Robust Reference-based Super-Resolution via C^2-Matching Supplementary File

Yuming Jiang[1] Kelvin C.K. Chan[1] Xintao Wang[2] Chen Change Loy[1] Ziwei Liu[1]✉
[1]S-Lab, Nanyang Technological University [2]Applied Research Center, Tencent PCG
{yuming002, chan0899, ccloy, ziwei.liu}@ntu.edu.sg xintao.wang@outlook.com

In this supplementary file, we will explain the network structures (*i.e.* Contrastive Correspondence Network and Restoration Network) and training details in Section A. Then we will introduce training losses we used in Section B. In Section C, the model size will be analyzed. In Section D, we will provide more visual comparisons with state-of-the-art methods. Finally, we will show more visual comparisons of ablation study in Section E.

A. Network Structures and Training Details

A.1. Contrastive Correspondence Network

Network Structure. Table 1 shows the detailed feature extractor structure of contrastive correspondence network. Since the resolutions of input image and reference image are different, we adopt two feature extractors for LR input image and HR reference image, respectively.

Table 1. **The feature extractor structure of contrastive correspondence network.** The kernel size of convolution layers is 3×3 and the MaxPool operation is with kernel size of 2×2.

#	Layer name(s)
0	Conv (3, 64), ReLU
1	Conv (64, 64), ReLU
2	MaxPool (2×2)
3	Conv (64, 128), ReLU
4	Conv (128, 128), ReLU
5	MaxPool (2×2)
6	Conv (128, 256)

Training Details. To enable teacher-student correlation distillation, a teacher contrastive correspondence network should be first trained. The hyperparameters for the training of teacher model are set as follows: the margin value m is 1.0, the threshold value T is 4.0, the batch size is set as 8, and the learning rate is 10^{-3}. We use the pretrained weights of VGG-16 to initialize the feature extractor. Then the student contrastive correspondence network is trained with the teacher network fixed. The margin value m, threshold value T, batch size and learning rate are the same as the teacher network. The temperature value τ is 0.15, and the weight α_{kl} for KL-divergence loss is 15.

A.2. Restoration Network

Network Structure. The restoration network consists of dynamic aggregation module and restoration module. For each image, three reference features (*i.e.* pretrained VGG relu3_1, relu2_1, relu1_1 feature [3]) are aggregated by dynamic aggregation module, and the aggregated reference features are denoted as Aggregated Reference Feature1, Aggregated Reference Feature2 and Aggregated Reference Feature3, respectively. The structure of restoration module is illustrated in Table. 2.

Table 2. **The structure of restoration module.** The kernel size of convolution layers is 3×3. PixelShuffle layers are $2\times$. RB denotes residual block. Aggregated Reference Feature denotes the reference feature aggregated by the dynamic aggregation module.

#	Layer name(s)
0	Conv(3, 64), LeakyReLU
1	RB [Conv(64, 64), ReLU, Conv(64, 64)] $\times$ 16
2	Concat [#1, Aggregated Reference Feature1]
3	Conv(320, 64), LeakyReLU
4	RB [Conv(64, 64), ReLU, Conv(64, 64)] $\times$ 16
5	ElementwiseAdd(#1, #4)
6	Conv(64, 256), PixelShuffle, LeakyReLU
7	Concat [#6, Aggregated Reference Feature2]
8	Conv(192, 64), LeakyReLU
9	RB [Conv(64, 64), ReLU, Conv(64, 64)] $\times$ 16
10	ElementwiseAdd(#6, #9)
11	Conv(64, 256), PixelShuffle, LeakyReLU
12	Concat [#11, Aggregated Reference Feature3]
13	Conv(128, 64), LeakyReLU
14	RB [Conv(64, 64), ReLU, Conv(64, 64)] $\times$ 16
15	ElementwiseAdd(#11, #14)
16	Conv(64, 32), LeakyReLU
17	Conv(32, 3)

Training Details. The learning rate is set as 10^{-4}. For the training of the network with adversarial loss and perceptual

loss, we adopt the same setting as [8] (*i.e.* the network is trained with only reconstruction loss for the first 10K iterations).

