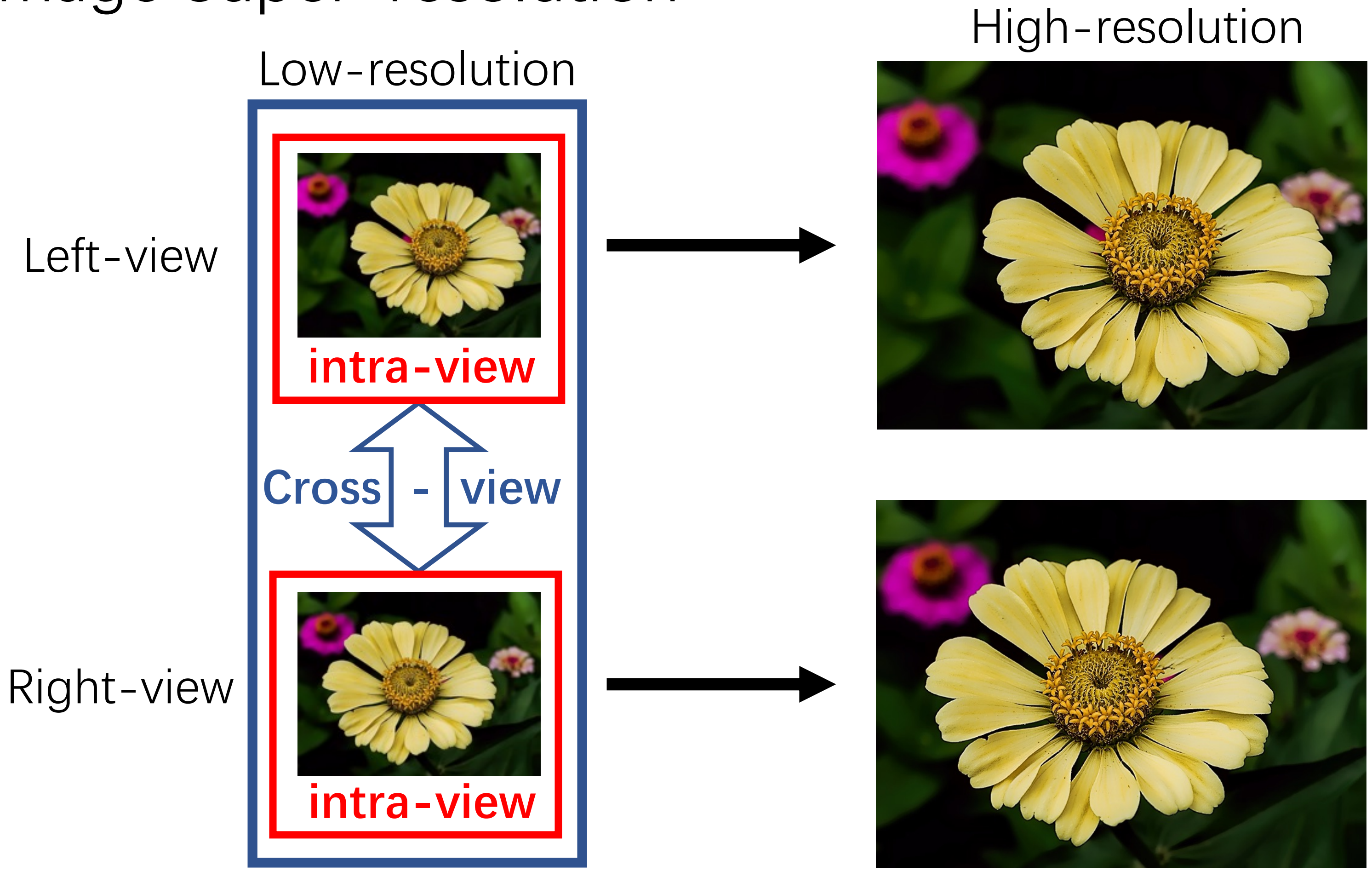
NAFSSR: Stereo Image Super-Resolution Using NAFNet

