北京交通大学
BEIJING JIAOTONG UNIVERSITY

MePro 数字媒体信息处理研究中心
Center of Digital Media Information Processing

Meeting of Paper Sharing

Image and Video Super-Resolution

Qi Tang

2023/10/21

Resolution

Super-Resolution

✓ To improve the resolution of image/video.

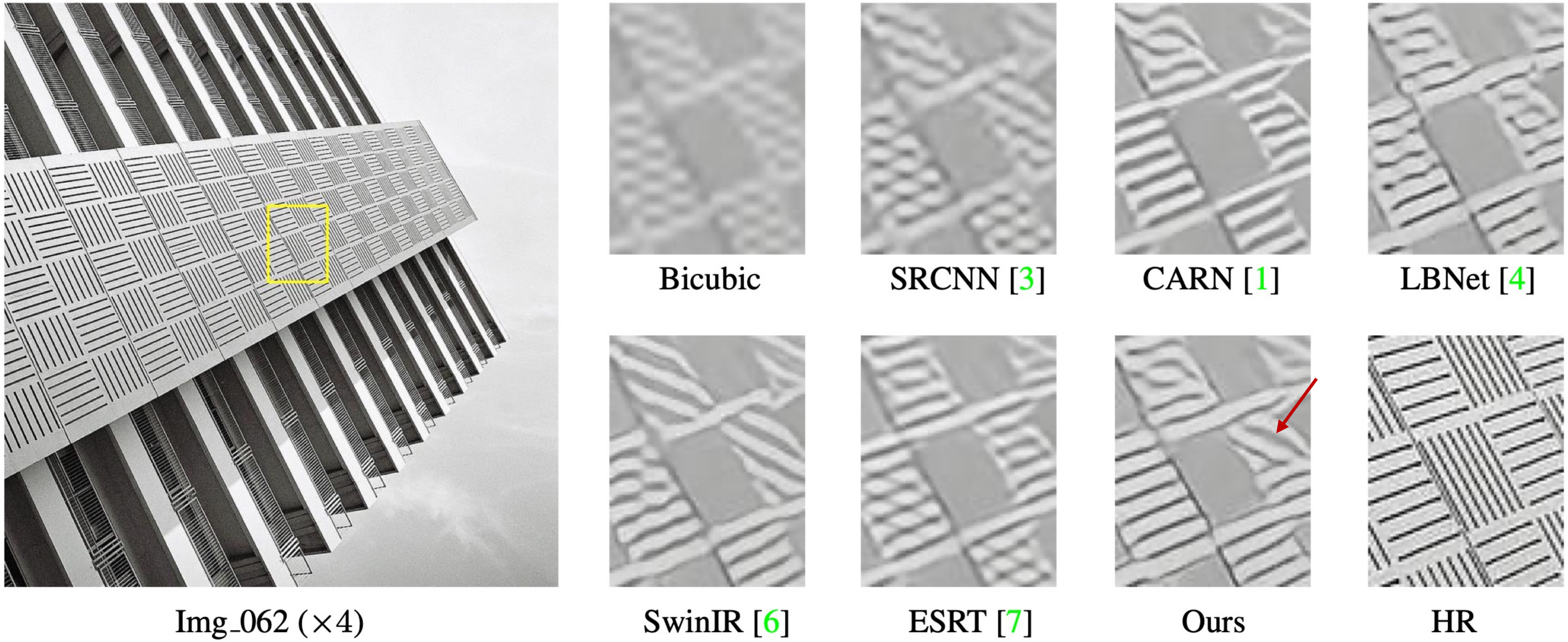
✓ To restore the high-resolution (HR) image consistent with the content of the low-resolution (LR) image. **Fidelity**

✓ To generate the realistic details of the image.

8×

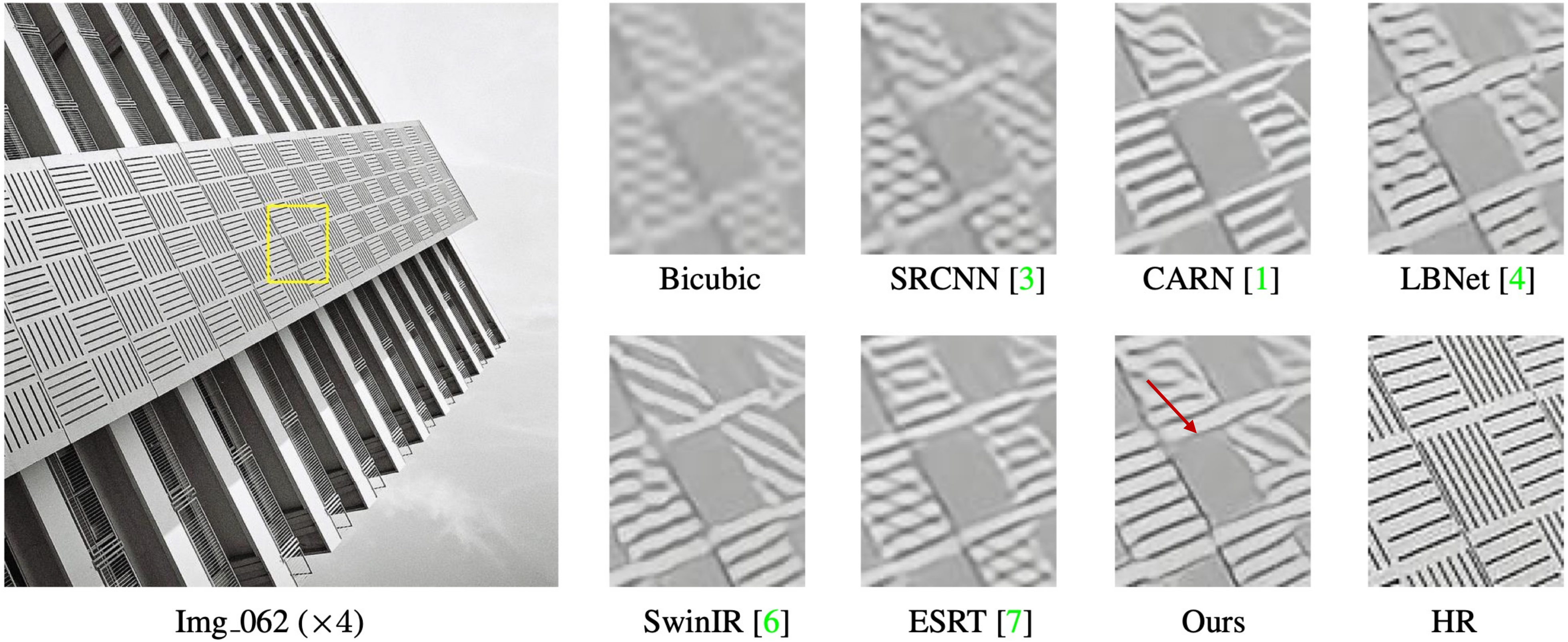
北京交通大學
BEIJING JIAOTONG UNIVERSITY

- ✓ To improve the resolution of image/video.
- ✓ To restore the high-resolution (HR) image consistent with the content of the low-resolution (LR) image. **Fidelity**
- ✓ To generate the realistic details of the image.



Super-Resolution

- ✓ To improve the resolution of image/video.
- ✓ To restore the high-resolution (HR) image consistent with the content of the low-resolution (LR) image. **Fidelity**
- ✓ To generate the realistic details of the image. **Photo-Realistic**



Super-Resolution

HD Reproduction of Classic Games

HD Reproduction of Animations

Restoration of old Photos

Super-Resolution

Save bandwidth for transmitting high-definition images

1,000×1,500, 100kb

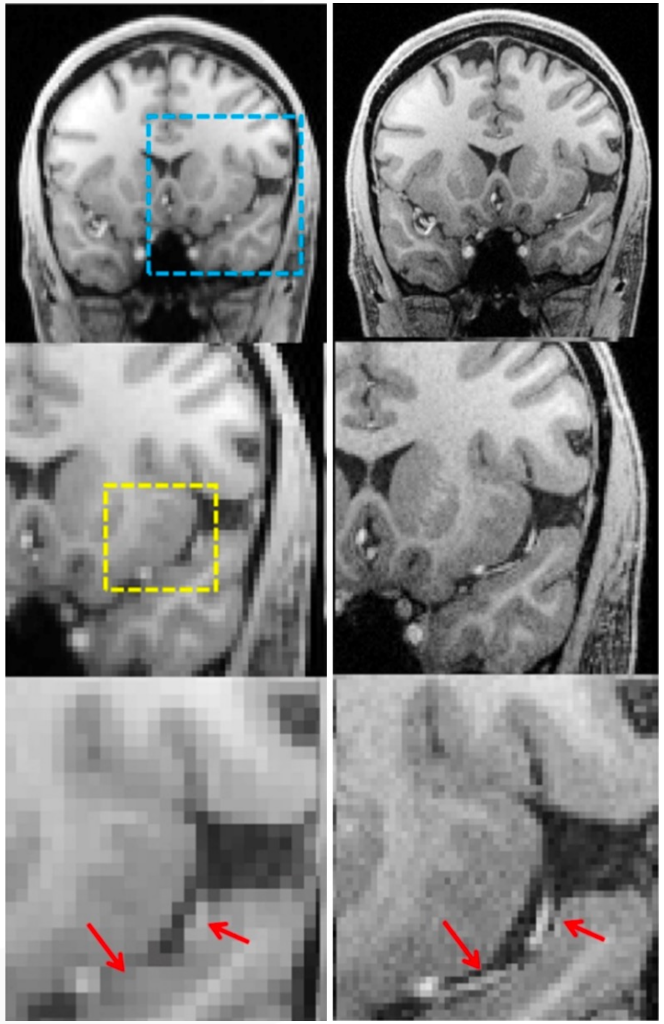
1/4 pixel of the original image is transmitted on the network

1,000×1,500, 25kb

high-definition image is restored on user equipment

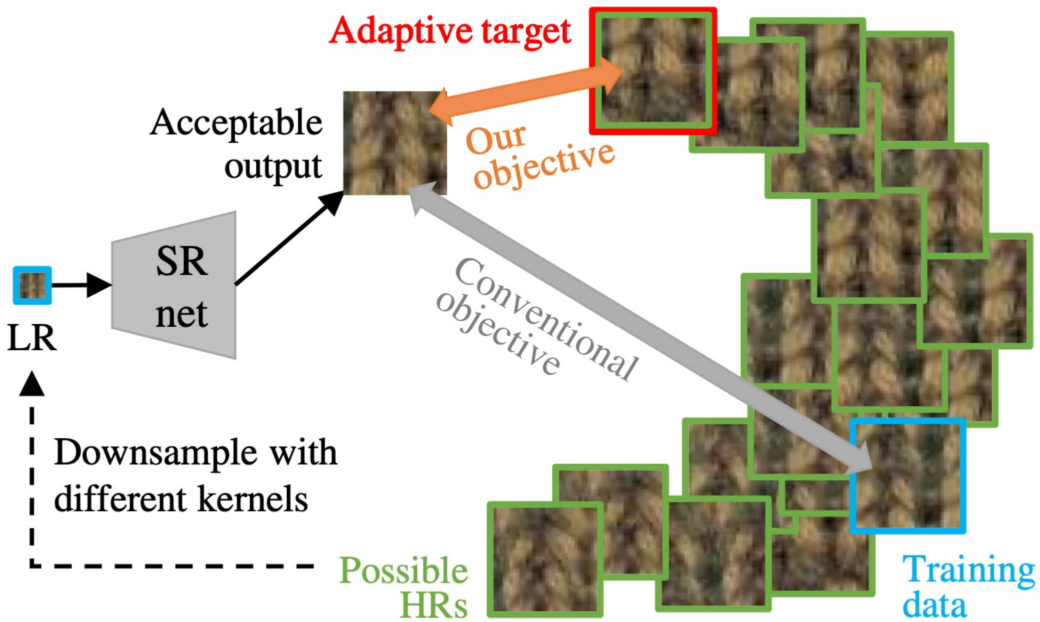
Super-Resolution

People's Wellbeing: MRI, Satellite Image, Surveillance…



Ill-Posed Problem

In most cases, there are several possible output images corresponding to a given input image and the problem can be seen as a task of selecting the most proper one from all the possible outputs. That is, the image restoration problem can be formulated as the problem of estimating the distribution conditioned on the input image.