B. Loss Functions

Reconstruction Loss. The ℓ_1-norm is adopted to keep the spatial structure of the LR images. It is defined as follows:

$$L_{rec} = \left\| I^{HR} - I^{SR} \right\|_1 . \tag{1}$$

Perceptual Loss. The perceptual loss [1] is employed to improve the visual quality. It is defined as follows:

$$L_{per} = \frac{1}{V} \sum_{i=1}^{C} \left\| \phi_i(I^{HR}) - \phi_i(I^{SR}) \right\|_F , \tag{2}$$

where V and C denotes the volume and channel number of feature maps. ϕ denotes the relu5_1 features of VGG19 model [3]. $\|\cdot\|_F$ denotes the Frobenius norm.

Adversarial Loss. The adversarial loss [2] is defined as follows:

$$L_{adv} = -D(I^{SR}). \tag{3}$$

The loss for training discriminator D is defined as follows:

$$L_D = D(I^{SR}) - D(I^{HR}) + \lambda(\left\| \nabla_{\hat{I}} D(\hat{I}) \right\|_2 - 1)^2. \tag{4}$$

where $\hat{I}$ is the random convex combination of I^{SR} and I^{HR}.

C. Comparison of Model Size

The comparison of model size (*i.e.* the number of trainable parameters) is shown in Table 3. Our proposed C^2-Matching has a total number of 8.9M parameters and achieves a PSNR of 28.24dB. For a fair comparison in terms of model size, we build a light version of C^2-Matching, which has fewer trainable parameters. The C^2-Matching-*light* is built by setting the number of residual blocks of layer #9 and layer #14 to 8 and 4, respectively, and removing the Aggregated Reference Feature1. The C^2-Matching-*light* has a total number of 4.8M parameters. The light version has fewer parameters than TTSR [5] but significantly better performance.

Table 3. **Model sizes of different methods.** PSNR / SSIM are adopted as the evaluation metrics.

Method	Params	PSNR/SSIM
RCAN [7]	16M	26.06 / .769
RankSRGAN [6]	1.5M	22.31 / .635
CrossNet [9]	33.6M	25.48 / .764
SRNTT [8]	4.2M	26.24 / .784
TTSR [5]	6.4M	27.09 / .804
C^2-Matching-*light*	4.8M	28.14 / .839
C^2-Matching	8.9M	28.24 / .841

D. More Visual Comparisons with State-of-the-art Methods

In Fig. 1 and Fig. 2, more visual comparisons with ESRGAN [4], RankSRGAN [6], SRNTT [8] and TTSR [5] are provided. The images restored by our proposed C^2-Matching have better visual quality.

E. More Visual Comparisons of Ablation Study

In this paper, the proposed C^2-Matching consists of three major components: Dynamic Aggregation Module (Dyn-Agg), Contrastive Correspondence Network (Contras) and Teacher-Student Correlation Distillation (TS Corr). On top of the baseline model, we progressively add the Dyn-Agg module, Contras network and TS Corr distillation. In Fig. 3, we show more visual comparisons with these proposed modules progressively added.

References

[1] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Eur. Conf. Comput. Vis.*, pages 694–711. Springer, 2016. 2

[2] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4681–4690, 2017. 2

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Int. Conf. Learn. Represent.*, 2015. 1, 2

[4] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Eur. Conf. Comput. Vis. Worksh.*, 2018. 2, 3, 4

[5] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5791–5800, 2020. 2, 3, 4

[6] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Int. Conf. Comput. Vis.*, pages 3096–3105, 2019. 2, 3, 4

[7] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Eur. Conf. Comput. Vis.*, pages 286–301, 2018. 2

[8] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7982–7991, 2019. 2, 3, 4

[9] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *Eur. Conf. Comput. Vis.*, pages 88–104, 2018. 2