Xiaojie Chu* Liangyu Chen* Wenqing Yu

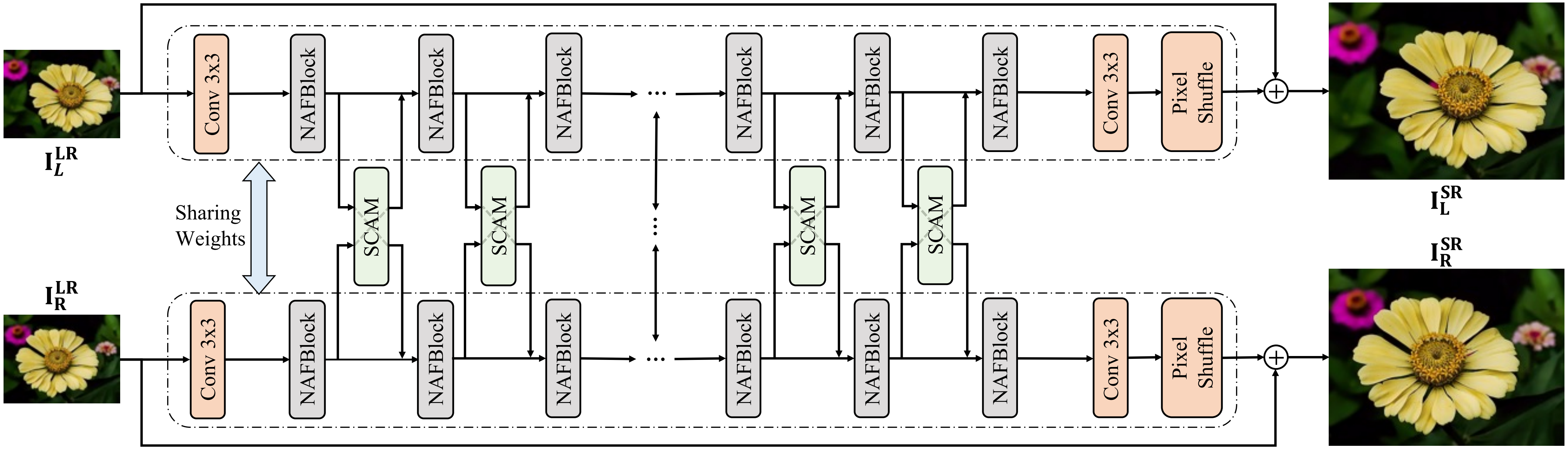
Background

- Stereo Image Super-resolution



Overview

➢NAFSSR

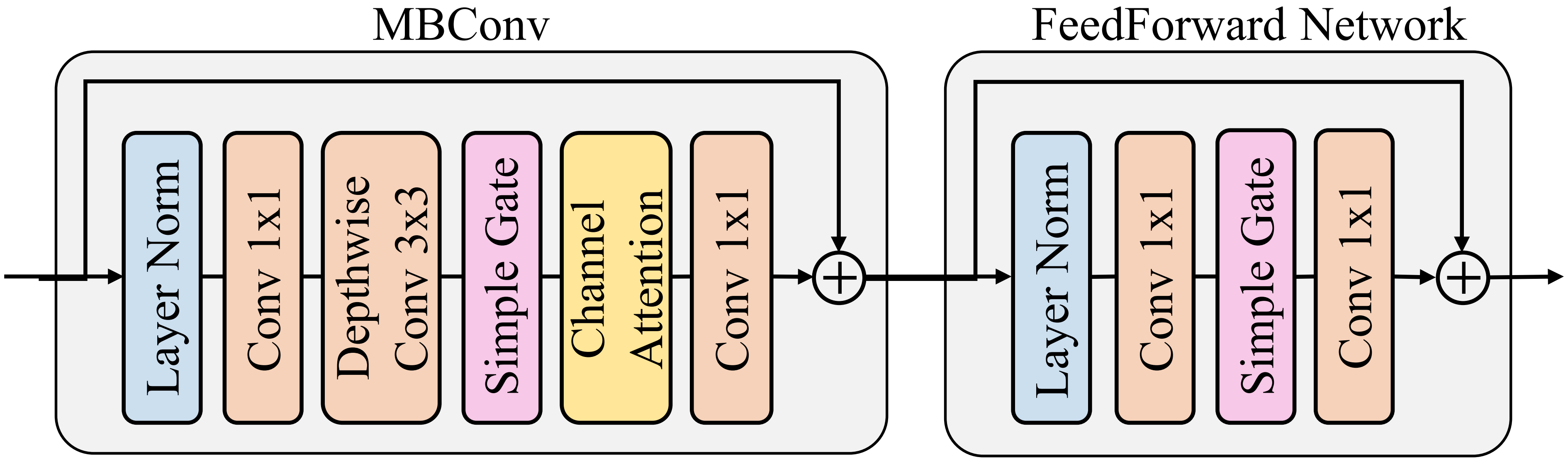


NAFBlock: Blocks from NAFNet

SCAM: Stereo Cross-Attention Modules

Architectures

- ➢NAFBlock [1]



SimpleGate(x) = x1*x2, where [x1,x2] = split(x, dim=channel)

[1] *Chen, Liangyu, et al. "Simple baselines for image restoration." arXiv preprint arXiv:2204.04676 (2022).*

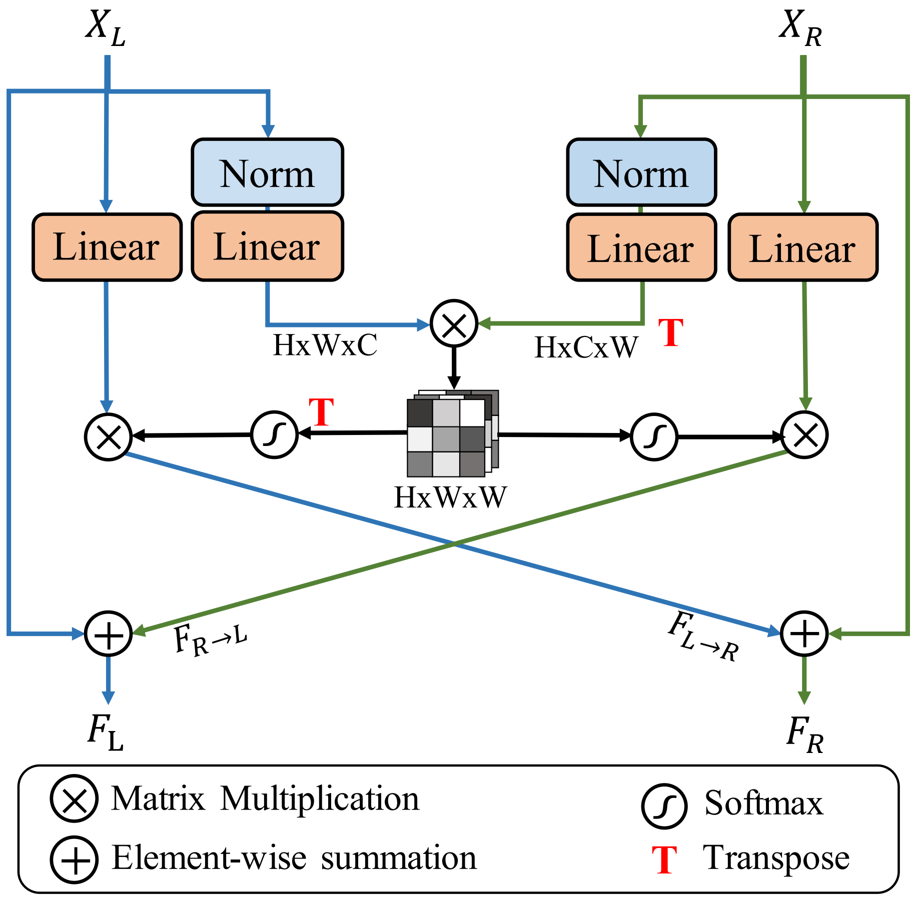
Architectures

➢Stereo Cross Attention Module (SCAM)

