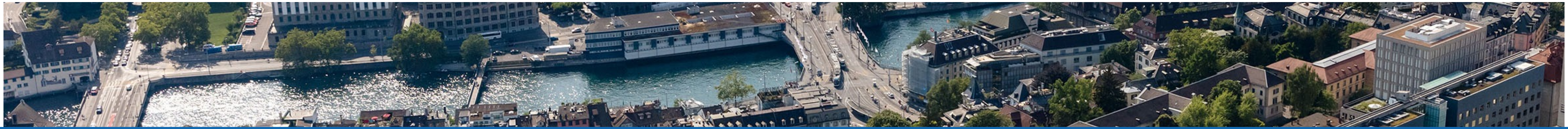
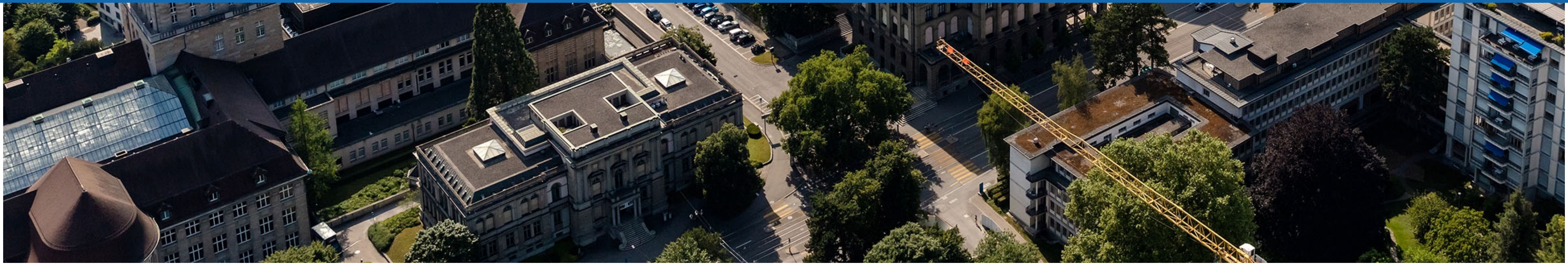
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA



CiaoSR: Continuous Implicit Attention-in-Attention Network for Arbitrary-Scale Image Super-Resolution

Jiezhang Cao, Qin Wang, Yongqin Xian, Yawei Li, Bingbing Ni, Zhiming Pi, Kai Zhang, Yulun Zhang, Radu Timofte, Luc Van Gool

TUE-AM-171

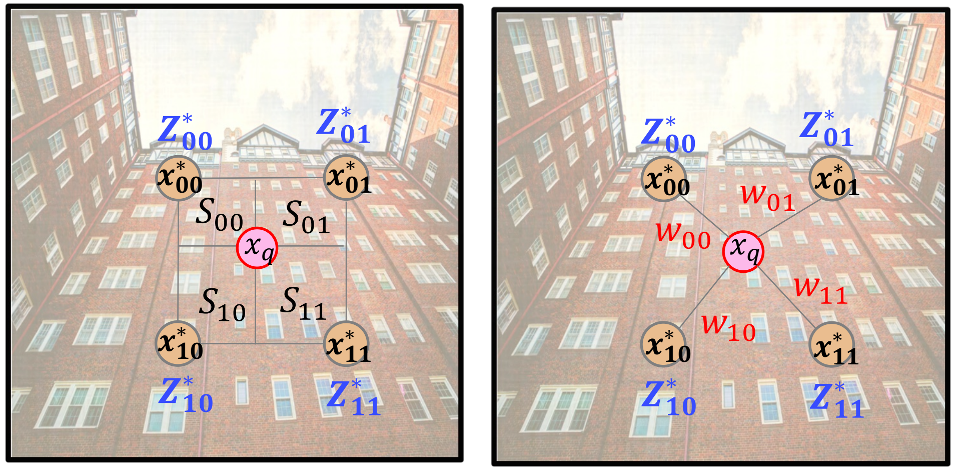


Motivation

- The RGB value at $\boldsymbol{x_q}$ can be predicted by directly ensembling its neighborhood information

$$I(\boldsymbol{x_q}) = \sum_{(i,j)\in\mathcal{I}} w_{i,j} \cdot f(\boldsymbol{Z}^*_{i,j}, \boldsymbol{x_q} - \boldsymbol{x}^*_{i,j})$$

