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

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

Reference-based Super-Resolution (Ref-SR)

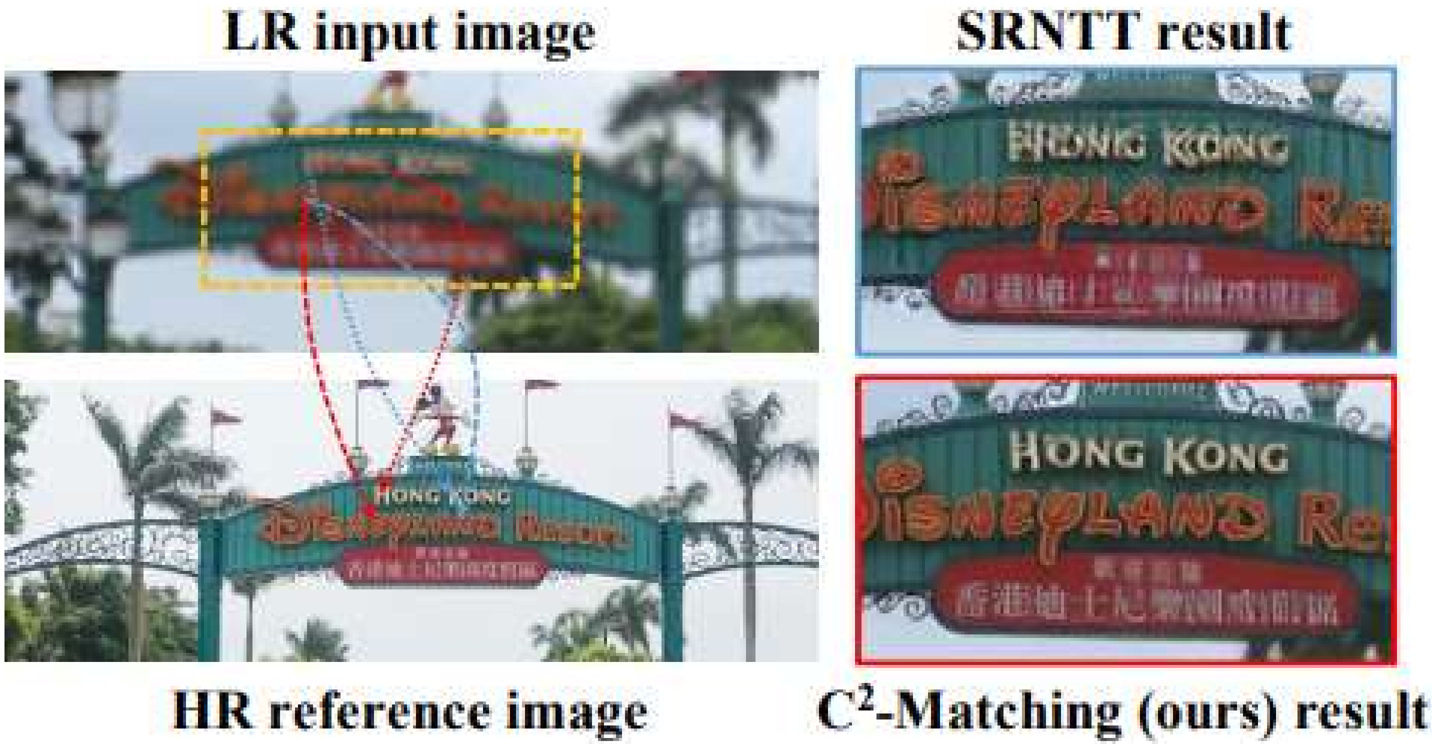
- Super-resolves input images by transferring HR details of reference images
  - Patch-Match method (TTSR, SRNTT)
  - Learnable feature extractor (SRNTT)
  - Deformable Convolution Network (SSEN)

- Performing local transfer is difficult because of two gaps between input and reference image.
  - Transformation gap (scale and rotation)
  - Resolution gap (HR and LR)