- Scaled dot-Product Attention
  - $\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{C})\boldsymbol{V}$

- Bidirectional Cross Attention
  - $F_{R\longrightarrow L} = \text{Attention}(\boldsymbol{W}_1^L\overline{\boldsymbol{X}_L}, \boldsymbol{W}_1^R\overline{\boldsymbol{X}_R}, \boldsymbol{W}_2^R\overline{\boldsymbol{X}_R})$
  - $F_{L\longrightarrow R} = \text{Attention}(\boldsymbol{W}_1^R\overline{\boldsymbol{X}_R}, \boldsymbol{W}_1^L\overline{\boldsymbol{X}_L}, \boldsymbol{W}_2^L\overline{\boldsymbol{X}_L})$

- Fusion
  - $F_L = \gamma_L F_{R\longrightarrow L} + X_L$
  - $F_R = \gamma_R F_{L\longrightarrow R} + X_R$



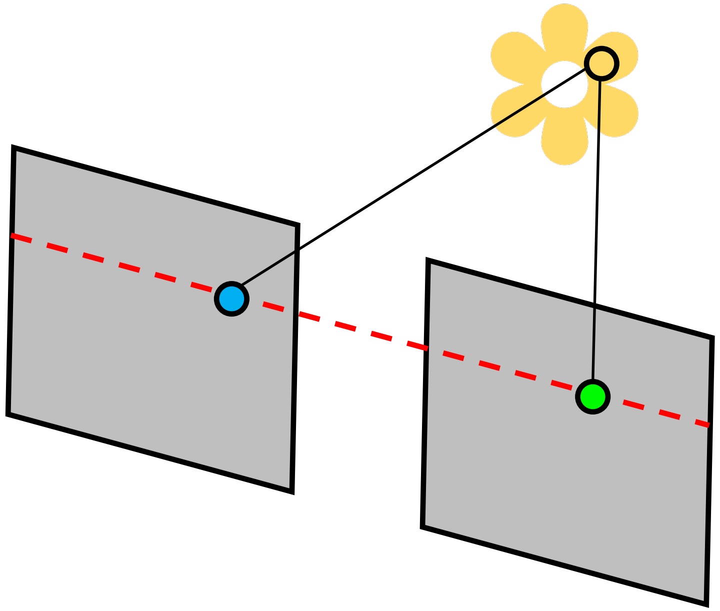
Architectures

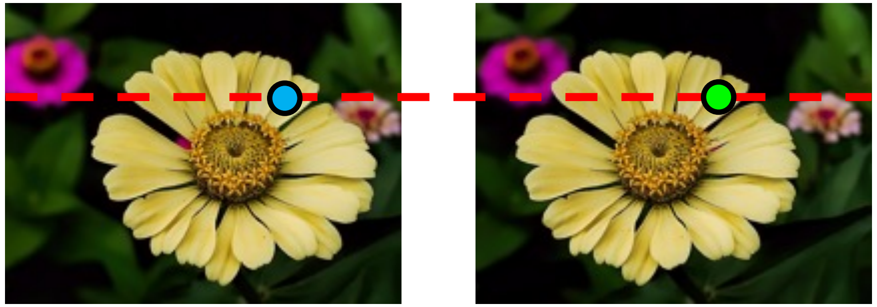
➢Stereo Cross Attention Module (SCAM)

- Scaled dot-Product Attention
  - $\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{C})\boldsymbol{V}$

- Bidirectional Cross Attention
  - $F_{R\longrightarrow L} = \text{Attention}(\boldsymbol{W}_1^L\overline{\boldsymbol{X}_L},\ \boldsymbol{W}_1^R\overline{\boldsymbol{X}_R},\ \boldsymbol{W}_2^R\overline{\boldsymbol{X}_R})$
  - $F_{L\longrightarrow R} = \text{Attention}(\boldsymbol{W}_1^R\overline{\boldsymbol{X}_R},\ \boldsymbol{W}_1^L\overline{\boldsymbol{X}_L},\ \boldsymbol{W}_2^L\overline{\boldsymbol{X}_L})$

