, G., Kim, D.S.: Progressive face super-resolution via attention to facial landmark. arXiv preprint arXiv:1908.08239 (2019)
14. 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, pp. 2492–2501 (2018)
15. Wang, C., Zhong, Z., Jiang, J., Zhai, D., Liu, X.: Parsing map guided multi-scale attention network for face hallucination. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2518–2522 (2020)

16. Hu, X., et al.: Face super-resolution guided by 3D facial priors. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 763–780. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_44
17. Yin, Y., Robinson, J., Zhang, Y., Fu, Y.: Joint super-resolution and alignment of tiny faces. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12693–12700 (2020)
18. Xin, J., Wang, N., Gao, X., Li, J.: Residual attribute attention network for face image super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9054–9061 (2019)
19. Ma, C., Jiang, Z., Rao, Y., Lu, J., Zhou, J.: Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In: Proceedings of the IEEE Computer Vision and Pattern Recognition, pp. 5569–5578 (2020)
20. Shen, Z., Lai, W. S., Xu, T., Kautz, J., Yang, M.H.: Deep semantic face deblurring. In: Proceedings of the IEEE Computer Vision and Pattern Recognition, pp. 8260–8269 (2018)
21. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Computer Vision and Pattern Recognition, pp. 4681–4690 (2017)
22. Kim, J., Lee, J. K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Computer Vision and Pattern Recognition, pp. 1646–1654 (2016)
23. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision, pp. 286–301 (2018)
24. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 318–333. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_20
25. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Computer Vision and Pattern Recognition, pp. 2337–2346 (2019)
26. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE Conference on Computer Vision, pp. 3730–3738 (2015)
27. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7574, pp. 679–692. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33712-3_49
28. Luo, L., Xue, D., Feng, X.: Ehanet: an effective hierarchical aggregation network for face parsing. Appl. Sci. **10**(9), 3135 (2020)
29. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
30. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
31. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144 (2016)
32. Liu, Z.S., Siu, W.C., Chan, Y.L.: Reference based face super-resolution. IEEE. Access **7**, 129112–129126 (2019)