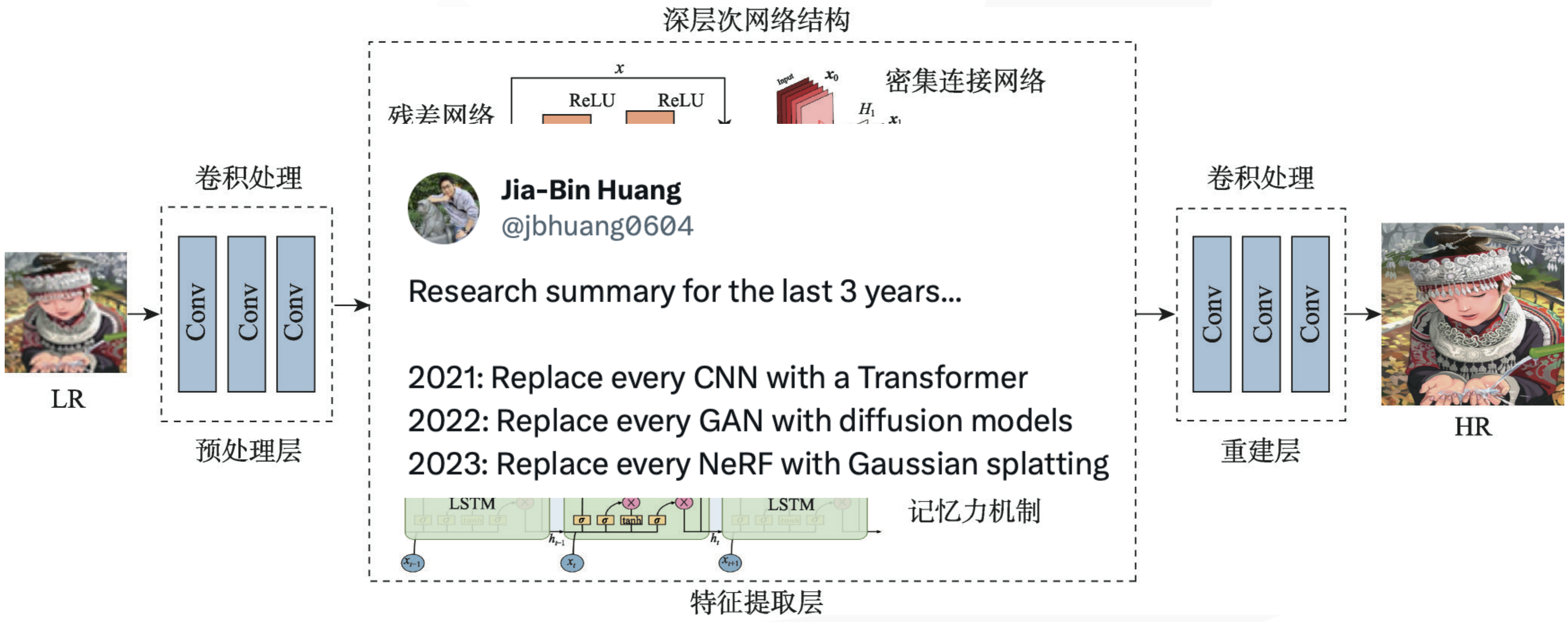
Younghyun Jo, Seoung Wug Oh, Peter Vajda, and Seon Joo Kim. Tackling the Ill-Posedness of Super-resolution Through Adaptive Target Generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 16236–16245, 2021

Jaeyoung Yoo, Sang-ho Lee, and Nojun Kwak. Image Restoration by Estimating Frequency Distribution of Local Patches. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 6684–6692, 2018.

Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering Realistic Texture in Image Super-Resolution by Deep Spatial Feature Transform. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 606–615, 2018.

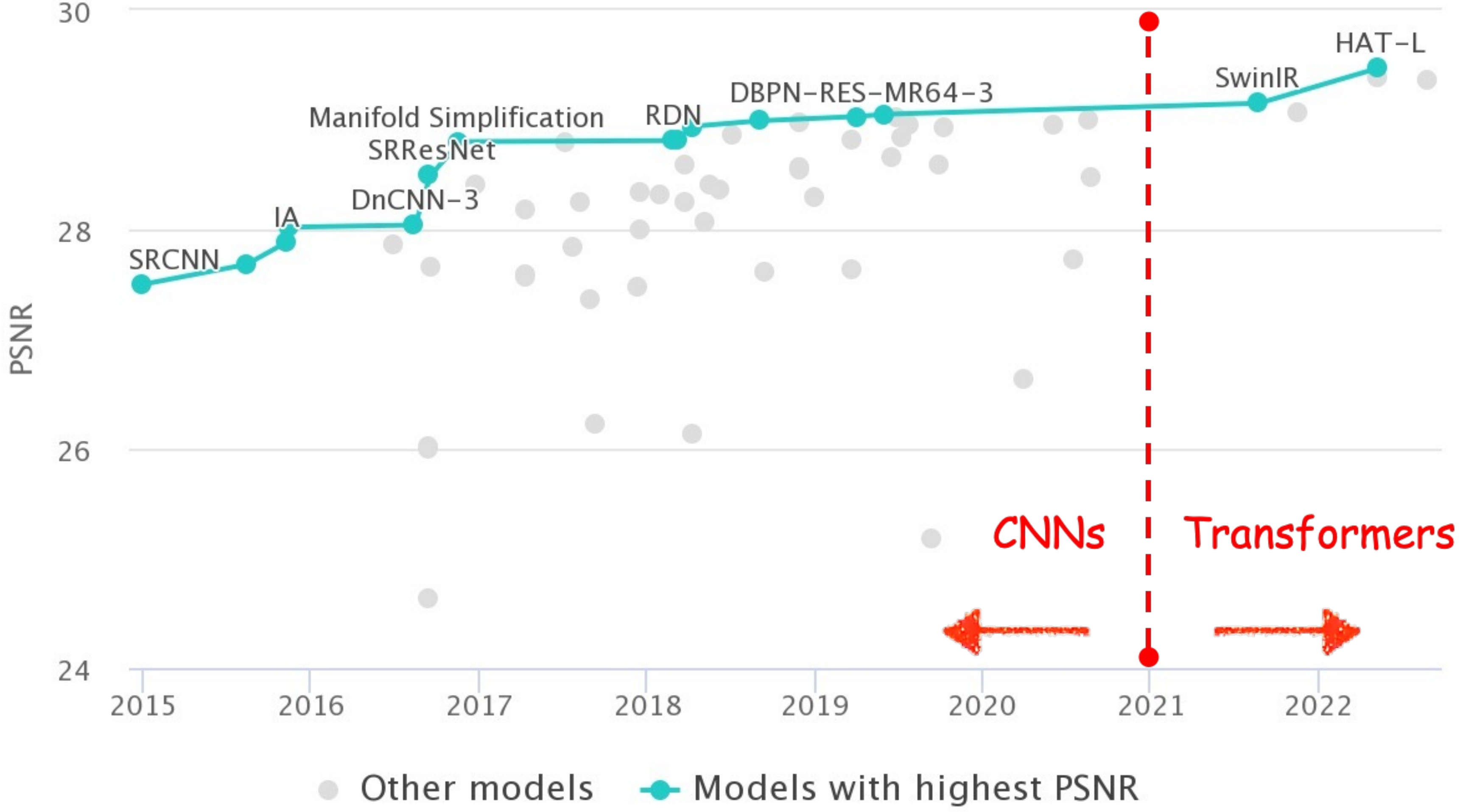
CNN-based Backbone

Shallow Feature Extractor	Deep Feature Extractor	Reconstruction



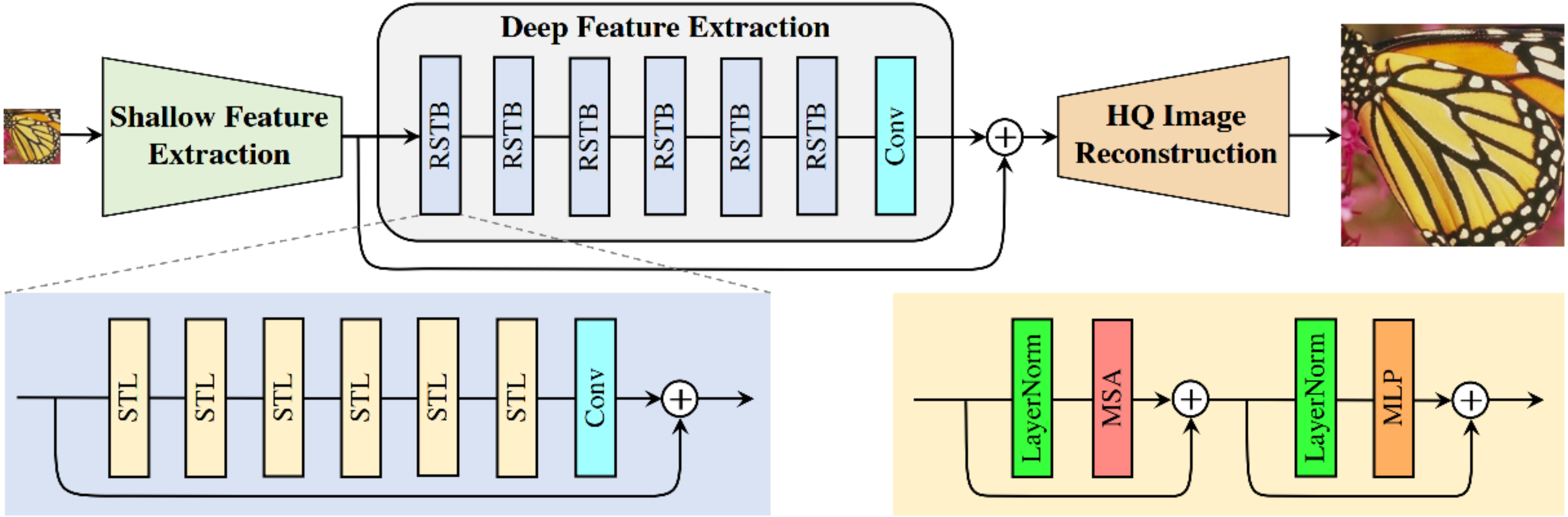
Transformer-based Backbone

Transformers refresh the state-of-the-art in Network designs.



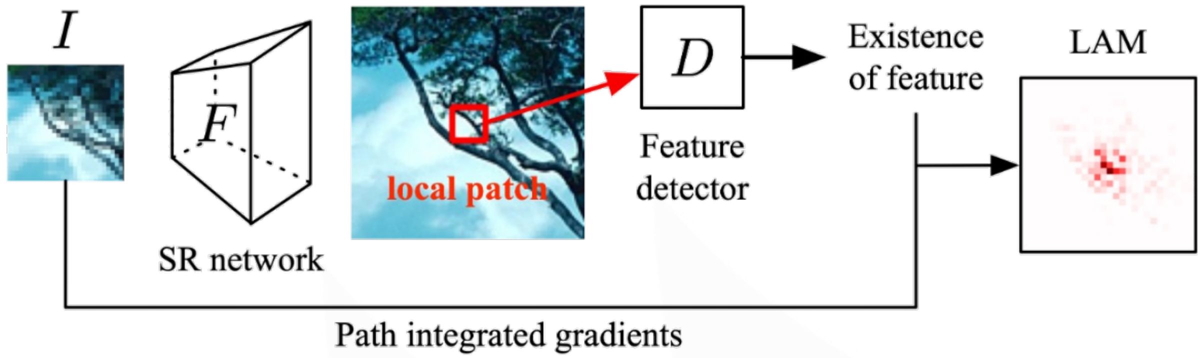
Transformer-based Backbone

Shallow Feature Extractor | Deep Feature Extractor | Reconstruction

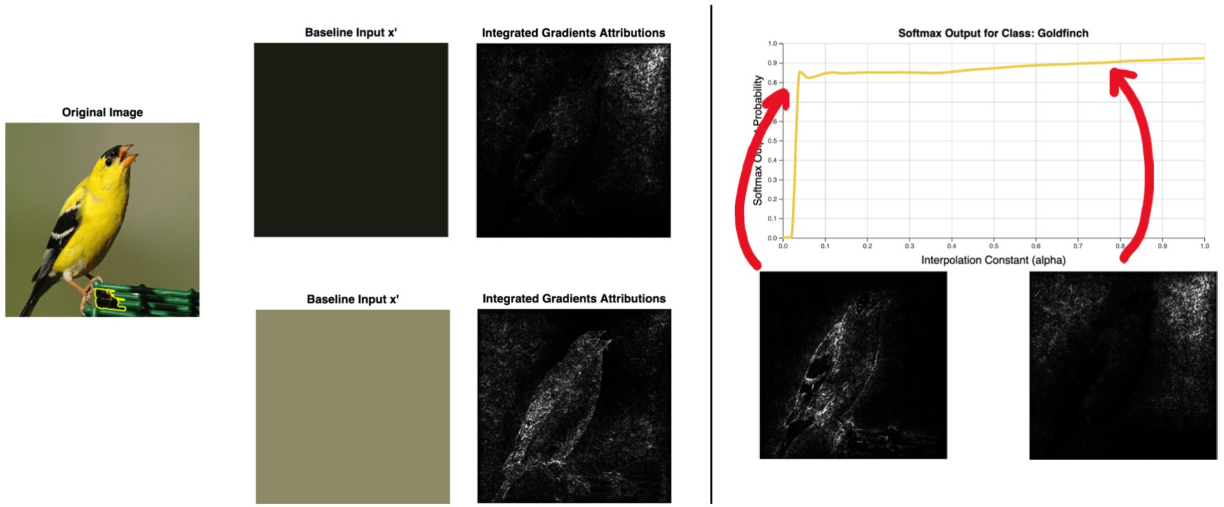


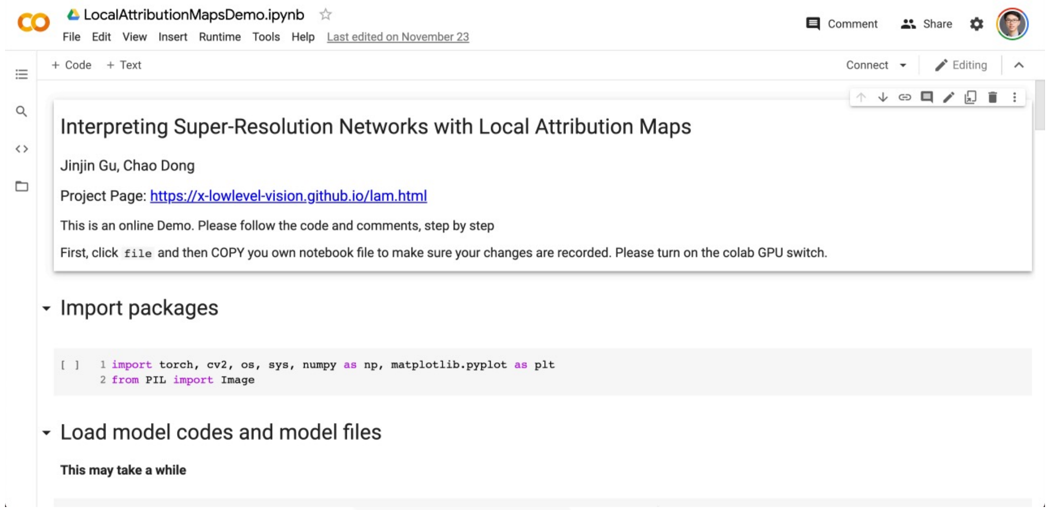
Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image Restoration Using Swin Transformer. In Proceedings of the IEEE International Conference on Computer Vision, pages 1833–1844, 2021