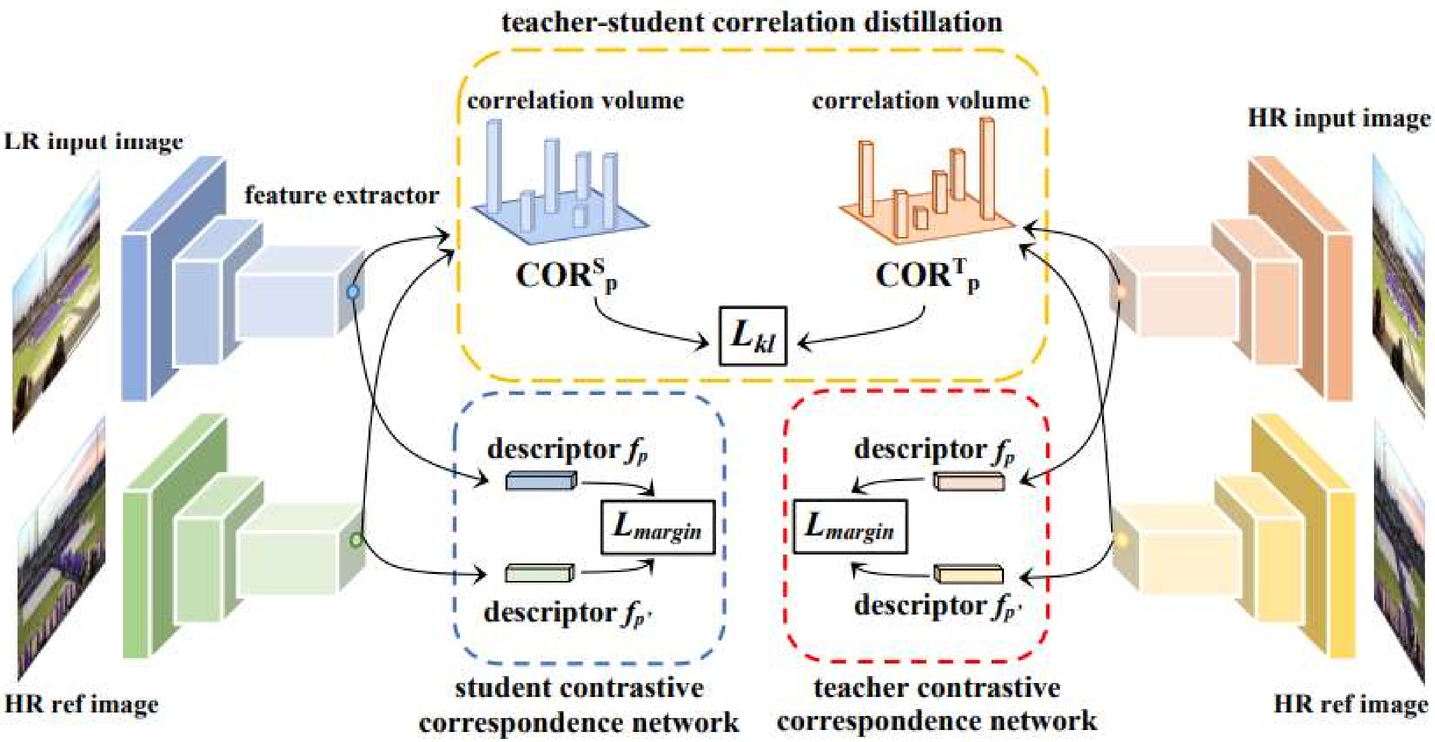
C^2- Matching contribution

→ Produce explicit robust matching crossing transformation and resolution

- Contrastive Correspondence network (for transformation gap)
  - Compute more robust to scale and rotation transformations
- Teacher-student correlation distillation (for resolution gap)
  - Boost the performance of LR – HR matching
- Dynamic Aggregation module for potential misalignment issue
- WR-SR dataset : New benchmark dataset