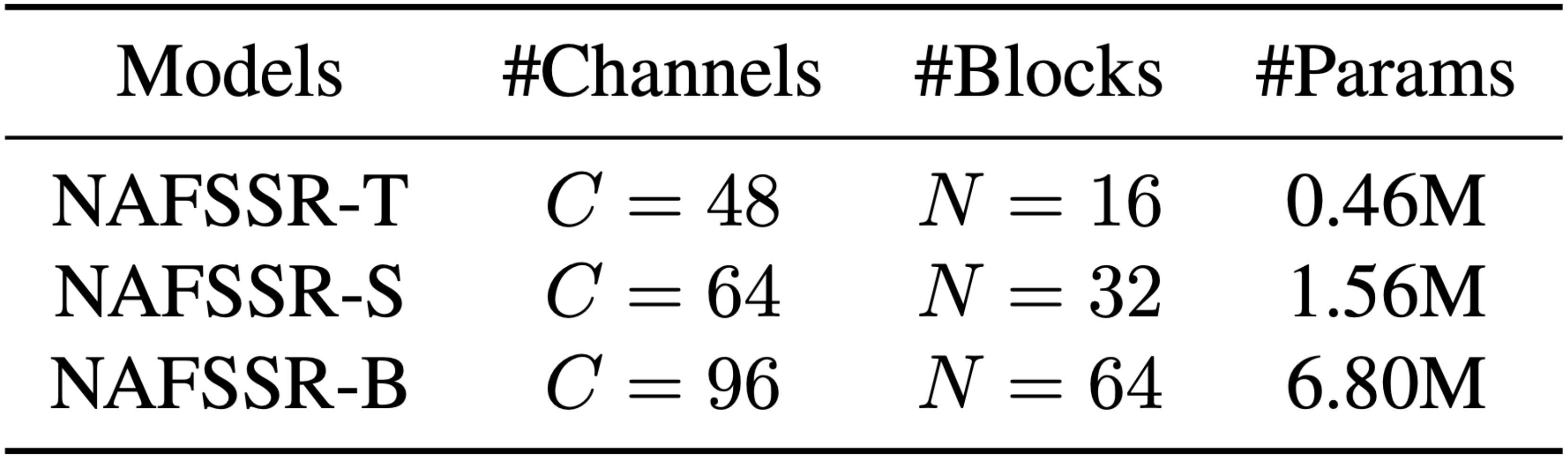
  - Attends to corresponding features along the **horizontal** epipolar line
  - since image pairs has **horizontal** disparities only





Architectures

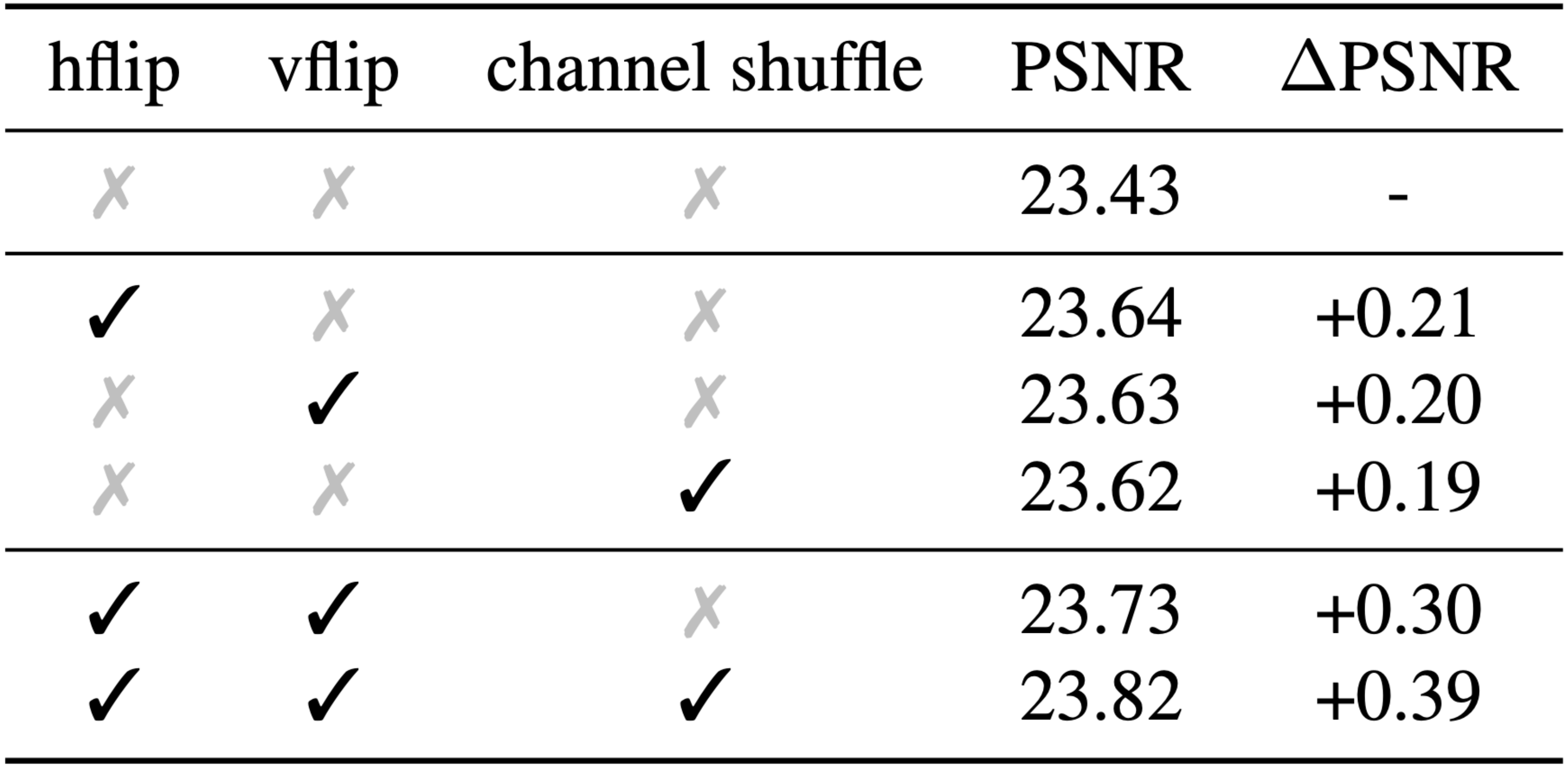
- NAFSSR Family



Models	#Channels	#Blocks	#Params
NAFSSR-T	$C = 48$	$N = 16$	0.46M
NAFSSR-S	$C = 64$	$N = 32$	1.56M
NAFSSR-B	$C = 96$	$N = 64$	6.80M