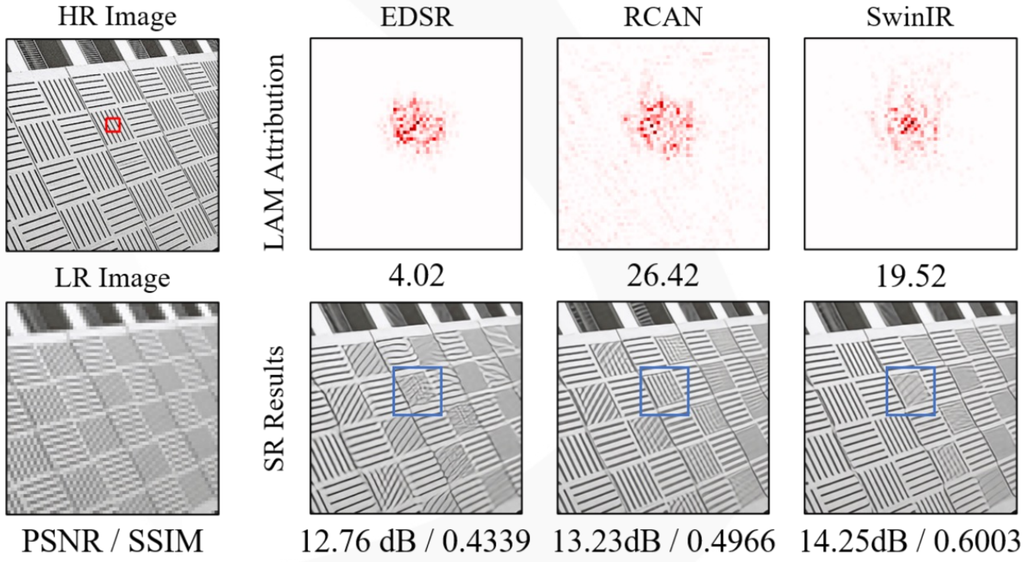
Local Attribution Map



$$\mathrm{LAM}_{F,D}(\gamma)_i := \int_0^1 \frac{\partial D(F(\gamma(\alpha))}{\partial \gamma(\alpha)_i} \times \frac{\partial \gamma(\alpha)_i}{\partial \alpha} d\alpha$$



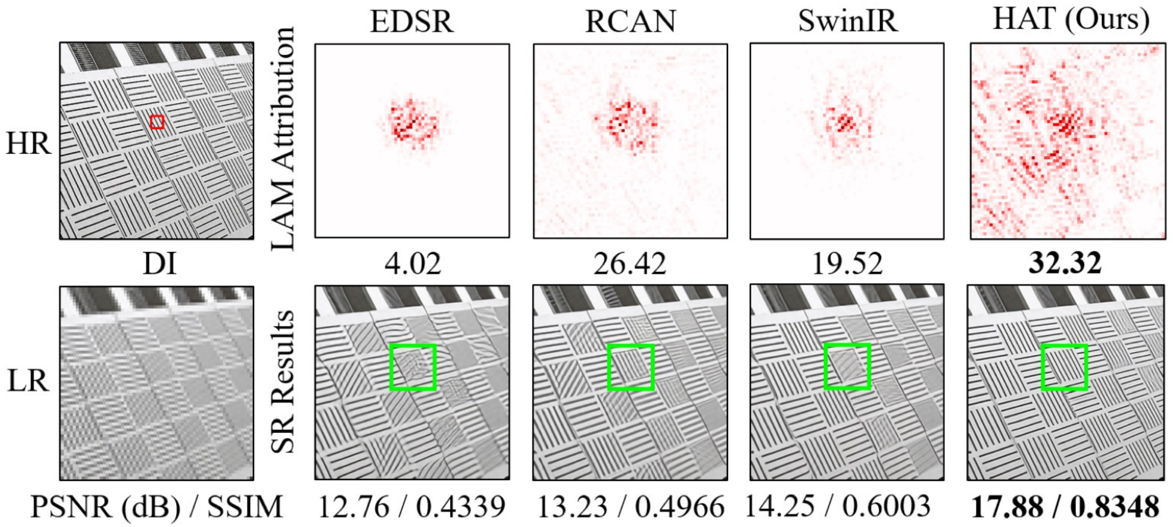




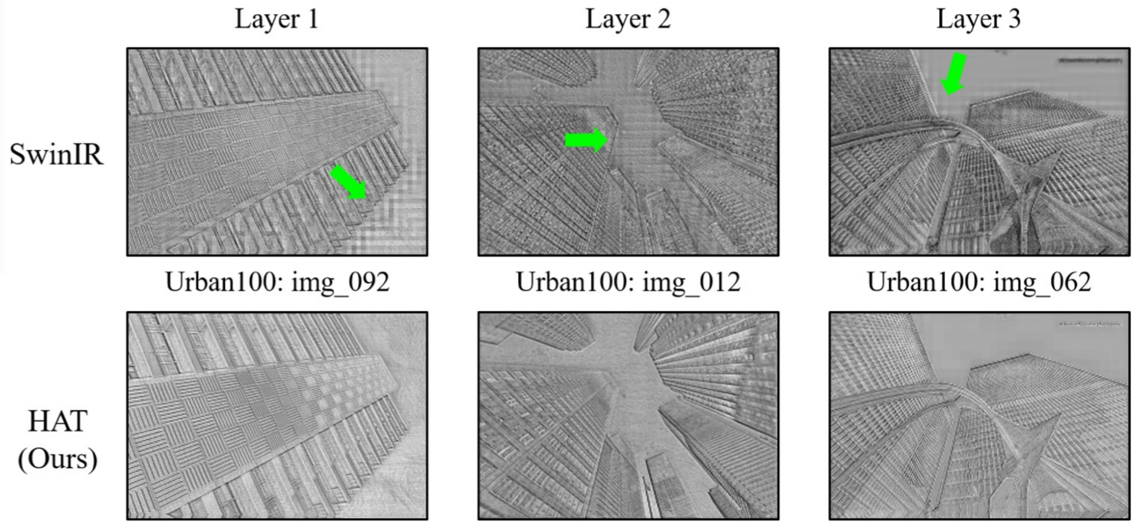
Jinjin Gu and Chao Dong. Interpreting Super-Resolution Networks with Local Attribution Maps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9199–9208, 2021.

Spatial and Channel Attention

- SwinIR utilizes less information compared to RCAN
- SwinIR has a much stronger mapping ability than CNN, and thus could use less information to achieve better performance



- Obvious blocking artifacts in the intermediate features of SwinIR, which are caused by the window partition mechanism. It suggests that the shifted window mechanism is inefficient to build the cross-window connection.



Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating More Pixels in Image Super-Resolution Transformer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages22367–22377, 2023.

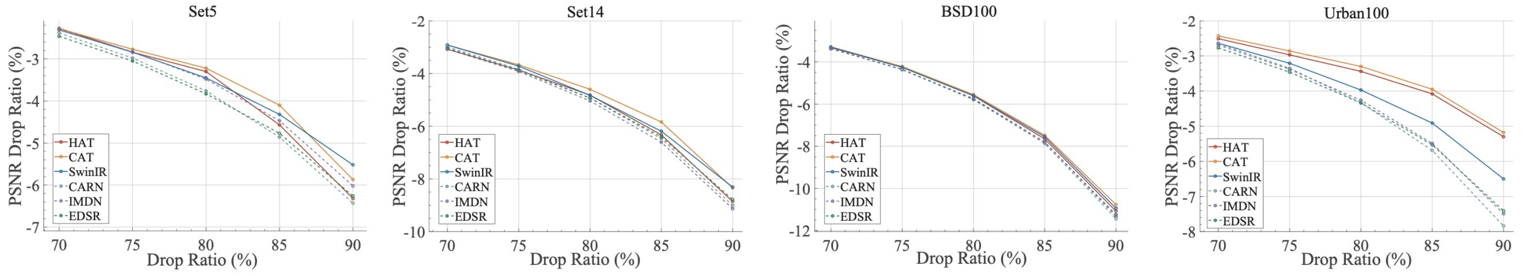
Feature Modulation Transformer

Transformer exhibits reduced sensitivity to high-frequency information and excels in capturing low-frequency information

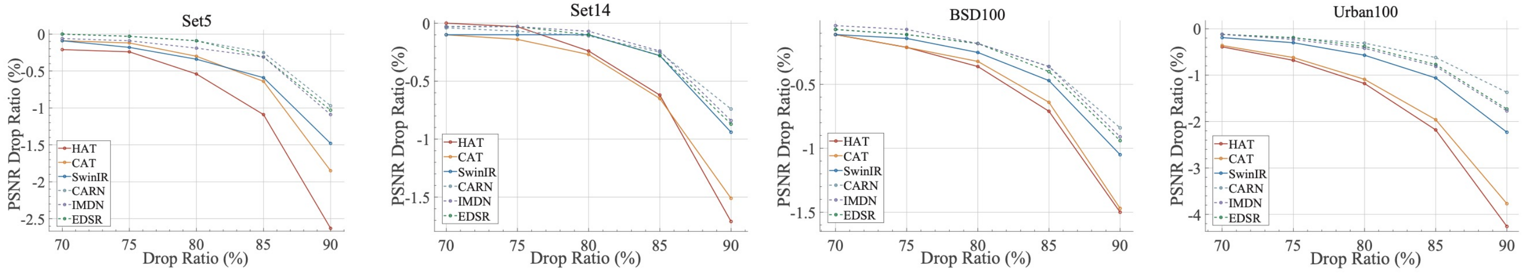
$$R_{drop}^{D}(\gamma) = \frac{P(0) - P^{D}(\gamma)}{P(0)}$$

$$R_{drop}^{E}(\gamma) = \frac{P^{E}(\gamma) - P(0)}{P(0)}$$

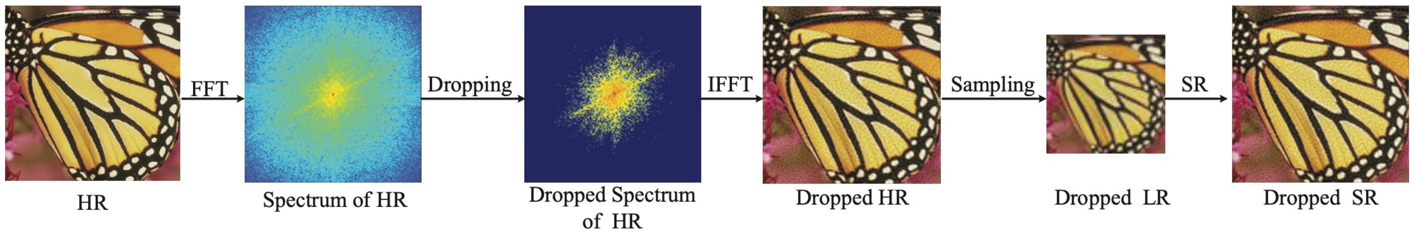
$$X_{\text{drop}}^{HR}(\gamma) = \text{IFFT}\left(\text{Drop}\left(\left|FFT\left(X^{HR}\right)\right|, \gamma\right)\right)$$



(a) Dependency of different structures on high-frequency information.



(b) Effectiveness of reconstructing high-frequency information.