- $w_{i,j}$ = bilinear interpolation
- no learnable parameters
- neglects the similarity of features
- $\boldsymbol{Z}^*_{i,j}$ has only neighboring features

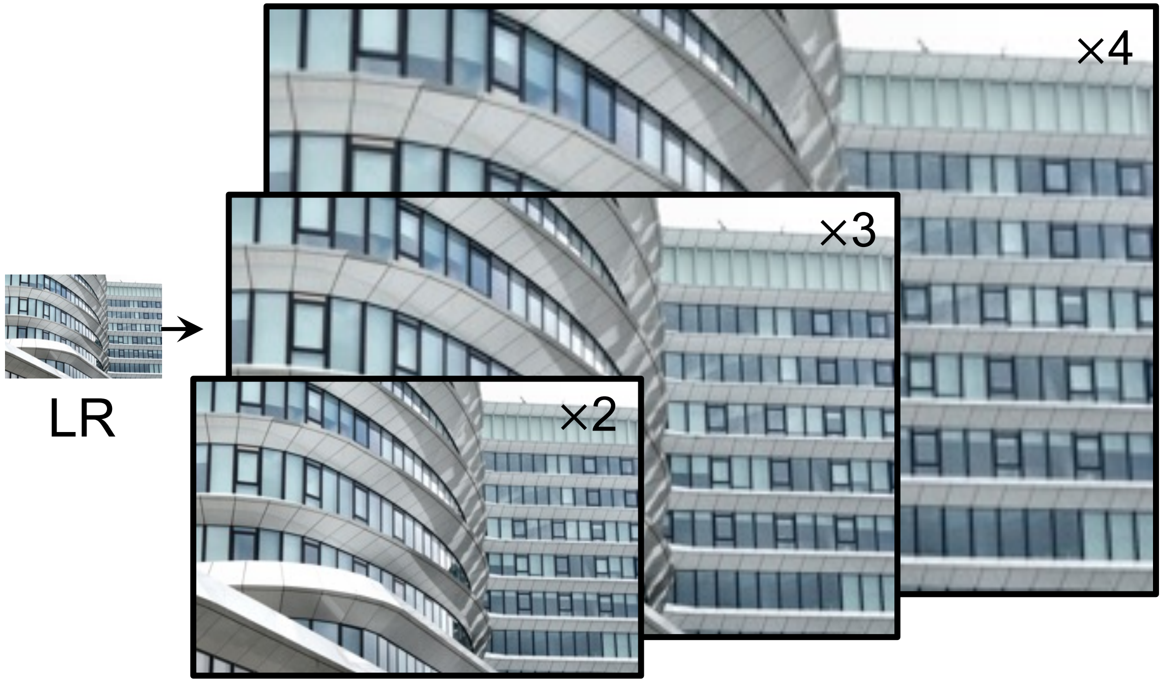


LIIF: $w_{i,j} = \frac{S_t}{\sum_{t'} S_{t'}}$ $\qquad$ $w_{i,j} = ?$

Forecast

This work designs a new implicit model to generate continuous scale images

Classical image SR

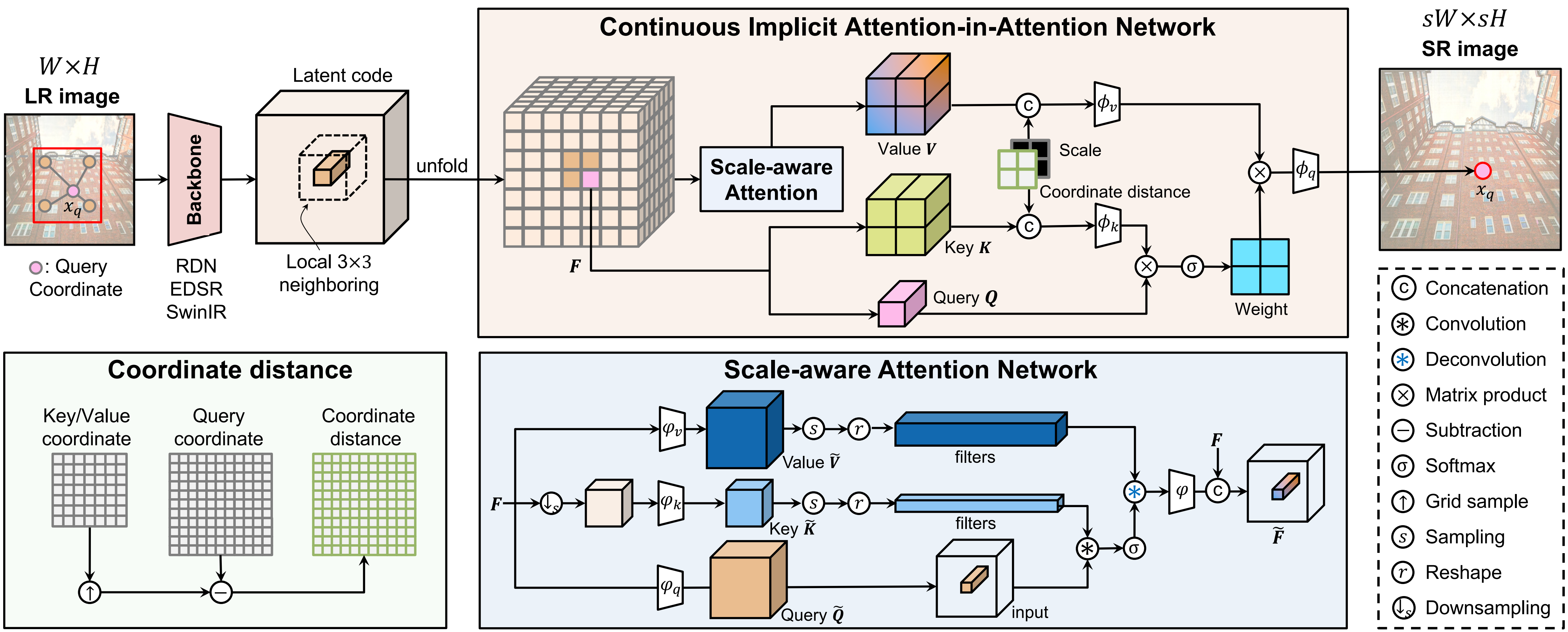