Tricks

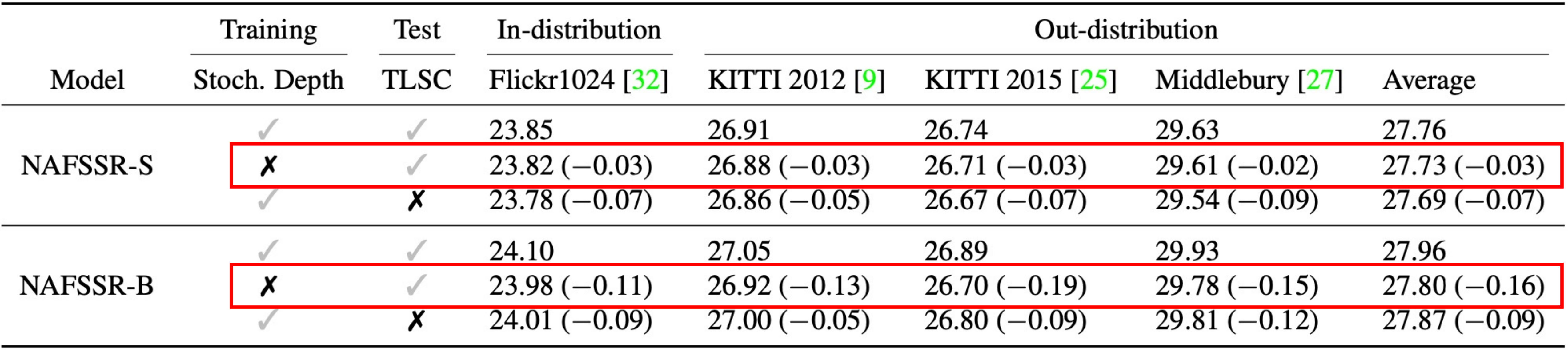
- Data Augmentation



hflip	vflip	channel shuffle	PSNR	ΔPSNR
✗	✗	✗	23.43	-
✓	✗	✗	23.64	+0.21
✗	✓	✗	23.63	+0.20
✗	✗	✓	23.62	+0.19
✓	✓	✗	23.73	+0.30
✓	✓	✓	23.82	+0.39

Tricks

- Stochastic depth [2] for better generality.

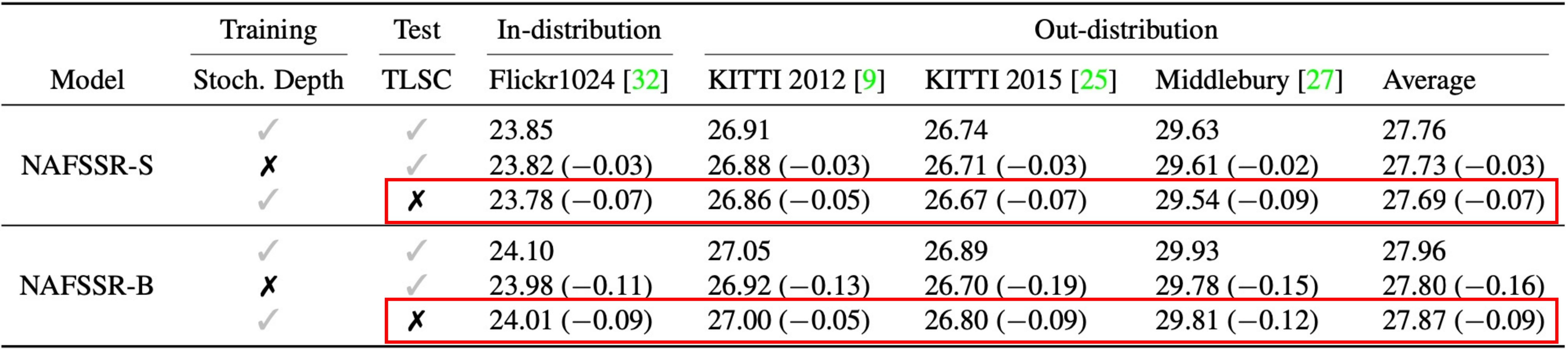


	Training	Test	In-distribution	Out-distribution			
Model	Stoch. Depth	TLSC	Flickr1024 [32]	KITTI 2012 [9]	KITTI 2015 [25]	Middlebury [27]	Average
NAFSSR-S	✓	✓	23.85	26.91	26.74	29.63	27.76
	✗	✓	23.82 (−0.03)	26.88 (−0.03)	26.71 (−0.03)	29.61 (−0.02)	27.73 (−0.03)
	✓	✗	23.78 (−0.07)	26.86 (−0.05)	26.67 (−0.07)	29.54 (−0.09)	27.69 (−0.07)
NAFSSR-B	✓	✓	24.10	27.05	26.89	29.93	27.96
	✗	✓	23.98 (−0.11)	26.92 (−0.13)	26.70 (−0.19)	29.78 (−0.15)	27.80 (−0.16)
	✓	✗	24.01 (−0.09)	27.00 (−0.05)	26.80 (−0.09)	29.81 (−0.12)	27.87 (−0.09)

[2] *Huang, Gao, et al. "Deep networks with stochastic depth." European conference on computer vision. Springer, Cham, 2016.*

Tricks

- Train-test Inconsistency: Patches vs. Image
- Inference: Test-time Local Statistics Converter (TLSC) [3]