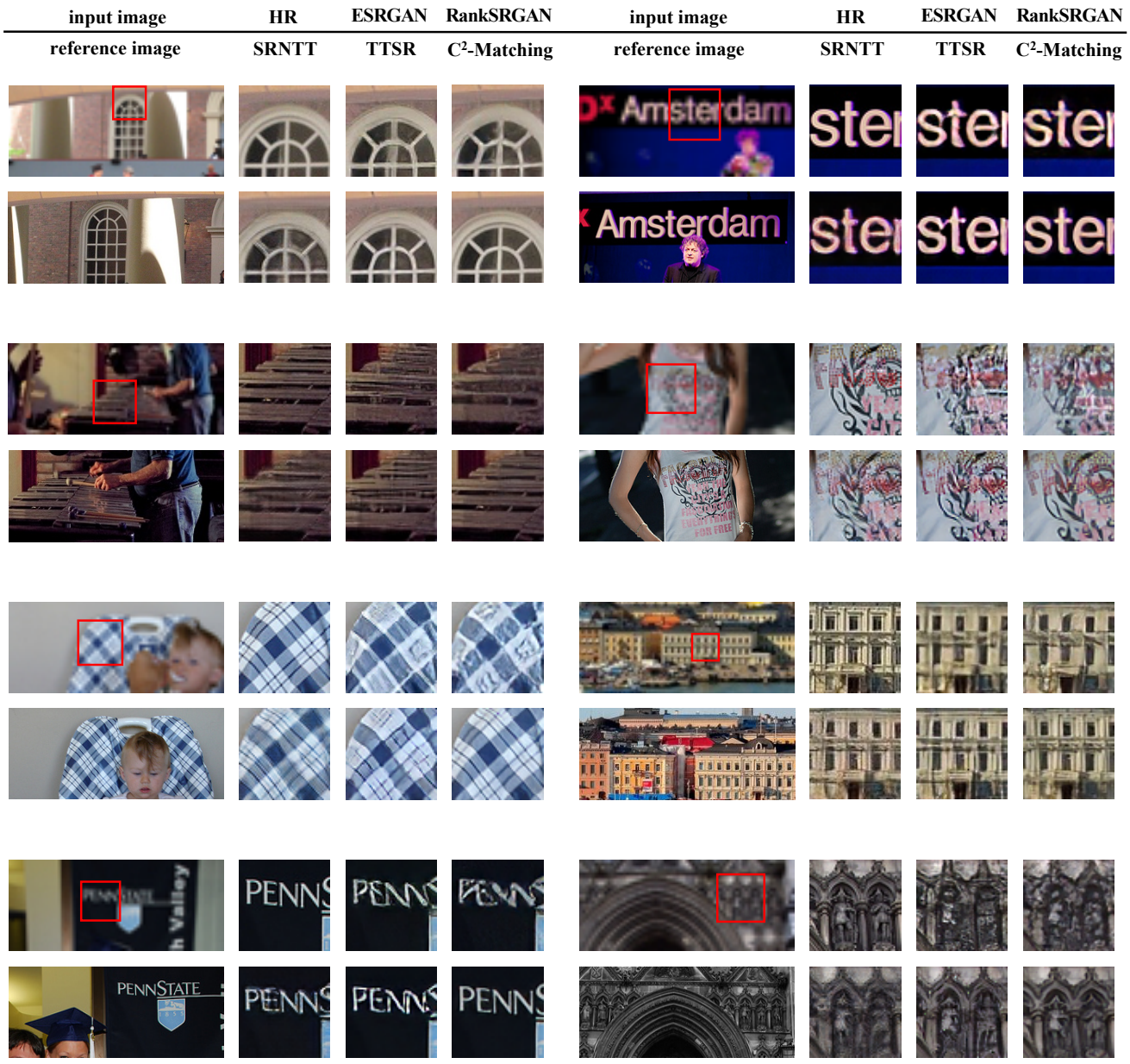

Figure 1. **More visual comparisons.** We compare our results with ESRGAN [4], RankSRGAN [6], SRNTT [8], and TTSR [5]. All these methods are trained with GAN loss. Our results have better visual quality with more texture details.