Discrete SR images

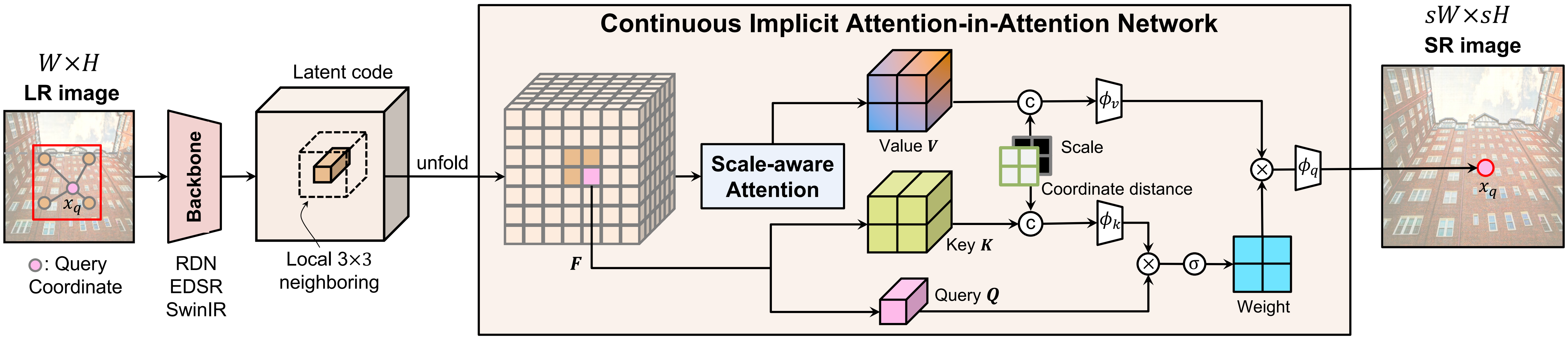
Arbitrary-scale image SR

Continuous SR images

Continuous Implicit Attention-in-Attention Network



Continuous Implicit Attention-in-Attention Network



Given pairs of coordinates and latent codes, we predict RGB values at the given query

$$I_q = \phi_q\left(\sum_{t \in \mathcal{T}(x_q)} \sigma(\boldsymbol{Q}^{\mathrm{T}} \boldsymbol{K}_t) \boldsymbol{V}_t\right)$$