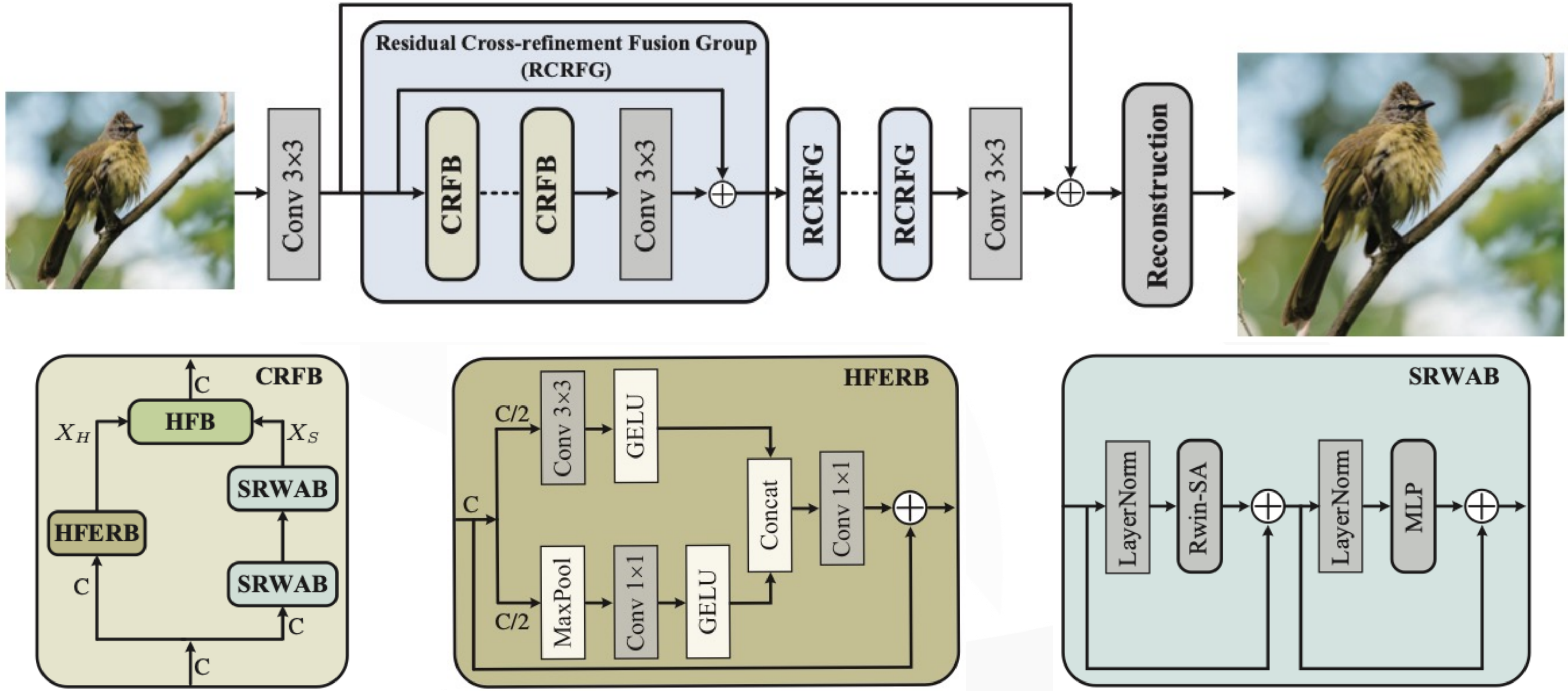


(c) The procedure of dropping high-frequency.

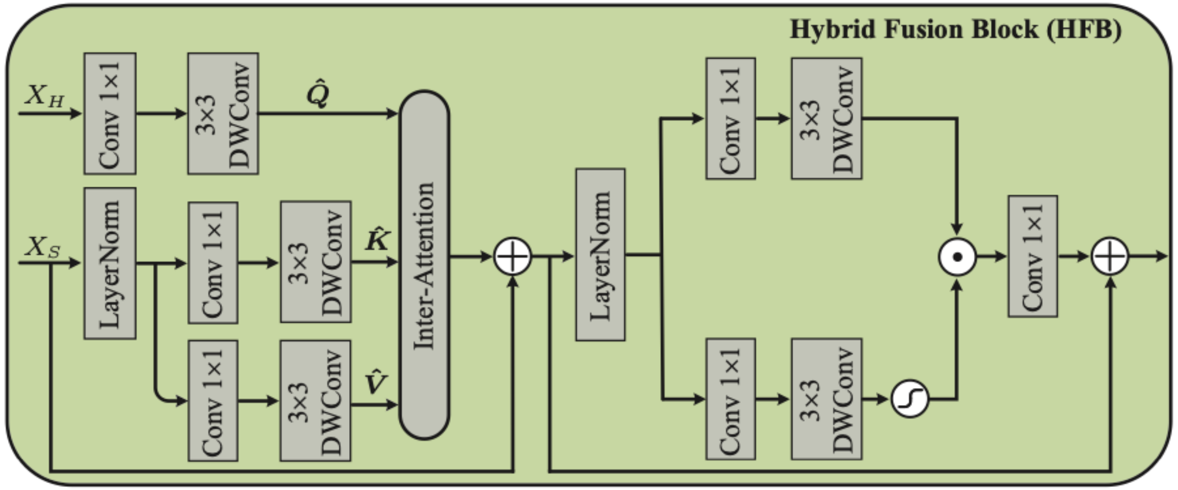
Ao Li, Le Zhang, Yun Liu, and Ce Zhu. Feature Modulation Transformer: Cross-Refinement of Global Representation via High-Frequency Prior for Image Super-Resolution. In Proceedings of the IEEE International Conference on Computer Vision, pages 12514–12524, 2023.

Feature Modulation Transformer

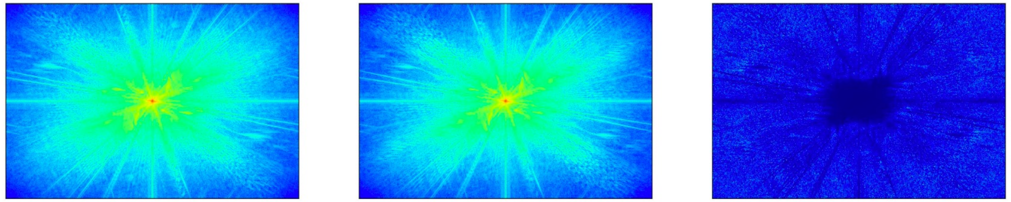
Shallow Feature Extractor | Deep Feature Extractor | Reconstruction



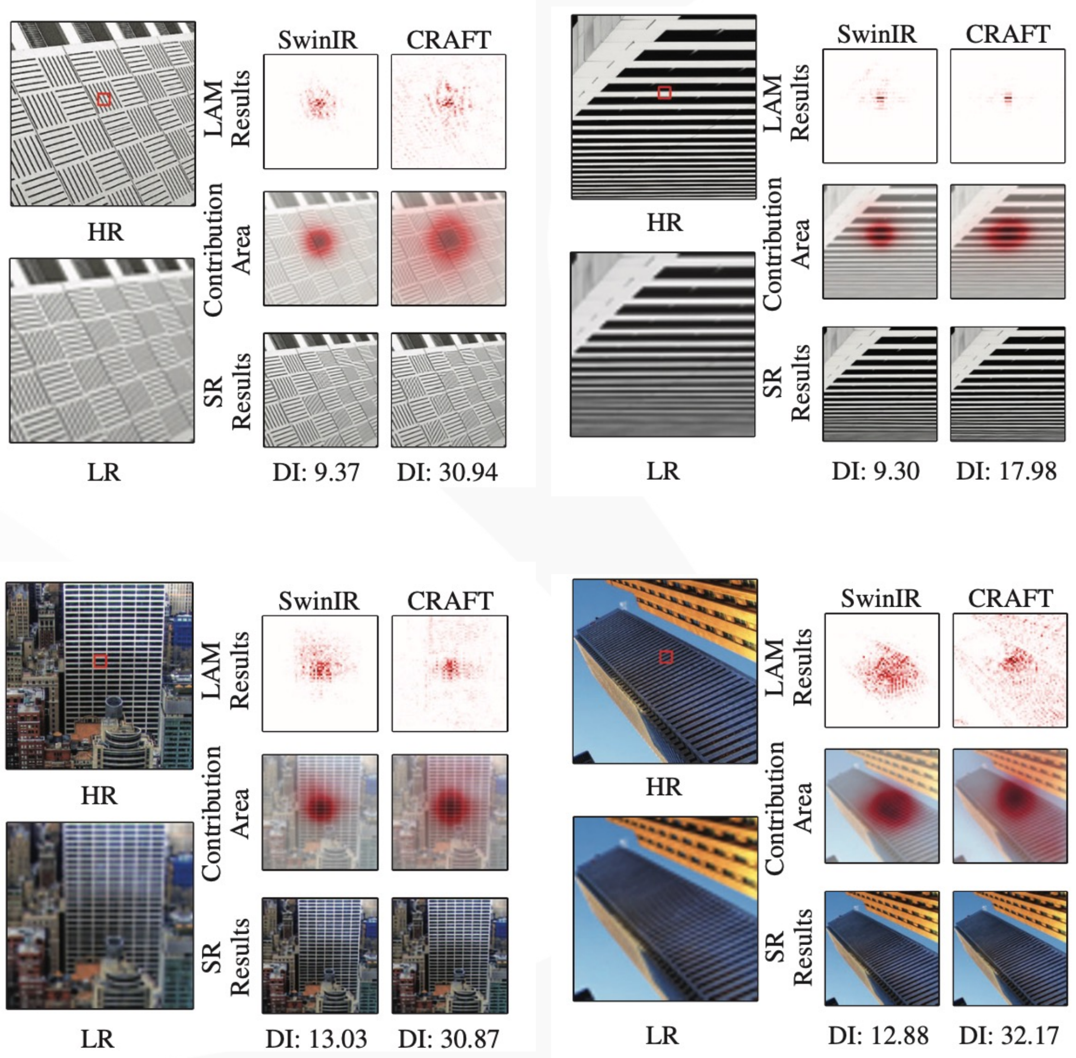
Feature Modulation Transformer



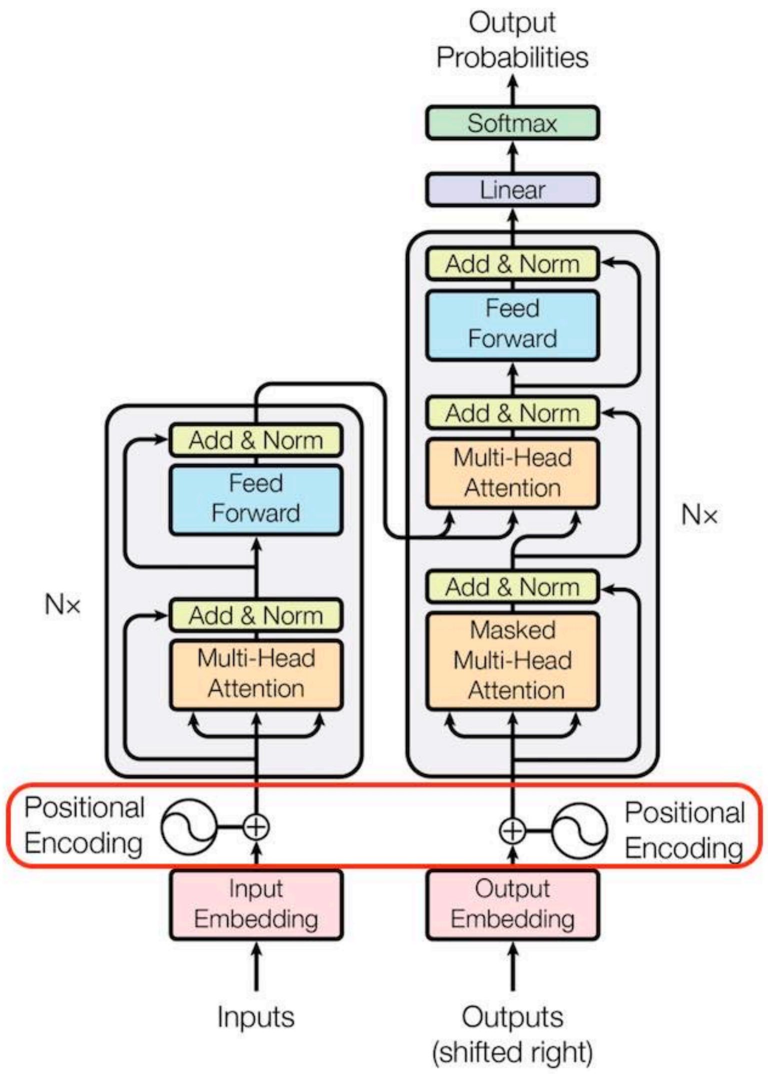
Bicubic | *w/o* H | *w/* H | HR



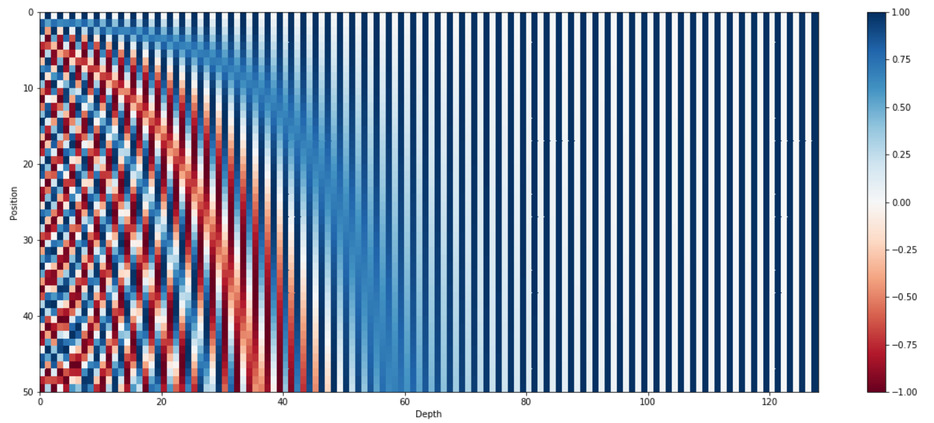
$\Phi(w/\ H)$ | $\Phi(w/o\ H)$ | $|\Phi(w/\ H) - \Phi(w/o\ H)|$



Position Embedding in Low-Level



Recent have shown that positional encoding can enhance the network in the high-frequency domain by expanding 2D coordinates into a high-dimensional periodic positional encoding.