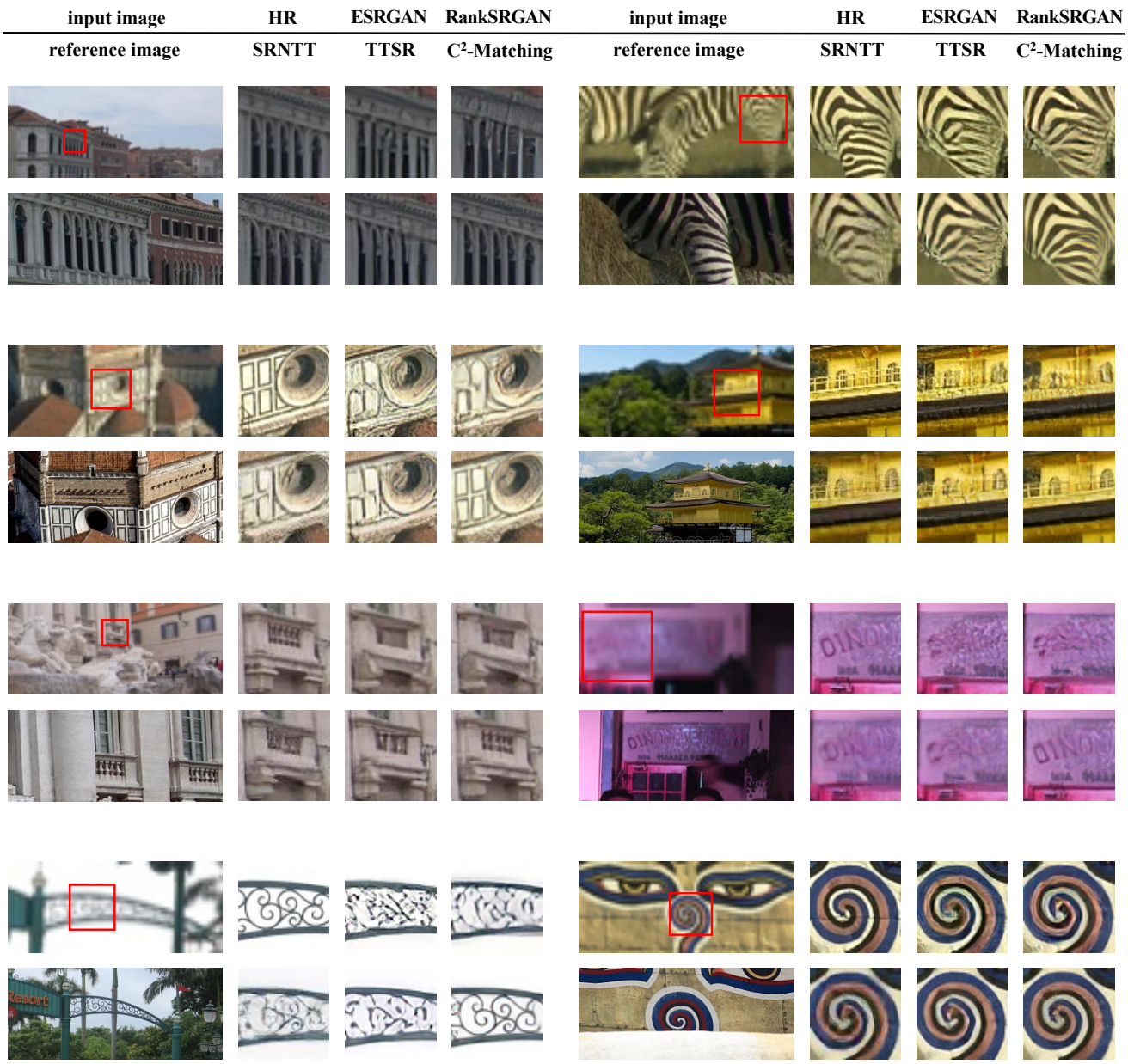

Figure 2. **More visual comparisons.** We compare our results with ESRGAN [4], RankSRGAN [6], SRNTT [8], and TTSR [5]. All these methods are trained with GAN loss. Our results have better visual quality with more texture details.