Query, Key and Value:

$$\begin{cases} \boldsymbol{Q} = \boldsymbol{F}^* \\ \boldsymbol{K} = \phi_k([\boldsymbol{F}_{i,j}, (\boldsymbol{r}_k)_{i,j}, \boldsymbol{s}]) \\ \boldsymbol{V} = \phi_v\left(\left[[\boldsymbol{F}_{i,j}, \widetilde{\boldsymbol{F}}_{i,j}], (\boldsymbol{r}_v)_{i,j}, \boldsymbol{s}\right]\right) \end{cases}$$

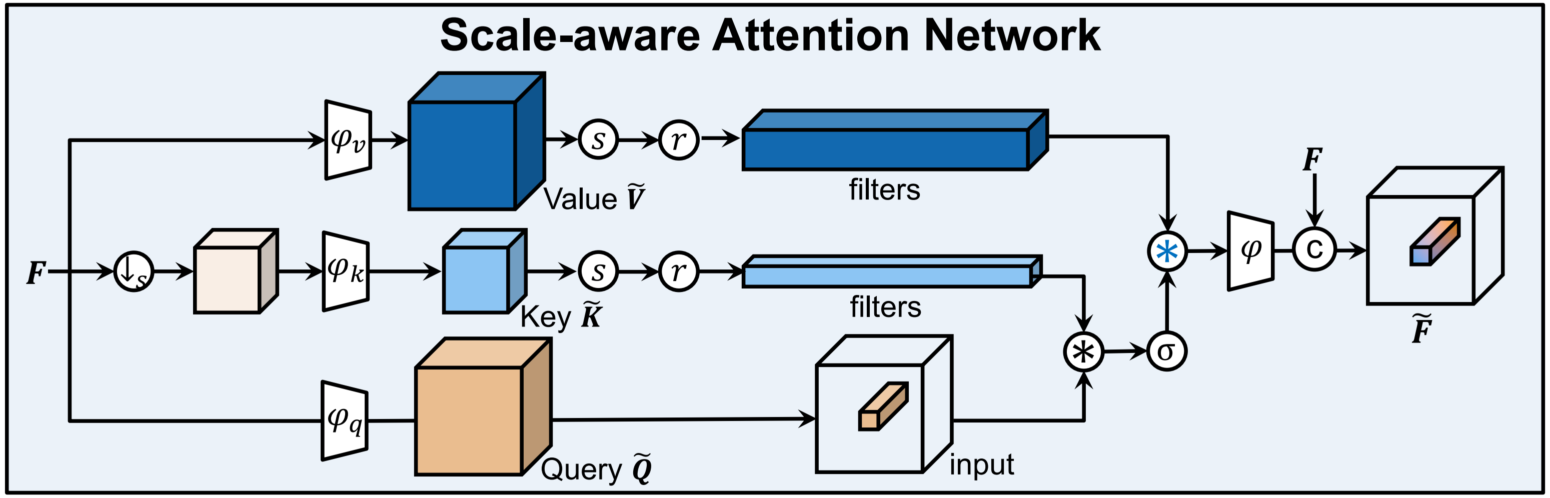
Scale-aware Attention Network

Given a local feature $\boldsymbol{F}$, we first downsample $\boldsymbol{F}$ with smaller scale, then calculate non-local features,

$$\widetilde{\boldsymbol{F}}_{i,j} = \varphi\left(\sum_{u,v} \frac{\exp\left(\widetilde{\boldsymbol{Q}}_{i,j}^{\mathrm{T}} \widetilde{\boldsymbol{K}}_{u,v}\right)}{\sum_{u',v'} \exp\left(\widetilde{\boldsymbol{Q}}_{i,j}^{\mathrm{T}} \widetilde{\boldsymbol{K}}_{u',v'}\right)} \widetilde{\boldsymbol{V}}_{s'u,\, s'v}^{s'p \times s'p}\right)$$

Query, Key and Value for non-local features:

$$\begin{cases} \widetilde{\boldsymbol{Q}} = \varphi_k(\boldsymbol{F}) \\ \widetilde{\boldsymbol{K}} = \varphi_k\left(\boldsymbol{F}_{\downarrow_s}\right) \\ \widetilde{\boldsymbol{V}} = \varphi_v(\boldsymbol{F}) \end{cases}$$