LR input image
SRNTT result
HR reference image
C2-Matching (ours) result

Design of C^2- Matching



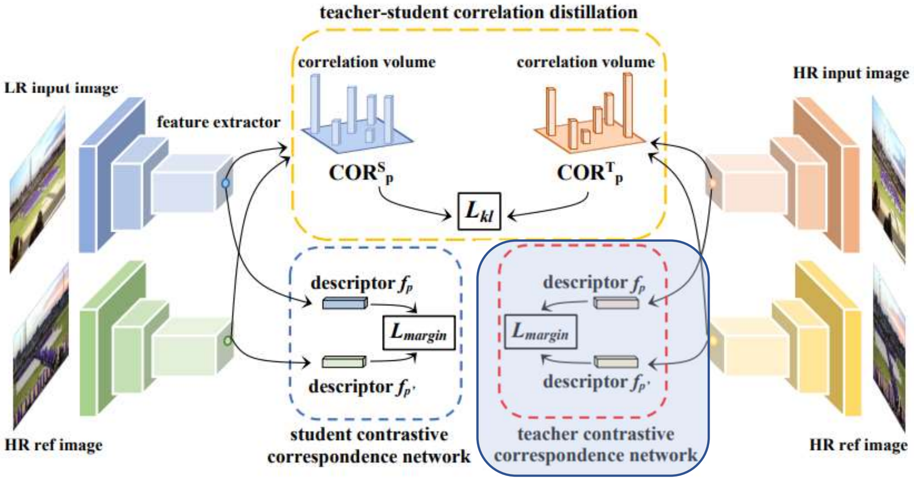
(a) Design of C^2-Matching

Contrastive Correspondence Network

→ Learns transformation-robust correspondence matching

- Synthesize HR reference images by applying homography transformation to original HR input images.
  - every position ***p*** in the LR input image ***I*,** we can compute its ground-truth correspondence point ***p'*** in the transformed image ***I'*** according to the homography transformation matrix.

- HR input, HR reference image description (e.g. VGGNet)
- LR input, HR reference image description + distillation



(a) Design of C²-Matching

Contrastive Correspondence Network

Triplet margin ranking loss

$$L_{margin} = \frac{1}{N} \sum_{p \in I} \max(0, m + \mathrm{Pos}(p) - \mathrm{Neg}(p)), \quad (1)$$

where N is the total number of points in image I and m is the margin value.

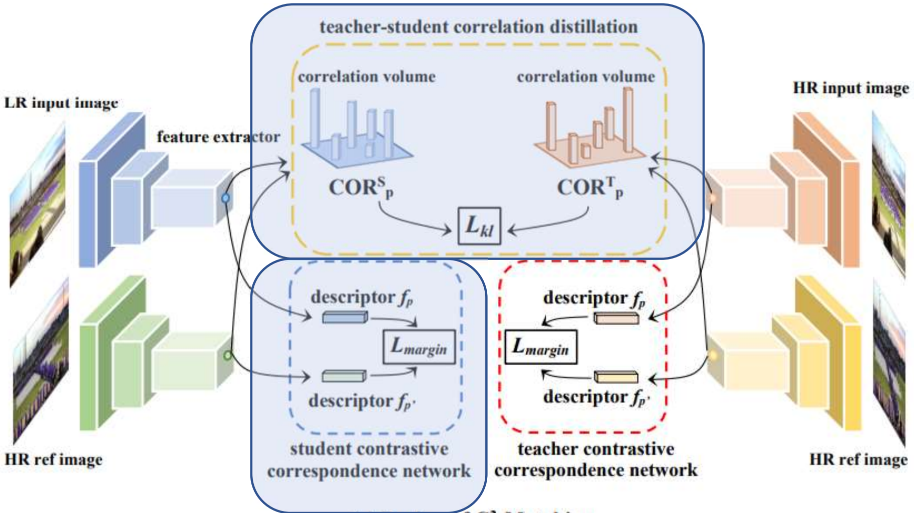
$$\mathrm{Pos}(p) = \| f_p - f_{p'} \|_2^2 . \quad (2)$$

$$\begin{aligned} \mathrm{Neg}(p) = \min(& \min_{k \in I', \|k - p'\|_\infty > T} \| f_p - f_k \|_2^2, \\ & \min_{k \in I, \|k - p\|_\infty > T} \| f_{p'} - f_k \|_2^2), \end{aligned} \quad (3)$$

Teacher-Student Correlation Distillation

→ HR – HR knowledge to LR – HR matching

- Since a lot of information is lost in LR input images, correspondence matching is difficult, especially for highly textured regions.
- Aim to transfer the matching ability of HR-HR matching to LR-HR matching



(a) Design of C^2-Matching

Teacher-Student Correlation Distillation

→ HR – HR knowledge to LR – HR matching

scriptors. By computing correlations between descriptors of input images and reference images, we can obtain an $N \times M$ matrix to represent the correlation volume, and view it as a probability distribution by applying a softmax function with temperature τ over it. To summarize, the correlation of the descriptor of input image at position p and the descriptor of reference image at position q is computed as follows:

$$\mathrm{cor}_{pq} = \frac{e^{\frac{f_p}{\|f_p\|} \cdot \frac{f_q}{\|f_q\|} / \tau}}{\sum_{k \in I'} e^{\frac{f_p}{\|f_p\|} \cdot \frac{f_k}{\|f_k\|} / \tau}}. \tag{4}$$

denote COR^T and COR^S as the teacher correlation volume and student correlation volume, respectively. For every descriptor p of input image, the divergence of teacher model's correlation and student model's correlation can be measured by Kullback Leibler divergence as follows:

$$\begin{aligned} \mathrm{Div}_p &= \mathrm{KL}(\mathrm{COR}_p^T || \mathrm{COR}_p^S) \\ &= \sum_{k \in I'} \mathrm{cor}_{pk}^T \, log(\frac{\mathrm{cor}_{pk}^T}{\mathrm{cor}_{pk}^S}). \end{aligned} \tag{5}$$