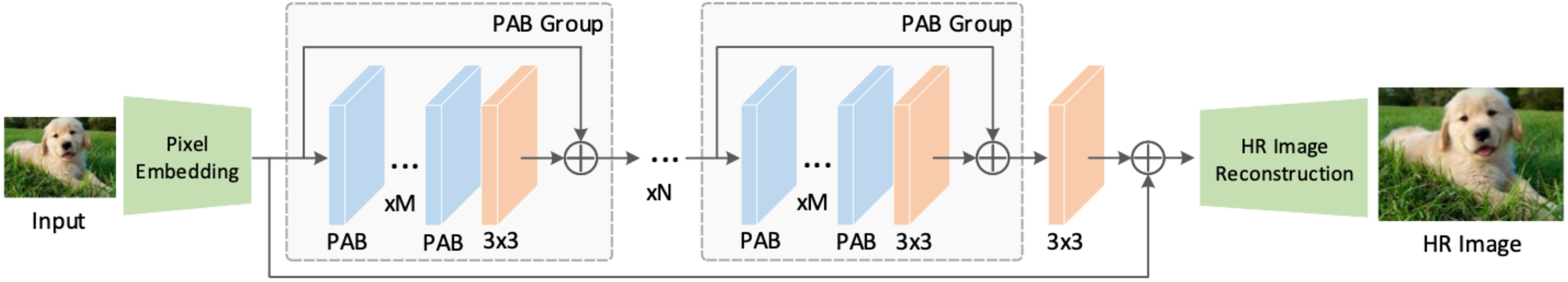


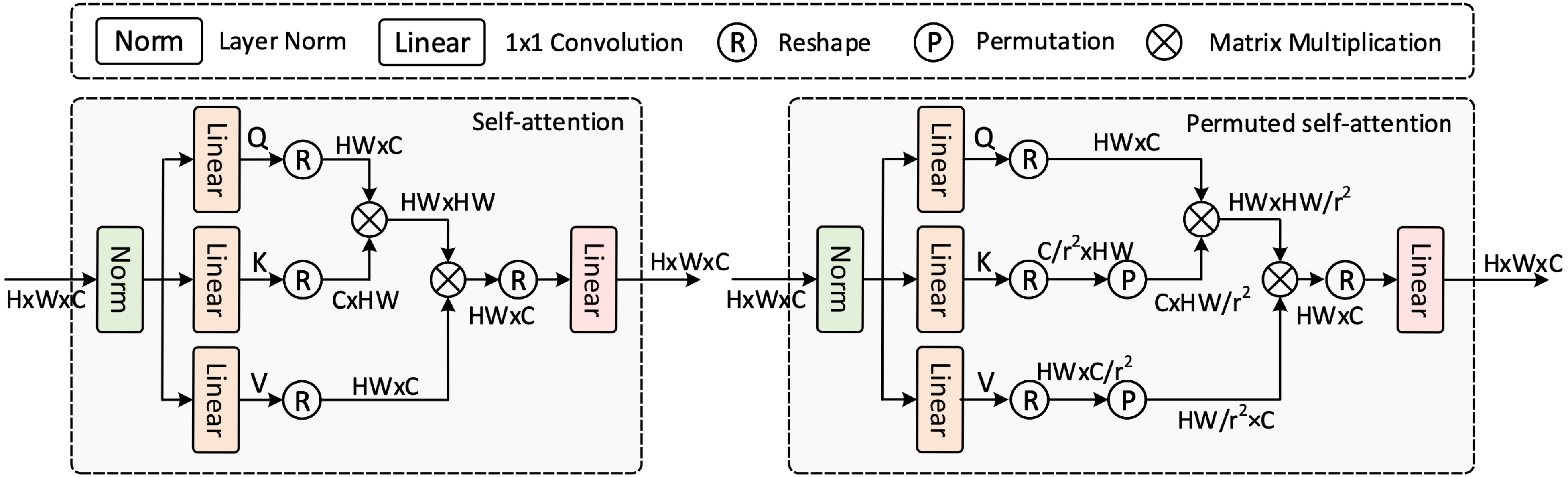
Jinsu Yoo, Taehoon Kim, Sihaeng Lee, Seung Hwan Kim, Honglak Lee, and Tae Hyun Kim. Enriched CNN-Transformer Feature Aggregation Networks for Super-Resolution. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, pages 4956 4965, 2023.

Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Ultrasr: Spatial Encoding is a Missing Key for Implicit Image Function-Based Arbitrary-Scale Super-Resolution. arXiv preprint arXiv:2103.12716, 2021.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing Scenes as Neural Radiance Fields for View Synthesis. Communications of the ACM, 65(1):99–106, 2021.

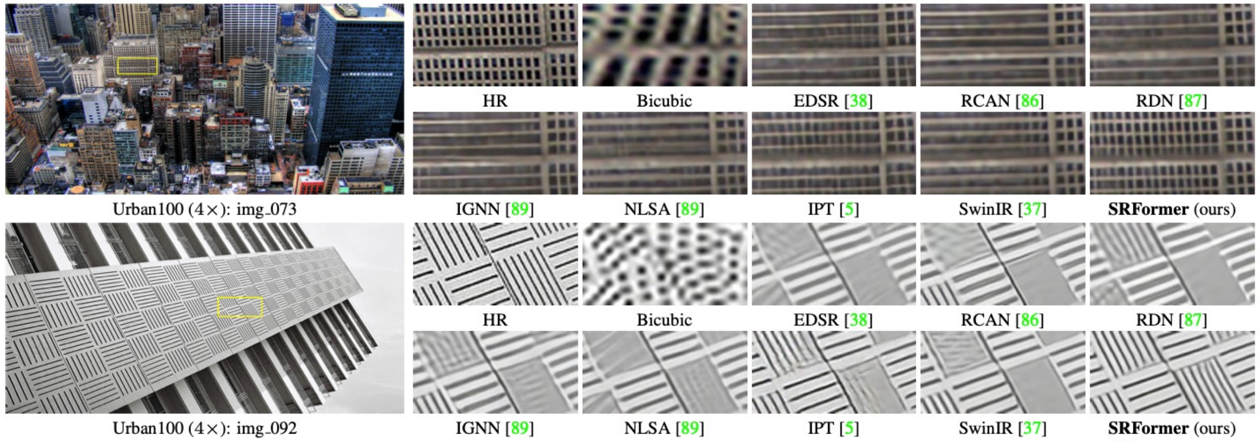
SRFormer

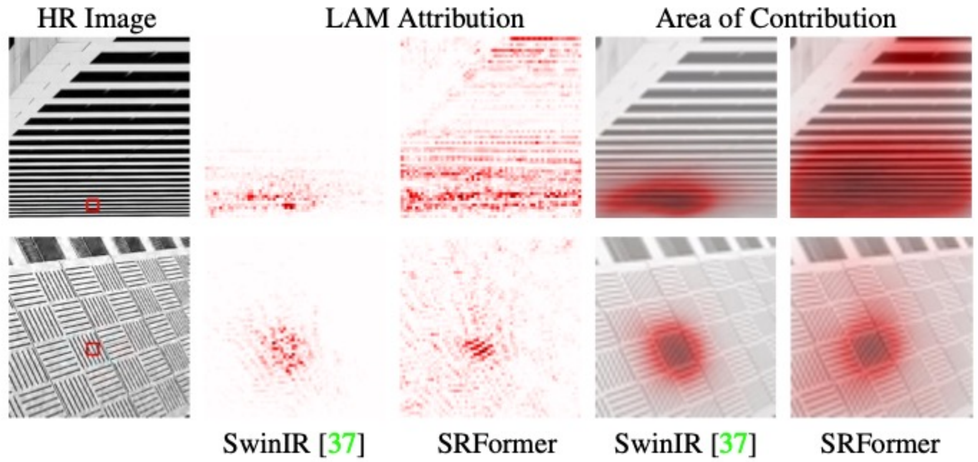




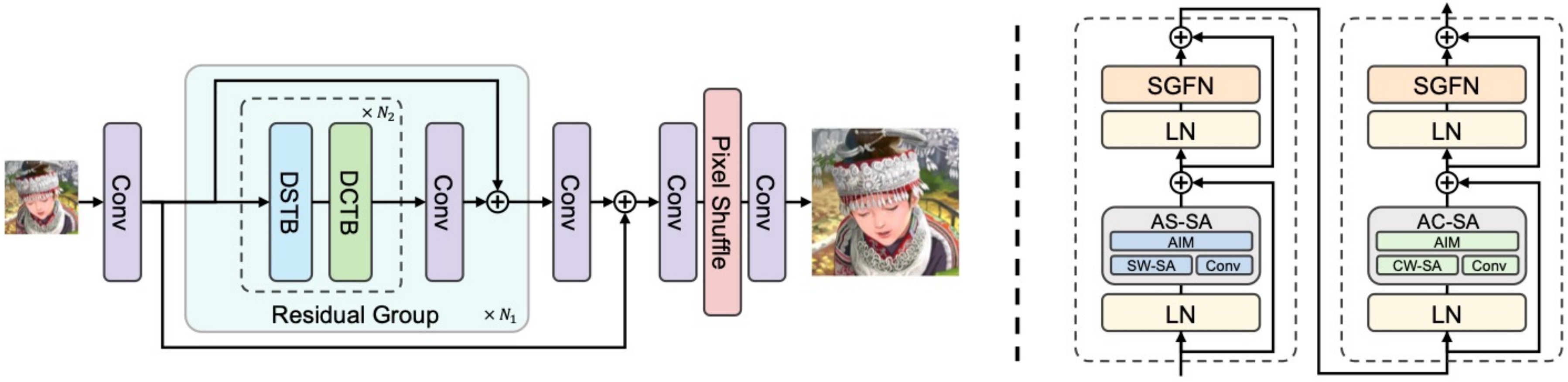
Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. SRFormer: Permuted Self-Attention for Single Image Super-Resolution. In Proceedings of the IEEE International Conference on Computer Vision, pages 12780–12791, 2023.

Method	Window size	Params	MACs	SET5 [3]		SET14 [77]		B100 [45]		Urban100 [20]		Manga109 [46]	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SwinIR [37]	8 × 8	11.75M	2868G	38.24	0.9615	33.94	0.9212	32.39	0.9023	33.09	0.9373	39.34	0.9784
	12 × 12	11.82M	3107G	38.30	0.9617	34.04	0.9220	32.42	0.9026	33.28	0.9381	39.44	0.9788
	16 × 16	11.91M	3441G	38.32	0.9618	34.00	0.9212	32.44	0.9030	33.40	0.9394	39.53	0.9791
SRFormer w/o ConvFFN	12 × 12	9.97M	2381G	38.23	0.9615	34.00	0.9216	32.37	0.9023	32.99	0.9367	39.30	0.9786
	16 × 16	9.99M	2465G	38.25	0.9616	33.98	0.9209	32.38	0.9022	33.09	0.9371	39.42	0.9789
	24 × 24	10.06M	2703G	38.30	0.9618	34.08	0.9225	32.43	0.9030	33.38	0.9397	39.44	0.9786
SRFormer	12 × 12	10.31M	2419G	38.22	0.9614	34.08	0.9220	32.38	0.9025	33.08	0.9372	39.13	0.9780
	16 × 16	10.33M	2502G	38.31	0.9617	34.10	0.9217	32.43	0.9026	33.26	0.9385	39.36	0.9785
	24 × 24	10.40M	2741G	38.33	0.9618	34.13	0.9228	32.44	0.9030	33.51	0.9405	39.49	0.9788