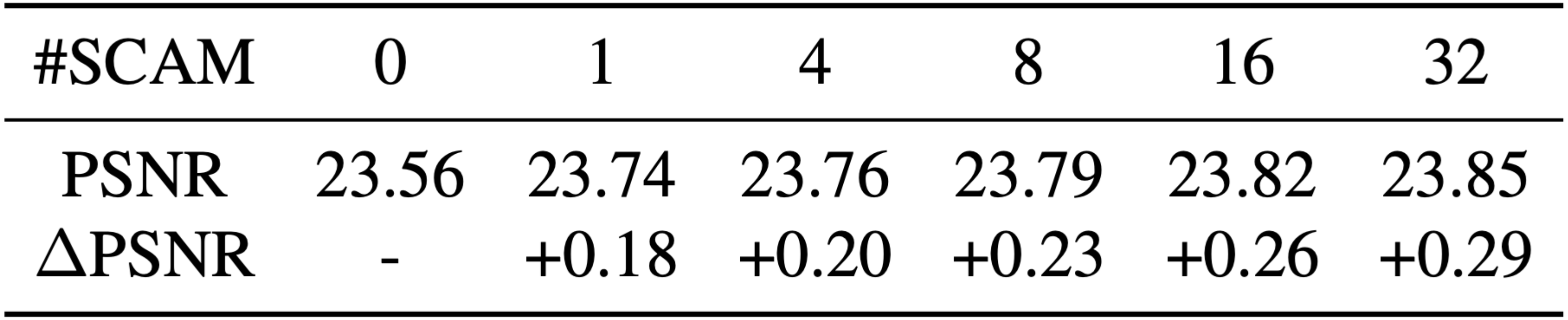


	Training	Test	In-distribution	Out-distribution			
Model	Stoch. Depth	TLSC	Flickr1024 [32]	KITTI 2012 [9]	KITTI 2015 [25]	Middlebury [27]	Average
NAFSSR-S	✓	✓	23.85	26.91	26.74	29.63	27.76
	✗	✓	23.82 (−0.03)	26.88 (−0.03)	26.71 (−0.03)	29.61 (−0.02)	27.73 (−0.03)
	✓	✗	23.78 (−0.07)	26.86 (−0.05)	26.67 (−0.07)	29.54 (−0.09)	27.69 (−0.07)
NAFSSR-B	✓	✓	24.10	27.05	26.89	29.93	27.96
	✗	✓	23.98 (−0.11)	26.92 (−0.13)	26.70 (−0.19)	29.78 (−0.15)	27.80 (−0.16)
	✓	✗	24.01 (−0.09)	27.00 (−0.05)	26.80 (−0.09)	29.81 (−0.12)	27.87 (−0.09)

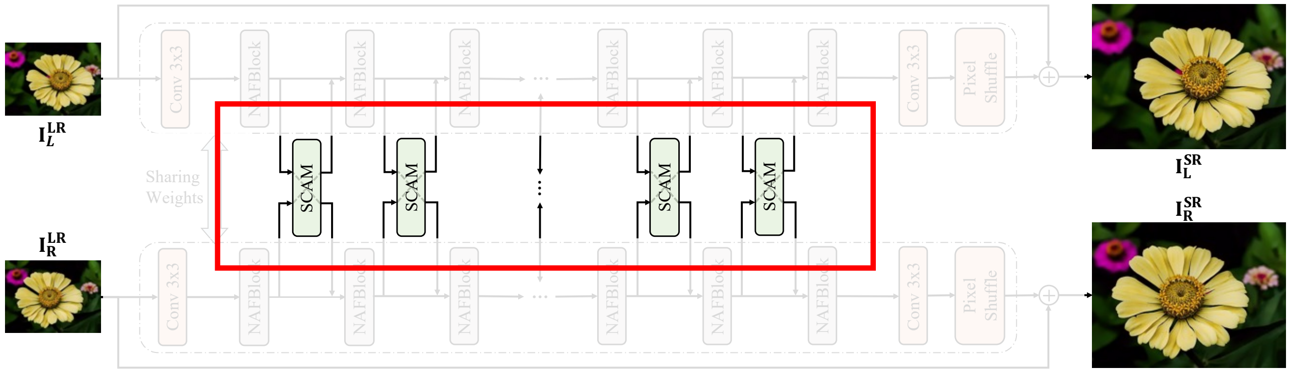
[3] *Chu, Xiaojie, et al. "Revisiting Global Statistics Aggregation for Improving Image Restoration." arXiv preprint arXiv:2112.04491 (2021).*