The correlation volume contains the knowledge of relationship between descriptors. By minimizing the divergence between two correlation volumes, the matching ability of teacher model can be transferred to the student model. This objective is defined as follows:

$$L_{kl} = \frac{1}{N} \sum_{p \in I} \mathrm{Div}_p. \tag{6}$$

With the teacher-student correlation distillation, the total loss used for training the contrastive correspondence network is:

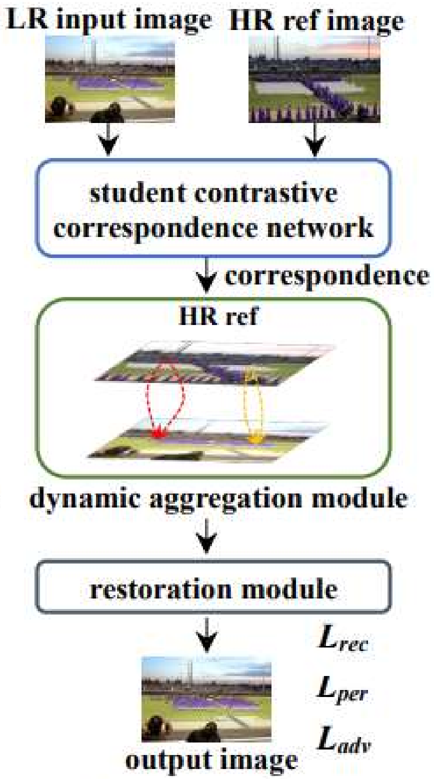
$$L = L_{margin} + \alpha_{kl} \cdot L_{kl}, \tag{7}$$

where α_{kl} is the weight for the KL-divergence loss.

Restoration module

→ dynamic aggregation module + restoration module

$$y(p) = \sum_{k=1}^{K} w_k \cdot x(p + p_0 + p_k + \Delta p_k) \cdot \Delta m_k,$$

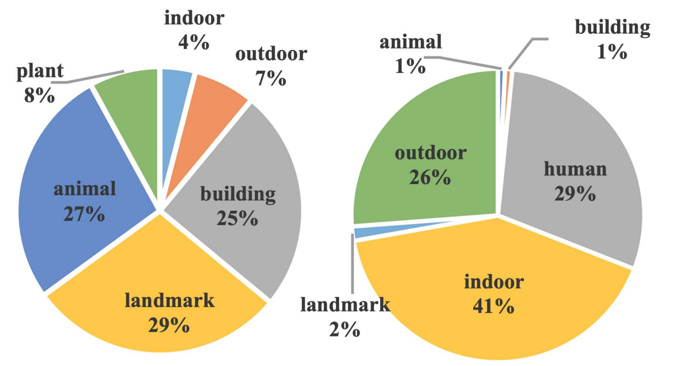


(b) Restoration Network

#	Layer name(s)
0	Conv(3, 64), LeakyReLU
1	RB [Conv(64, 64), ReLU, Conv(64, 64)] × 16
2	Concat [#1, Aggregated Reference Feature1]
3	Conv(320, 64), LeakyReLU
4	RB [Conv(64, 64), ReLU, Conv(64, 64)] × 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)] × 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)] × 16
15	ElementwiseAdd(#11, #14)
16	Conv(64, 32), LeakyReLU
17	Conv(32, 3)

Webly-Referenced SR Dataset

- The pair of input images and reference images are collected in a more realistic way.
- More diverse than CUFED5 datasets.