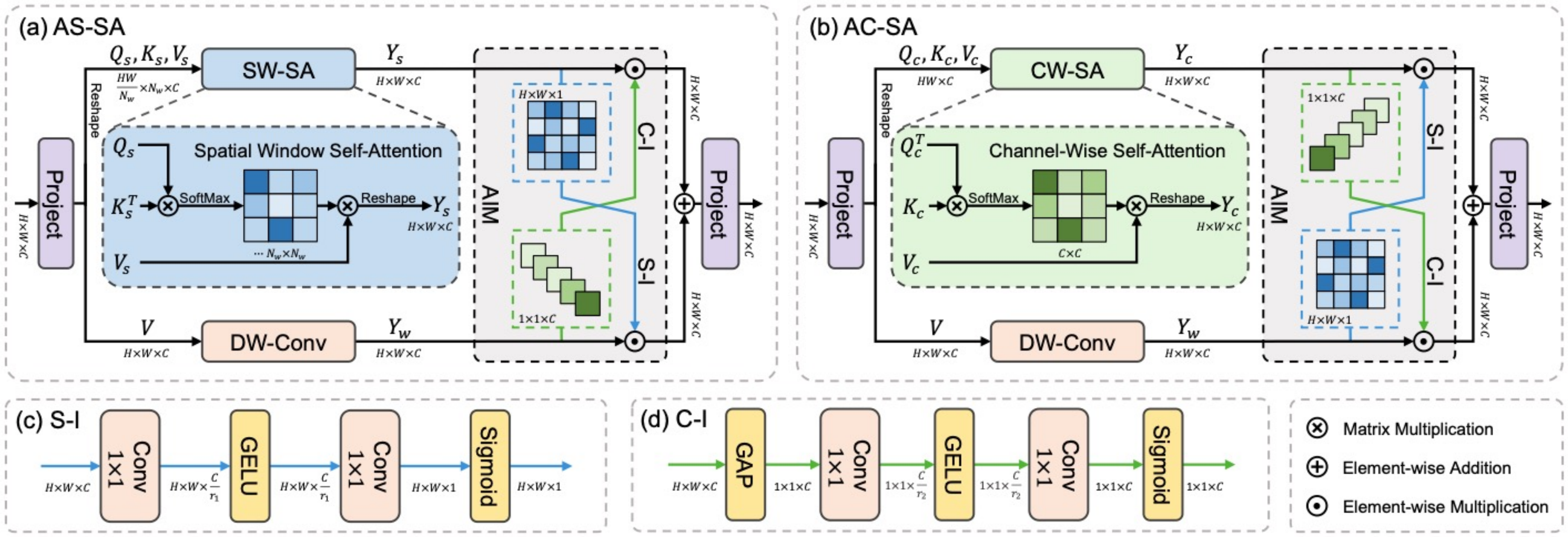




Spatial and Channel Attention



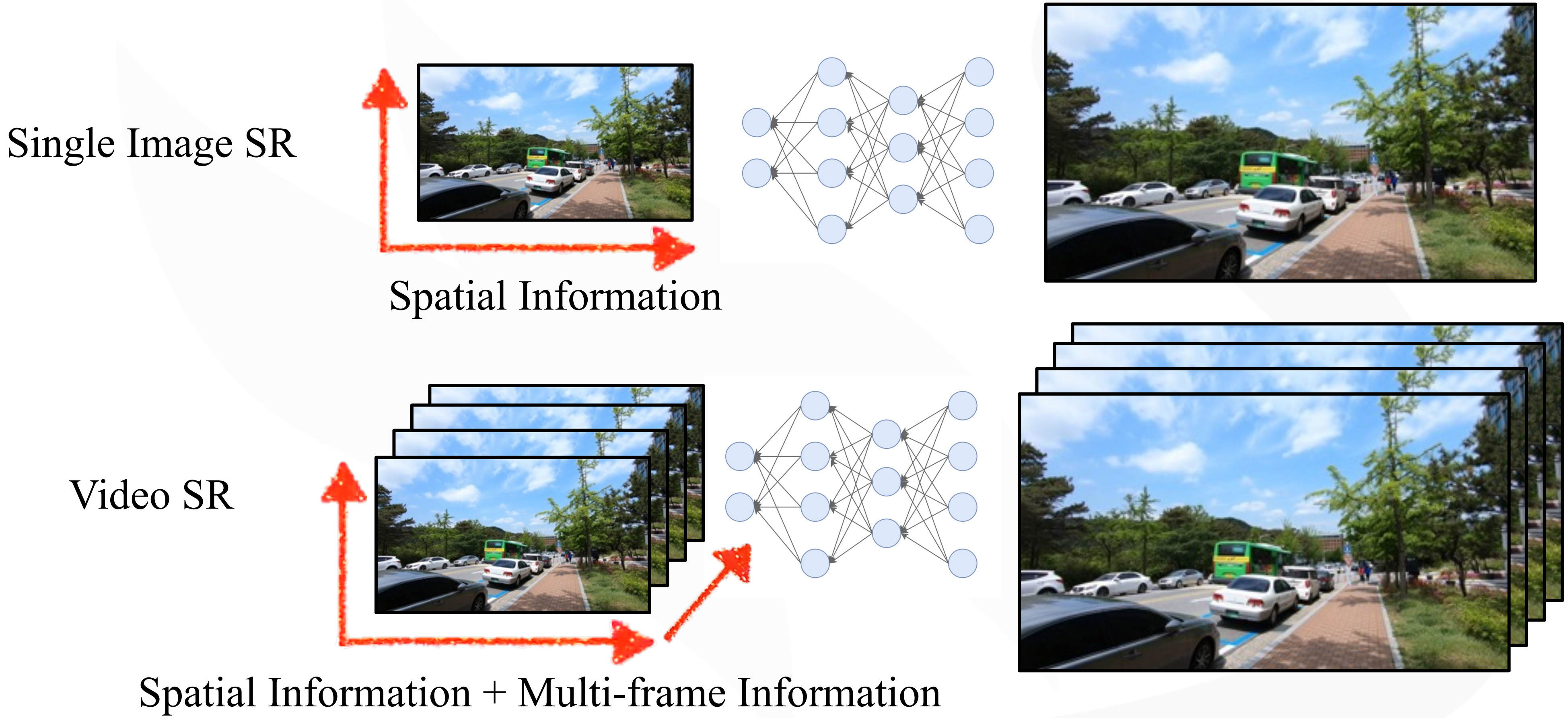
(a) DAT (b) DSTB (c) DCTB



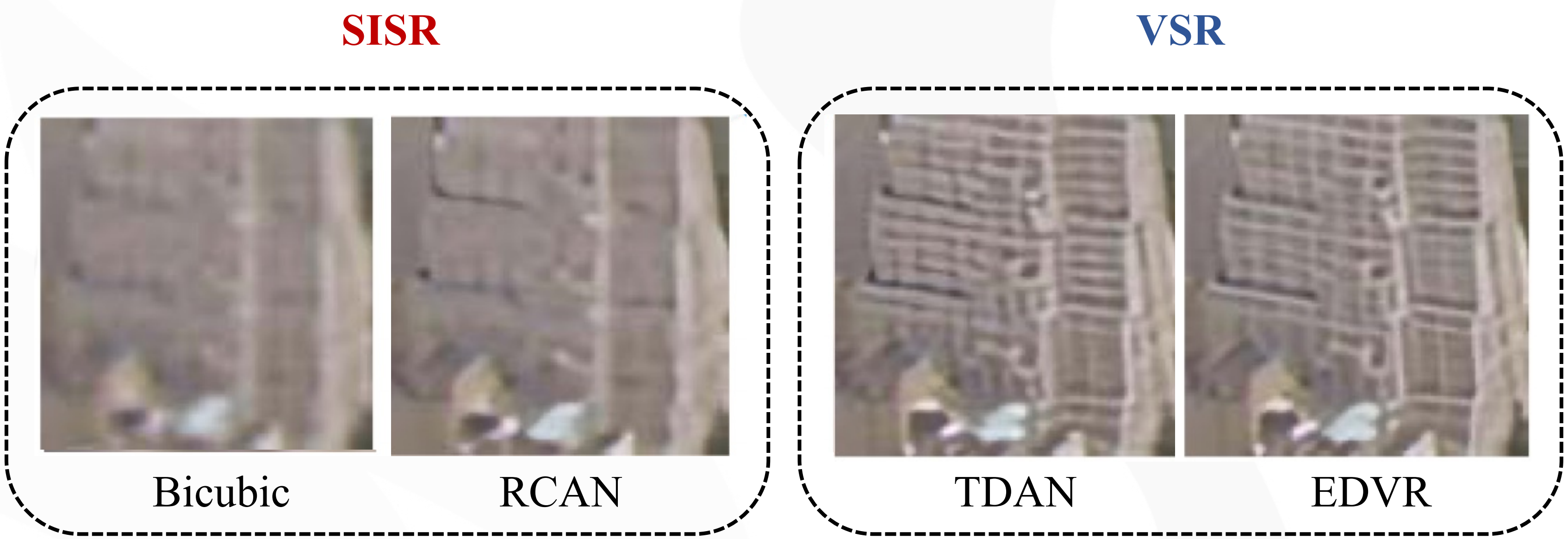
Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual Aggregation Transformer for Image Super-Resolution. In Proceedings of the IEEE International Conference on Computer Vision, pages 12312–12321, 2023.

Video Super-Resolution

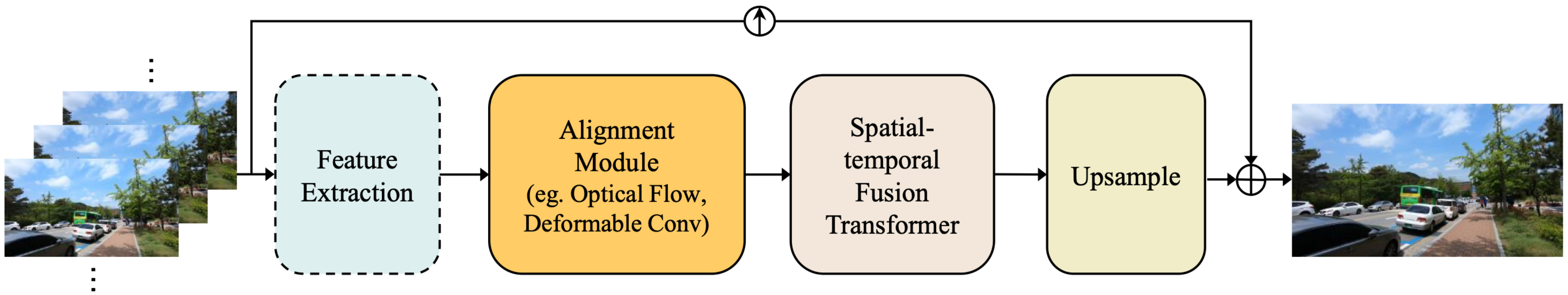
Video SR exploit the complementary sub-pixel information from multiple frames.