Results

- More Stereo Cross Attention Modules (SCAM), better results

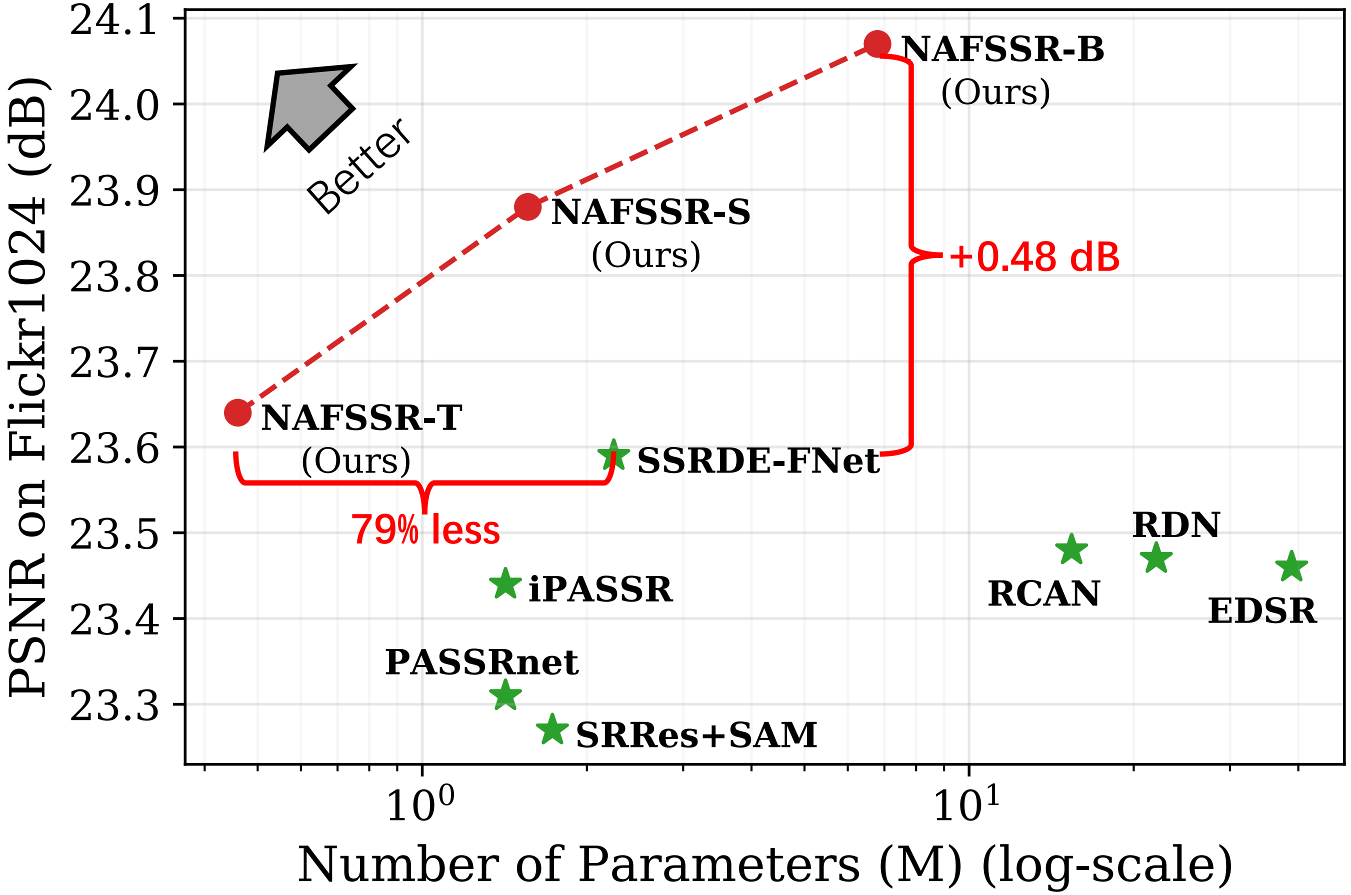


#SCAM	0	1	4	8	16	32
PSNR	23.56	23.74	23.76	23.79	23.82	23.85
ΔPSNR	-	+0.18	+0.20	+0.23	+0.26	+0.29



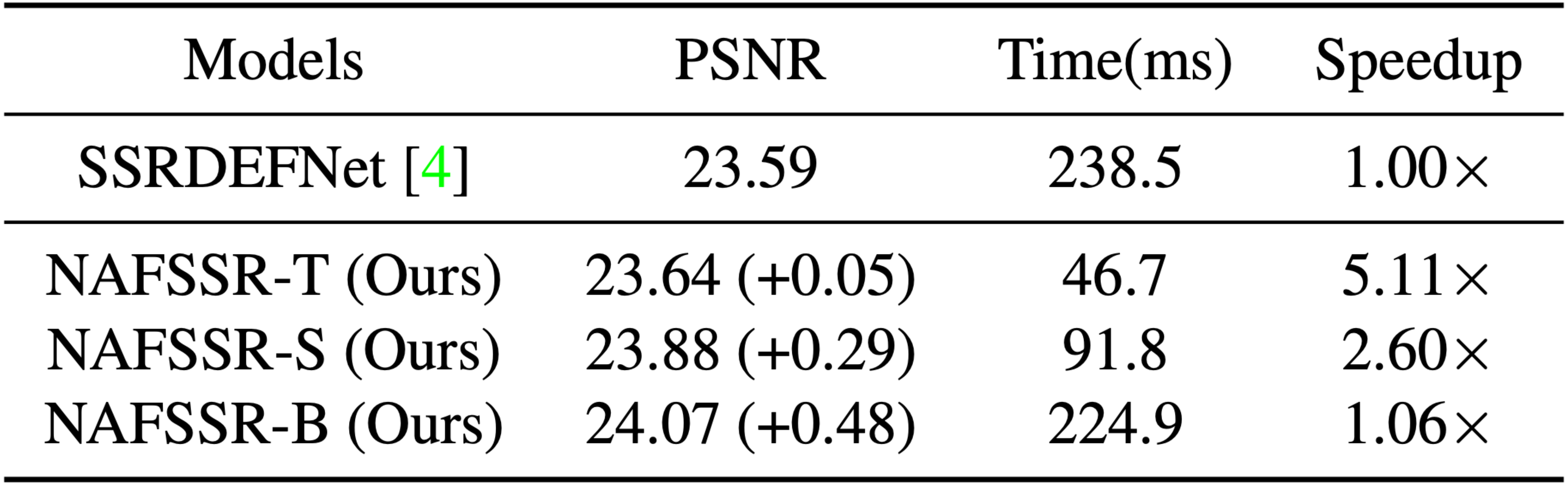
Results

- #Parameters vs. PSNR



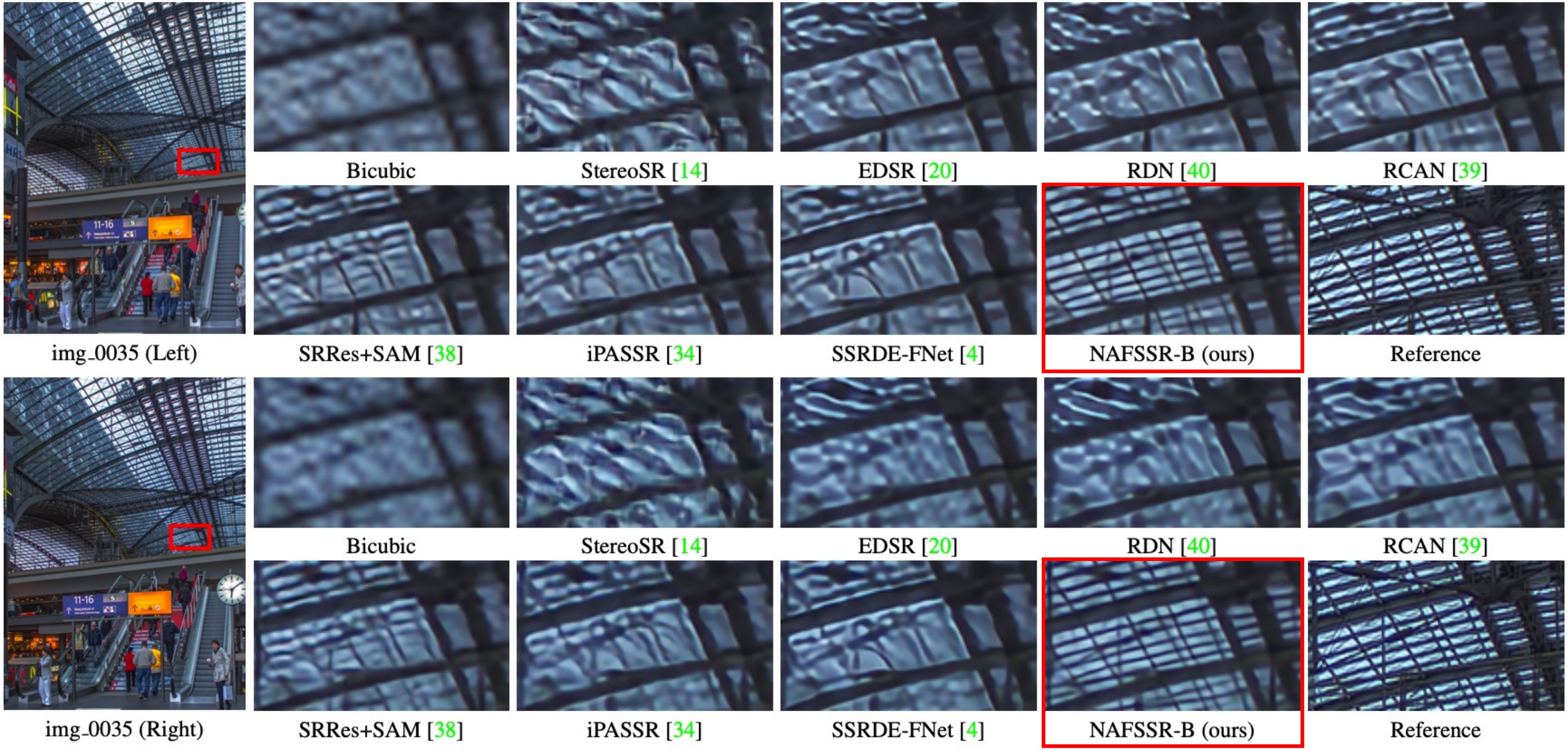
Results

- Runtime speedup



Models	PSNR	Time(ms)	Speedup
SSRDEFNet [4]	23.59	238.5	1.00×
NAFSSR-T (Ours)	23.64 (+0.05)	46.7	5.11×
NAFSSR-S (Ours)	23.88 (+0.29)	91.8	2.60×
NAFSSR-B (Ours)	24.07 (+0.48)	224.9	1.06×

Visual Examples



Bicubic StereoSR [14] EDSR [20] RDN [40] RCAN [39]

img_0035 (Left) SRRes+SAM [38] iPASSR [34] SSRDE-FNet [4] NAFSSR-B (ours) Reference

Bicubic StereoSR [14] EDSR [20] RDN [40] RCAN [39]

img_0035 (Right) SRRes+SAM [38] iPASSR [34] SSRDE-FNet [4] NAFSSR-B (ours) Reference

NTIRE Stereo Image Super-Resolution Challenge

- Additional Tricks for challenge
    - Further enlarge model by increasing its depth and width
    - Test-time data augmentations for self-ensemble [4]
    - Ensemble multiple models trained with various hyper-parameters [5]

- Result
    - 24.239 dB PSNR on the validation set
    - 23.787 dB PSNR on the test set (**First place**)

[4] *Lim, Bee, et al. "Enhanced deep residual networks for single image super-resolution." CVPRW, 2017.*
[5] *Wortsman, Mitchell, et al. "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time." arXiv preprint arXiv:2203.05482 (2022).*

Summary

- NAFSSR
  - Single View: NAFNet Block [1]
  - Cross-view: Stereo Cross Attention Module

- Tricks:
  - Training
    - Data augmentation: flip +RGB shuffle
    - Regularization: stochastic depth [2]
  - Inference
    - Test-time Local Statistics Converter [3]

Code (GitHub)

Scan Me

[1] *Chen, Liangyu, et al. "Simple baselines for image restoration." arXiv preprint arXiv:2204.04676 (2022).*

[2] *Huang, Gao, et al. "Deep networks with stochastic depth." ECCV, 2016.*

[3] *Chu, Xiaojie, et al. "Revisiting Global Statistics Aggregation for Improving Image Restoration." arXiv preprint arXiv:2112.04491 (2021).*

https://github.com/megvii-research/NAFNet