Experiment Results

- **Datasets:**
  - Training set: DIV2K (with continuous scales [1, 4])
  - Testing set: Set5, Set14, B100, Urban100, Manga109
- **Backbones:**
  - RDN, SwinIR
- **Compared methods:**
  - MetaSR, LIIF, ITSRN, LTE

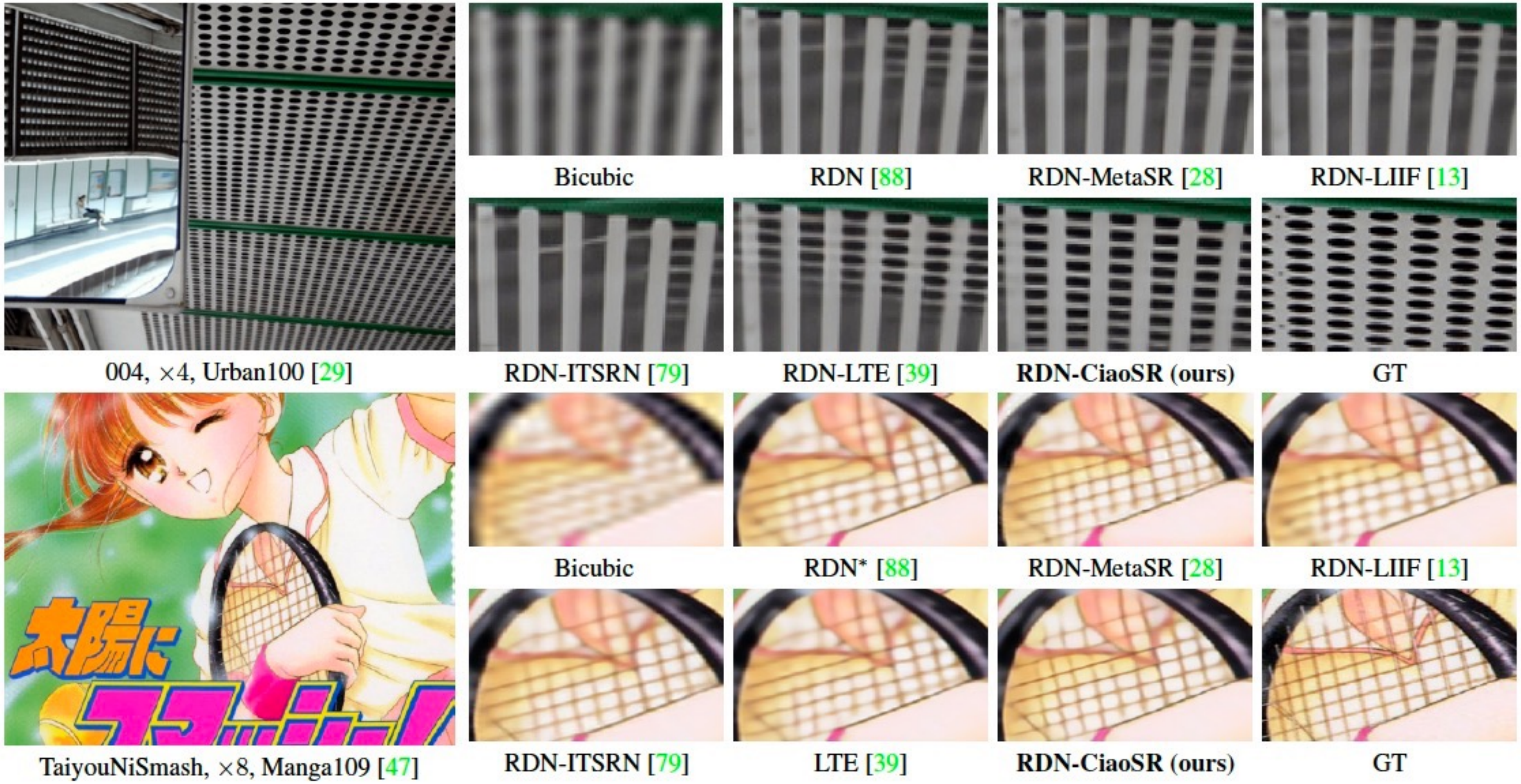
Quantitative Results

Methods	Set5 [3]			Set14 [80]			B100 [46]			Urban100 [29]			Manga109 [47]		
	×2	×3	×4	×2	×3	×4	×2	×3	×4	×2	×3	×4	×2	×3	×4
RDN [88]	38.24	34.71	32.47	34.01	30.57	28.81	32.34	29.26	27.72	32.89	28.80	26.61	39.18	34.13	31.00
RDN-MetaSR [28]	38.22	34.63	32.38	33.98	30.54	28.78	32.33	29.26	27.71	32.92	28.82	26.55	-	-	-
RDN-LIIF [13]	38.17	34.68	32.50	33.97	30.53	28.80	32.32	29.26	27.74	32.87	28.82	26.68	39.26	34.21	31.20
RDN-ITSRN† [79]	38.23	34.76	32.55	34.19	30.59	28.88	32.38	29.32	27.79	33.07	28.96	26.77	39.34	34.39	31.37
RDN-LTE [39]	38.23	34.72	32.61	34.09	30.58	28.88	32.36	29.30	27.77	33.04	28.97	26.81	39.28	34.32	31.30
RDN-CiaoSR (ours)	**38.29**	**34.85**	**32.66**	**34.22**	**30.65**	**28.93**	**32.41**	**29.34**	**27.83**	**33.30**	**29.17**	**27.11**	**39.51**	**34.57**	**31.57**
SwinIR [40]	38.35	34.89	32.72	34.14	30.77	28.94	32.44	29.37	27.83	33.40	29.29	27.07	39.60	34.74	31.67
SwinIR-MetaSR [28]	38.26	34.77	32.47	34.14	30.66	28.85	32.39	29.31	27.75	33.29	29.12	26.76	39.46	34.62	31.37
SwinIR-LIIF [13]	38.28	34.87	32.73	34.14	30.75	28.98	32.39	29.34	27.84	33.36	29.33	27.15	39.57	34.68	31.71
SwinIR-ITSRN† [79]	38.22	34.75	32.63	34.26	30.75	28.97	32.42	29.38	27.85	33.46	29.34	27.12	39.60	34.75	31.74
SwinIR-LTE [39]	38.33	34.89	32.81	34.25	30.80	29.06	32.44	29.39	27.86	33.50	29.41	27.24	39.63	34.79	31.79