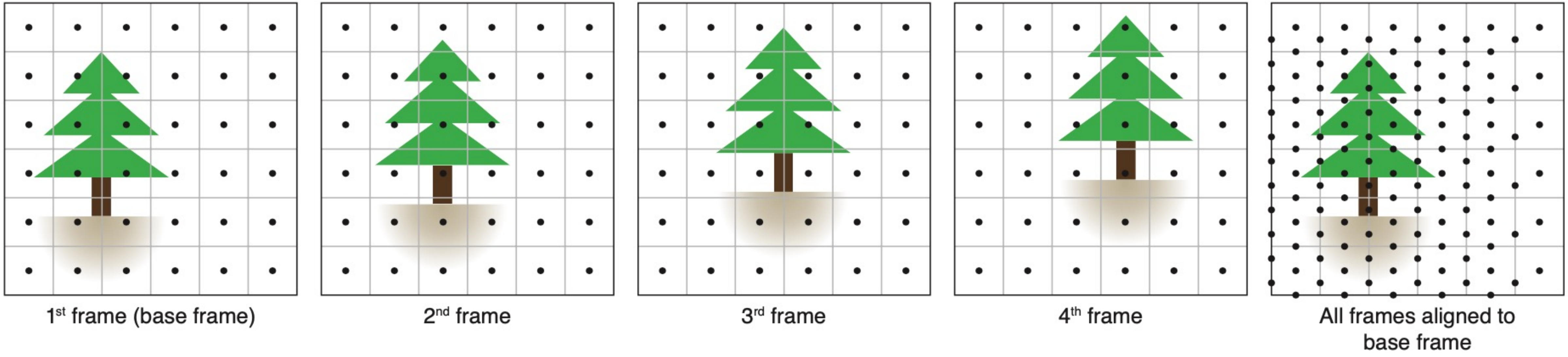


Video Super-Resolution



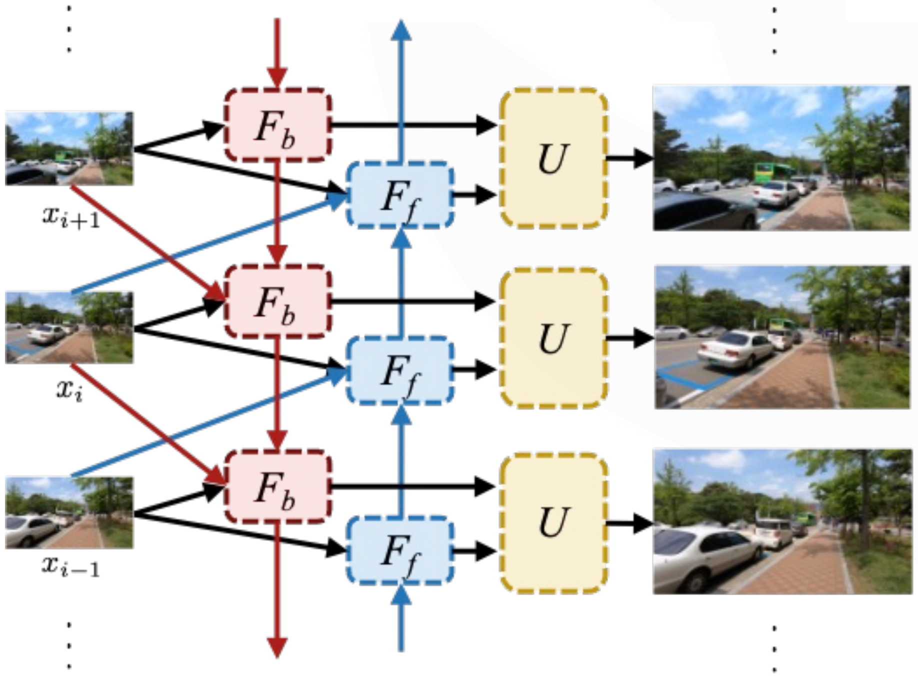
Inter-Frame Alignment

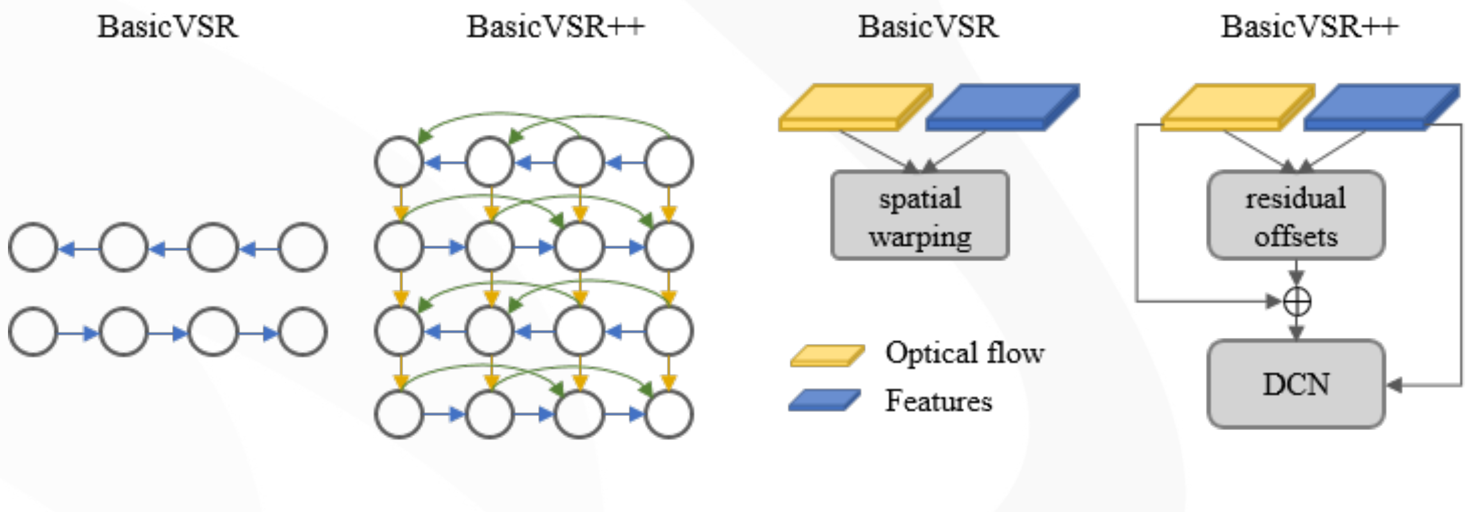




Video Super-Resolution Backbone

	Sliding-Window			Recurrent				
	EDVR	MuCAN	TDAN	BRCN	FRVSR	RSDN	**BasicVSR**	**IconVSR**
Propagation	Local	Local	Local	**Bidirectional**	Unidirectional	Unidirectional	**Bidirectional**	**Bidirectional** (coupled)
Alignment	**Yes** (DCN)	**Yes** (correlation)	**Yes** (DCN)	No	**Yes** (flow)	No	**Yes** (flow)	**Yes** (flow)
Aggregation	Concatenate + **TSA**	Concatenate	Concatenate	Concatenate	Concatenate	Concatenate	Concatenate	Concatenate + **Refill**
Upsampling	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle



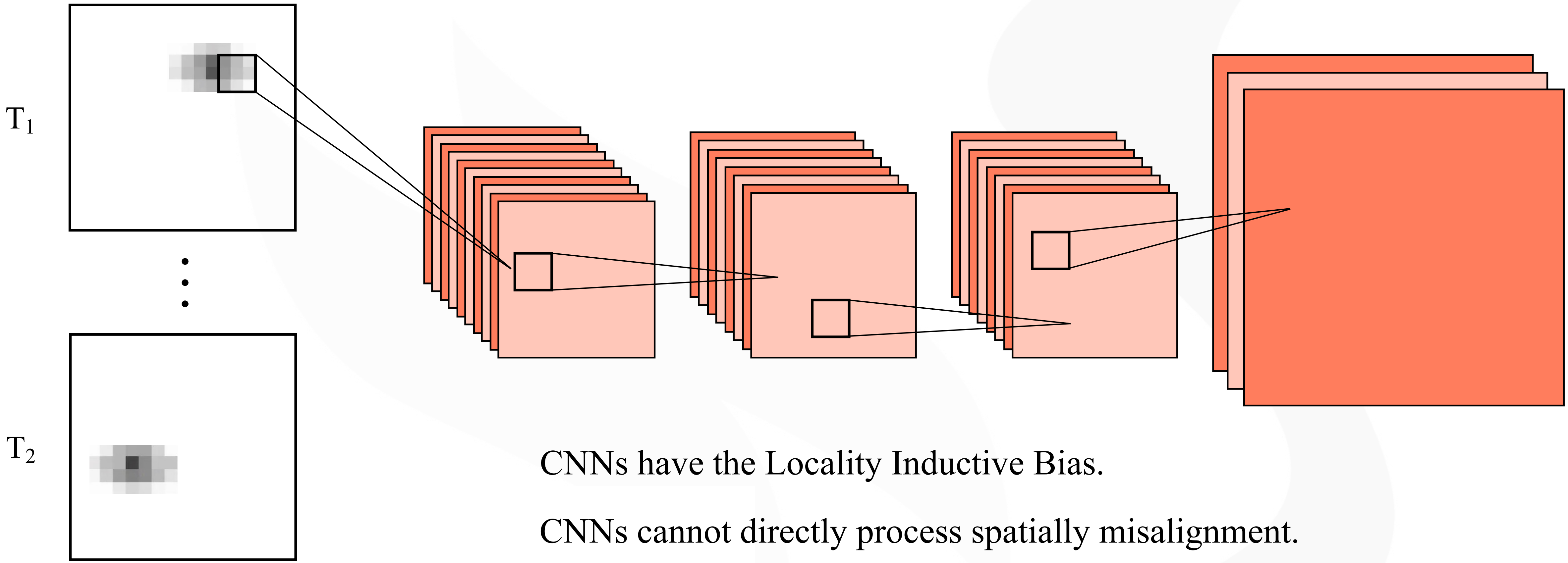


(a) Propagation (b) Alignment

Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The Search for Essential Components in Video Super-Resolution and Beyond. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4947–4956, 2021.

Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving Video Super-Resolution with Enhanced Propagation and Alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5972–5981, 2022

Video Super-Resolution Backbone



CNNs have the Locality Inductive Bias.

CNNs cannot directly process spatially misalignment.

Video Super-Resolution Backbone

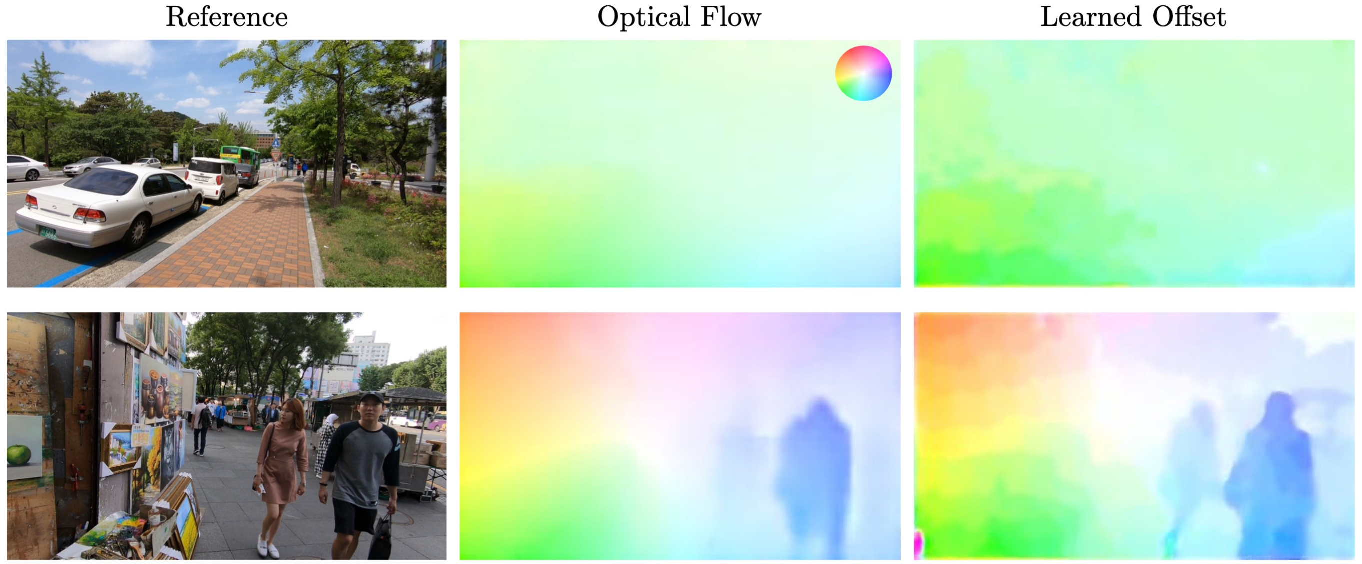
Alignment Methods:

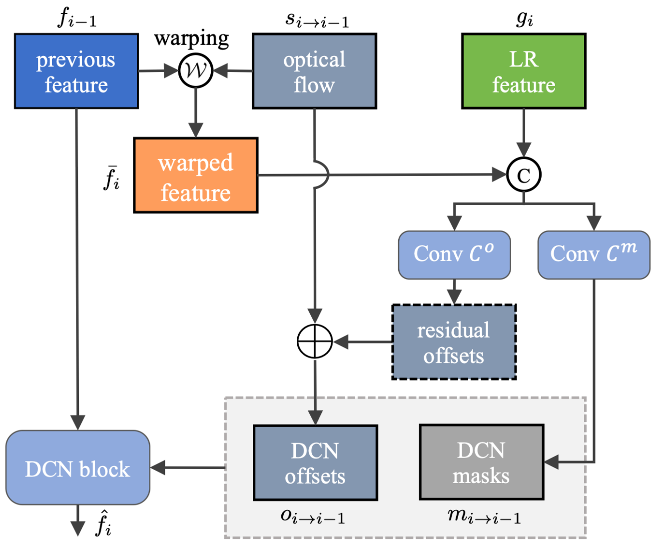
1. Image Alignment.

2. Feature Alignment.

3. Flow Guided Deformable Convolution.

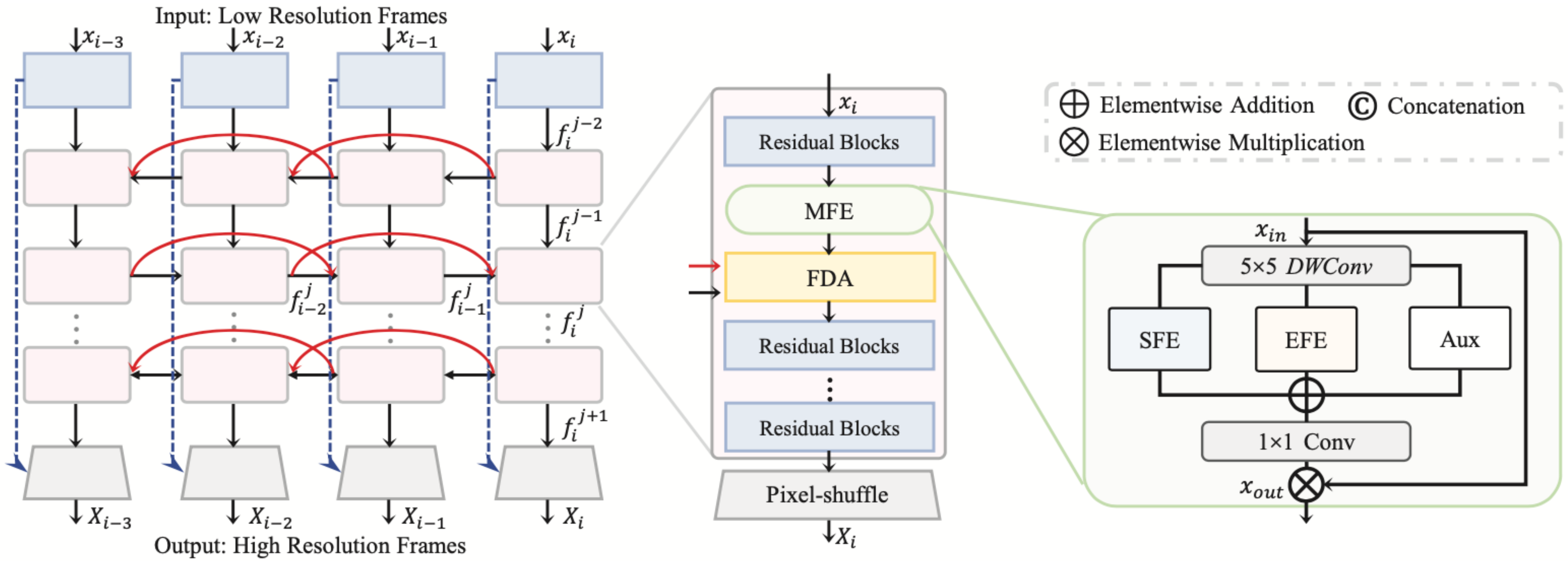
4. No Alignment.