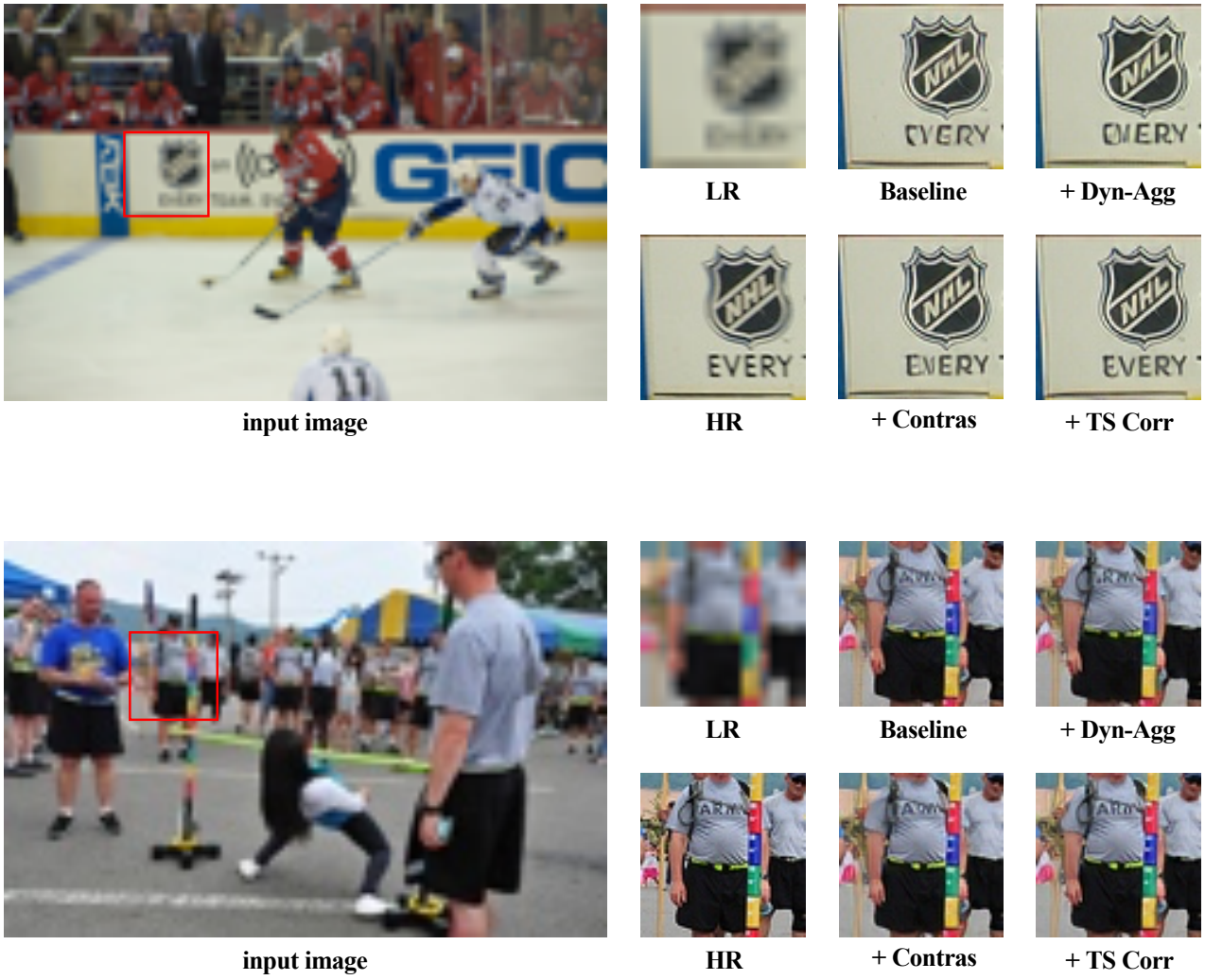

Figure 3. **More visual comparisons of ablation study.** On top of the baseline model, Dynamic Aggregation Module (Dyn-Agg), Contrastive Correspondence Network (Contras) and Teacher-Student Correlation Distillation (TS Corr) are progressively added.