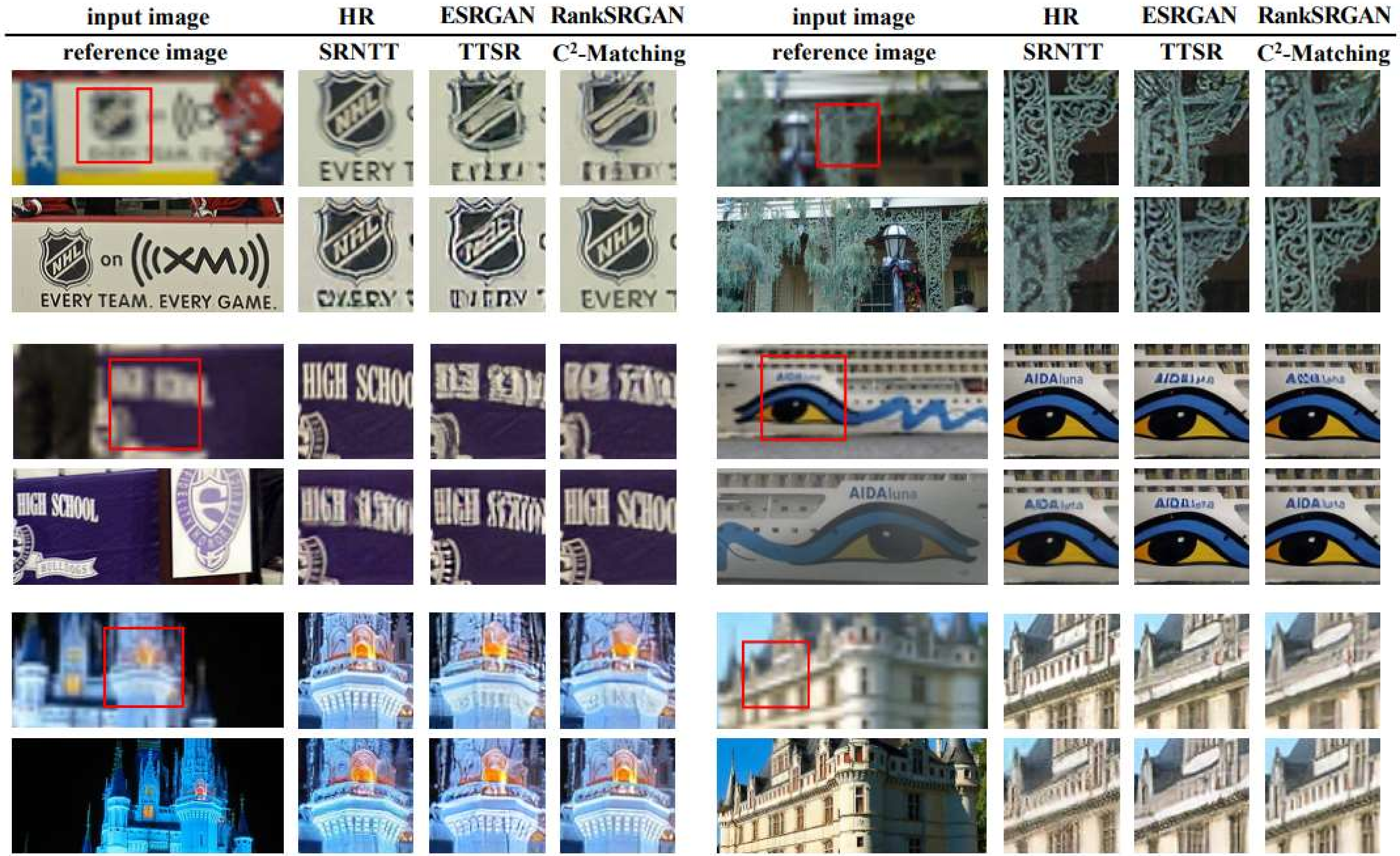


(b) WR-SR distribution (c) CUFED5 distribution

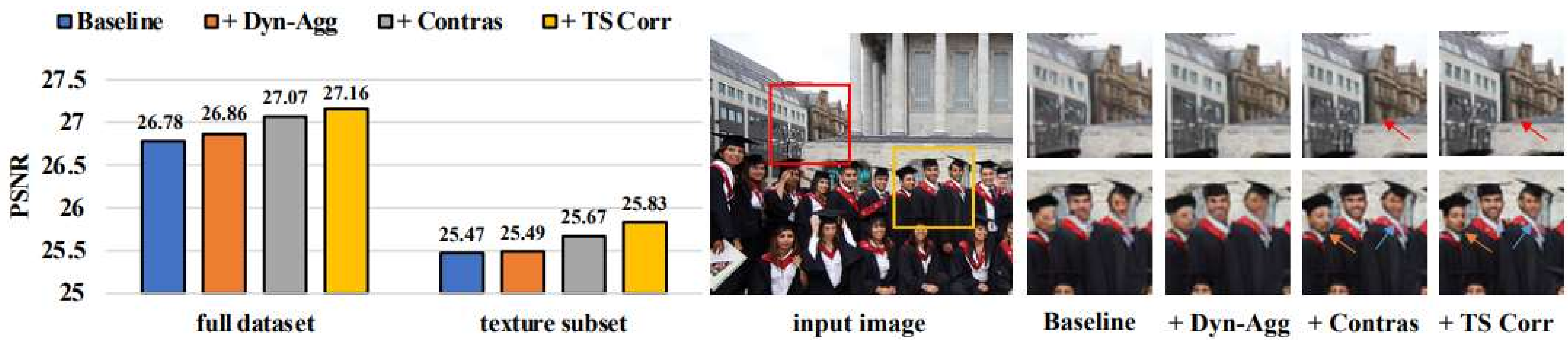
Experiments

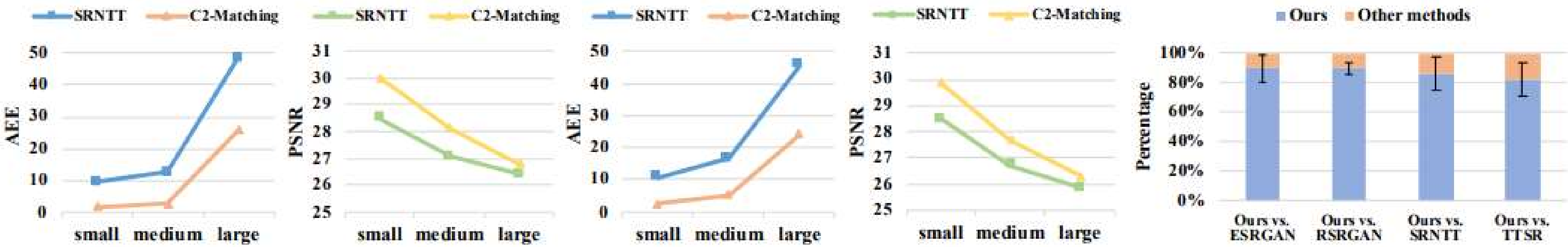
	Method	CUFED5	Sun80	Urban100	Manga109	WR-SR
SISR	SRCNN [6]	25.33 / .745	28.26 / .781	24.41 / .738	27.12 / .850	26.75 / .754
	EDSR [17]	25.93 / .777	28.52 / .792	25.51 / .783	28.93 / .891	27.36 / .773
	RCAN [36]	26.06 / .769	29.86 / .810	25.42 / .768	29.38 / .895	27.46 / .777
	SRGAN [15]	24.40 / .702	26.76 / .725	24.07 / .729	25.12 / .802	25.64 / .699
	ENet [23]	24.24 / .695	26.24 / .702	23.63 / .711	25.25 / .802	25.24 / .701
	ESRGAN [30]	21.90 / .633	24.18 / .651	20.91 / .620	23.53 / .797	25.37 / .691
	RankSRGAN [35]	22.31 / .635	25.60 / .667	21.47 / .624	25.04 / .803	25.98 / .722
Ref-SR	CrossNet [40]	25.48 / .764	28.52 / .793	25.11 / .764	23.36 / .741	-