SwinIR-CiaoSR (ours)	**38.38**	**34.91**	**32.84**	**34.33**	**30.82**	**29.08**	**32.47**	**29.42**	**27.90**	**33.65**	**29.52**	**27.42**	**39.67**	**34.84**	**31.91**

Methods	Set5 [3]			Set14 [80]			B100 [46]			Urban100 [29]			Manga109 [47]		
	×6	×8	×12	×6	×8	×12	×6	×8	×12	×6	×8	×12	×6	×8	×12
RDN-MetaSR [28]	29.04	29.96	-	26.51	24.97	-	25.90	24.83	-	23.99	22.59	-	-	-	-
RDN-LIIF [13]	29.15	27.14	24.86	26.64	25.15	23.24	25.98	24.91	23.57	24.20	22.79	21.15	27.33	25.04	22.36
RDN-ITSRN† [79]	29.32	27.25	24.86	26.68	25.17	23.28	26.01	24.93	23.58	24.23	22.81	21.16	27.45	25.04	23.35
RDN-LTE [39]	29.32	27.26	24.79	26.71	25.16	23.31	26.01	24.95	23.60	24.28	22.88	21.22	27.49	25.12	22.43
RDN-CiaoSR (ours)	**29.46**	**27.36**	**24.92**	**26.79**	**25.28**	**23.37**	**26.07**	**25.00**	**23.64**	**24.58**	**23.13**	**21.42**	**27.70**	**25.40**	**22.63**
SwinIR-MetaSR [28]	29.09	27.02	24.82	26.58	25.09	23.33	25.94	24.87	23.59	24.16	22.75	21.31	27.29	24.96	22.35
SwinIR-LIIF [13]	29.46	27.36	-	26.82	25.34	-	26.07	25.01	-	24.59	23.14	-	27.69	25.28	-
SwinIR-ITSRN† [79]	29.31	27.24	24.79	26.71	25.32	23.30	26.05	24.96	23.57	24.50	23.06	21.34	27.72	25.23	22.47
SwinIR-LTE [39]	29.50	27.35	-	26.86	**25.42**	-	26.09	25.03	-	24.62	23.17	-	27.83	25.42	-
SwinIR-CiaoSR (ours)	**29.62**	**27.45**	**24.96**	**26.88**	**25.42**	**23.38**	**26.13**	**25.07**	**23.68**	**24.84**	**23.34**	**21.60**	**28.01**	**25.61**	**22.79**

- CiaoSR achieves the best performance with all backbones on both in-scale and out-of-scale distributions

Qualitative comparison



- Our model is able to synthesize the SR images with sharper textures than other methods

Ablation Study

Table 4. Ablation study on each component of our networks on Urban100. We use RDN [88] as the backbone.