	SRNTT	25.61 / .764	27.59 / .756	25.09 / .774	27.54 / .862	26.17 / .744
	SRNTT-*rec* [39]	26.24 / .784	28.54 / .793	25.50 / .783	28.95 / .885	27.21 / .775
	TTSR	25.53 / .765	28.59 / .774	24.62 / .747	28.70 / .886	26.50 / .762
	TTSR-*rec* [34]	27.09 / .804	30.02 / .814	25.87 / .784	30.09 / .907	27.75 / .794
	SSEN	25.35 / .742	-	-	-	-
	SSEN-*rec* [26]	26.78 / .791	-	-	-	-
	E2ENT^2	24.01 / .705	28.13 / .765	-	-	-
	E2ENT^2-*rec* [32]	24.24 / .724	28.50 / .789	-	-	-
	CIMR	26.16 / .781	29.67 / .806	25.24 / .778	-	-
	CIMR-*rec* [33]	26.35 / .789	30.07 / .813	25.77 / **.792**	-	-
Ours	C^2-Matching	27.16 / .805	29.75 / .799	25.52 / .764	29.73 / .893	27.54 / .780
	C^2-Matching-*rec*	**28.24 / .841**	**30.18 / .817**	**26.03** / .785	**30.47 / .911**	**28.07 / .802**

Experiments



Ablation study





(a) Robustness to Scale Transformation

(b) Robustness to Rotation Transformation

(c) User Study