Attention-in-attention		✗	✓	✓
Scale-aware Attention Network		✗	✗	✓
In-scale	×2	32.87	33.24	**33.30**
	×3	28.82	29.10	**29.17**
	×4	26.69	26.96	**27.11**
Out-of-scale	×6	24.22	24.50	**24.58**
	×8	22.80	22.98	**23.13**

Table 5. Ablation study on training our implicit model with different types of scales on Urban100.

Type	Training scale s	In-scale			Out-of-scale	
		×2	×3	×4	×6	×8
Discrete	$s \in \{2\}$	33.13	27.01	25.60	22.27	22.09
	$s \in \{3\}$	31.39	29.06	25.77	23.44	22.16
	$s \in \{4\}$	31.42	27.87	26.88	24.28	22.85
	$s \in \{2, 3, 4\}$	33.15	29.14	27.02	24.47	23.03
Continuous	$s \in [1, 4]$	**33.30**	**29.17**	**27.11**	**24.58**	**23.13**

Table 6. Comparison of (PSNR and SSIM) for different synthesis steps on Urban100 [29] and Manga109 [47].

Type	Synthesis steps	Urban100 [29]		Manga109 [47]	
		PSNR	SSIM	PSNR	SSIM
Multiple steps	→×2→×4→×12	21.28	0.557	22.46	0.720
	→×2→×12	21.32	0.558	22.53	0.721
One step	→×12	**21.42**	**0.561**	**22.63**	**0.723**

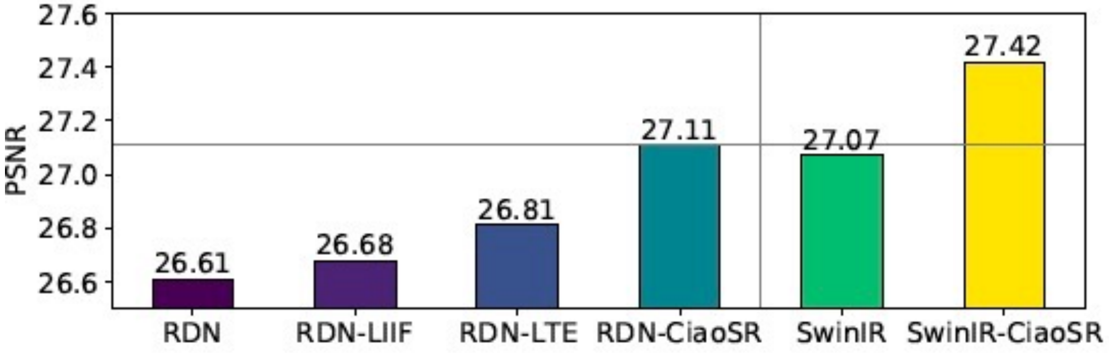
Table 7. Comparisons of model size, inference time and performance gain of different models.

Different models	Meta-SR [28]	LIIF [13]	ITSRN [79]	LTE [39]	CiaoSR
Model size (M)	1.7	1.6	0.7	1.7	1.4
Inference time (ms)	237	171	343	148	528
PSNR (dB)	26.55	26.68	26.77	26.81	27.11
Performance gain (dB)	-0.06	0.07	0.16	0.2	**0.5**

- Training with continuous scales can boost the performance
- Synthesis with one step is better than more steps
- Best performance, but with more inference time

Summary

- **New architecture:**
  - Propose a novel continuous implicit attention-in-attention network for arbitrary-scale image super-resolution
- **Best performance:**
  - Outperform all state-of-the-art methods
- **Good generalization ability:**
  - Generalize well on both in-scale and out-of-scale distributions



- **Good flexibility and applicability:**
  - Can be used behind any SR backbone to boost the performance

Thanks for your attention!