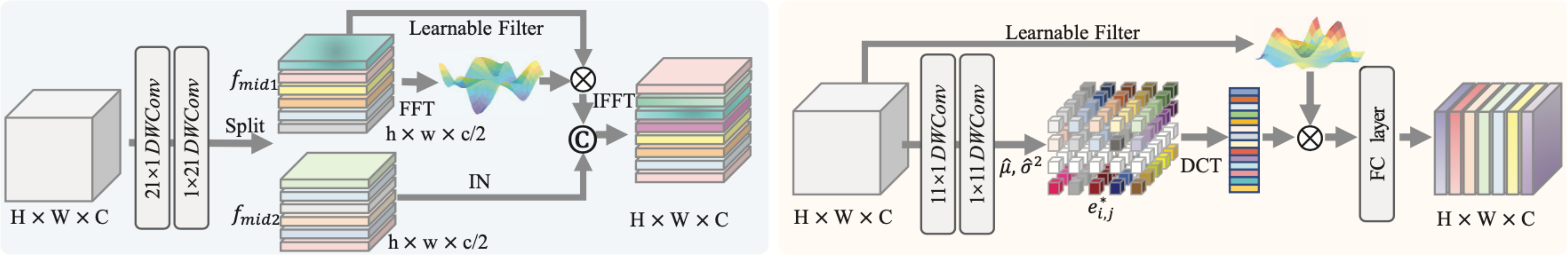
Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding Deformable Alignment in Video Super-Resolution. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 973–981, 2021.

Video Super-Resolution



(a) The Overall Framework of **MFPI**

(b) **MFE**: Multi-frequency Representation Enhancement



(c) **SFE**: Spatial Frequency Representation Enhancement

(d) **EFE**: Energy Frequency Representation Enhancement

Fei Li, Linfeng Zhang, Zikun Liu, Juan Lei, and Zhenbo Li. Multi-frequency Representation Enhancement with Privilege Information for Video Super-Resolution. In Proceedings of the IEEE International Conference on Computer Vision, pages 12814–12825, 2023.

Table 1: Quantitative comparison (PSNR/SSIM). All results are calculated on Y-channel except REDS4 [38] (RGB-channel). The runtime is computed on an LR size of 180×320. A 4× upsampling is performed following previous studies. Blanked entries correspond to results not reported in previous works. Numbers in bold indicate the best performance.

Model	Params (M)	Runtime (ms)	BI degradation			BD degradation		
			REDS4 [38]	Vimeo-90K-T [60]	Vid4[28]	UDM10 [65]	Vimeo-90-T [60]	Vid4 [28]
Bicubic	-	-	26.14/0.7292	31.32/0.8684	23.78/0.6347	28.47/0.8253	31.30/0.8687	21.80/0.5246
VESPCN [1]	-	-	-	-	25.35/0.7557	-	-	-
SPMC [49]	-	-	-	-	25.88/0.7752	-	-	-
TOFlow [61]	-	-	27.98/0.7990	33.08/0.9054	25.89/0.7651	36.26/0.9438	34.62/0.9212	-
FRVSR [45]	5.1	137	-	-	-	37.09/0.9522	35.64/0.9319	26.69/0.8103
DUF [21]	5.8	974	28.63/0.8251	-	-	38.48/0.9605	36.87/0.9447	27.38/0.8329
RBPN [11]	12.2	1507	30.09/0.8590	37.07/0.9435	27.12/0.8180	38.66/0.9596	37.20/0.9458	-
EDVR-M [57]	3.3	118	30.53/0.8699	37.09/0.9446	27.10/0.8186	39.40/0.9663	37.33/0.9484	27.45/0.8406
EDVR [57]	20.6	378	31.09/0.8800	37.61/0.9489	27.35/0.8264	39.89/0.9686	37.81/0.9523	27.85/0.8503
PFNL [65]	3.0	295	29.63/0.8502	36.14/0.9363	26.73/0.8029	38.74/0.9627	-	27.16/0.8355
MuCAN [27]	-	-	30.88/0.8750	37.32/0.9465	-	-	-	-
TGA [18]	5.8	384	-	-	-	-	37.59/0.9516	27.63/0.8423
RLSP [5]	4.2	49	-	-	-	38.48/0.9606	36.49/0.9403	27.48/0.8388
RSDN [17]	6.2	94	-	-	-	39.35/0.9653	37.23/0.9471	27.92/0.8505
RRN [19]	3.4	45	-	-		38.96/0.9644	-	27.69/0.8488
BasicVSR [2]	6.3	63	31.42/0.8909	37.18/0.9450	27.24/0.8251	39.96/0.9694	37.53/0.9498	27.96/0.8553
IconVSR [2]	8.7	70	31.67/0.8948	37.47/0.9476	27.39/0.8279	40.03/0.9694	37.84/0.9524	28.04/0.8570
BasicVSR++ [4]	7.3	77	32.39/0.9069	37.79/0.9530	27.79/0.8400	40.72/0.9722	38.21/0.9550	29.04/0.8753
PSRT [47]	13.4	812	32.72/0.9106	38.27/**0.9536**	28.07/**0.8485**	-	-	-
MFPI (Ours)	7.3	76	**32.81/0.9106**	**38.28**/0.9534	**28.11**/0.8481	**41.08/0.9741**	**38.70/0.9579**	**29.34/0.8781**

Video Super-Resolution

Base	MFE: SFE	MFE: EFE	MFE: Aux	Training: KD	Training: PT	PSNR (dB)	Params (M)	FLOPs (G)
✓						32.39	7.32	280.59
✓	✓					32.55	7.34	280.76
✓		✓				32.52	7.33	280.68
✓	✓	✓				32.67	7.34	280.86
✓	✓	✓	✓			32.69	7.34	281.25
✓	✓	✓	✓	✓		32.31	7.34	281.25
✓	✓	✓	✓		✓	**32.81**	7.34	281.25

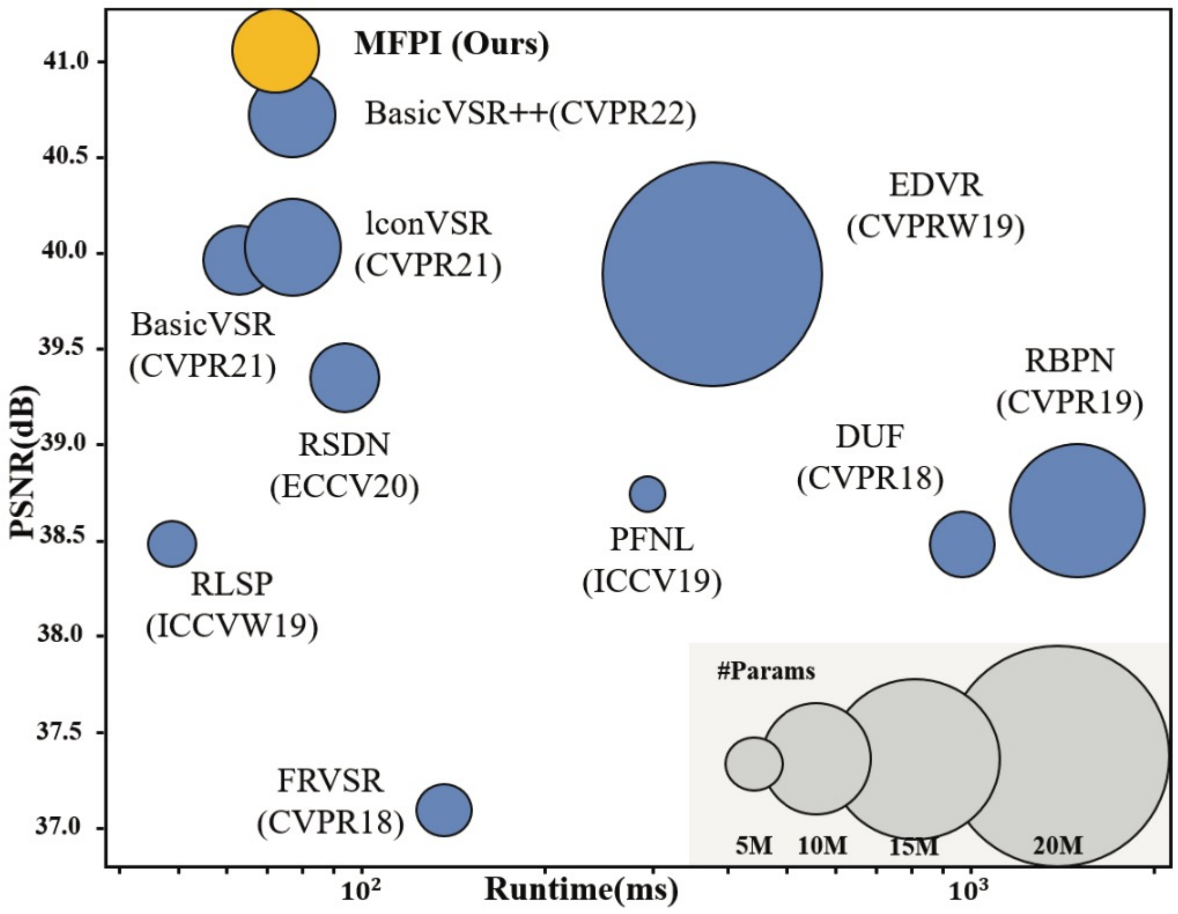
(a) Individual components.

Variant	PSNR
Base + FFT	32.35
Base + FFT w/. Learnable filter	32.48
Base + FFT w/. Learnable filter + original feature	32.49
Base + FFT w/. Learnable filter + BN	32.43
Base + FFT w/. Learnable filter + IN	32.52
Base + FFT w/. Learnable filter + IN + *DWConv* 3 × 3	32.48
Base + FFT w/. Learnable filter + IN + *DWConv* 7 × 7	32.52
Base + FFT w/. Learnable filter + IN + *DWConv* 11 × 11	32.51
Base + FFT w/. Learnable filter + IN + *DWConv* 21 × 21 (Our SFE)	**32.55**
Base + w/. Learnable filter + IN + *DWConv* 21 × 21	31.93

(b) Effects of the SFE branch.

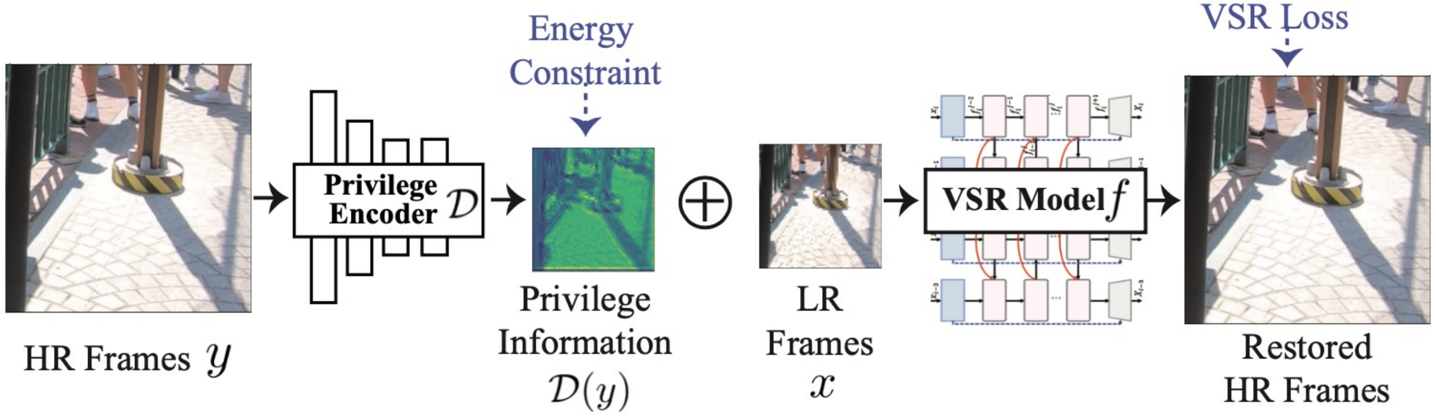
Variant	PSNR
Base + Energy function	32.40
Base + DCT w/. Fixed coefficients	32.32
Base + DCT w/. Learnable filter	32.41
Base + DCT w/. Fixed coefficients + Energy function	32.37
Base + DCT w/. Learnable filter + Energy function	32.50
Base + DCT w/. Learnable filter + Energy function + *DWConv* 3 × 3	32.42
Base + DCT w/. Learnable filter + Energy function + *DWConv* 7 × 7	32.40
Base + DCT w/. Learnable filter + Energy function + *DWConv* 11 × 11 (Our EFE)	**32.52**
Base + DCT w/. Learnable filter + Energy function + *DWConv* 21 × 21	32.43
Base + w/. Learnable filter + Energy function + *DWConv* 11 × 11	32.07

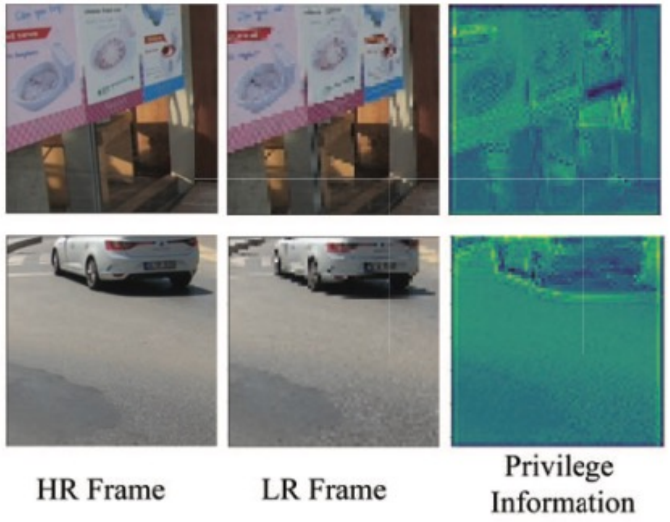
(c) Effects of the EFE branch.



Training Strategy

✓ Privilege Training



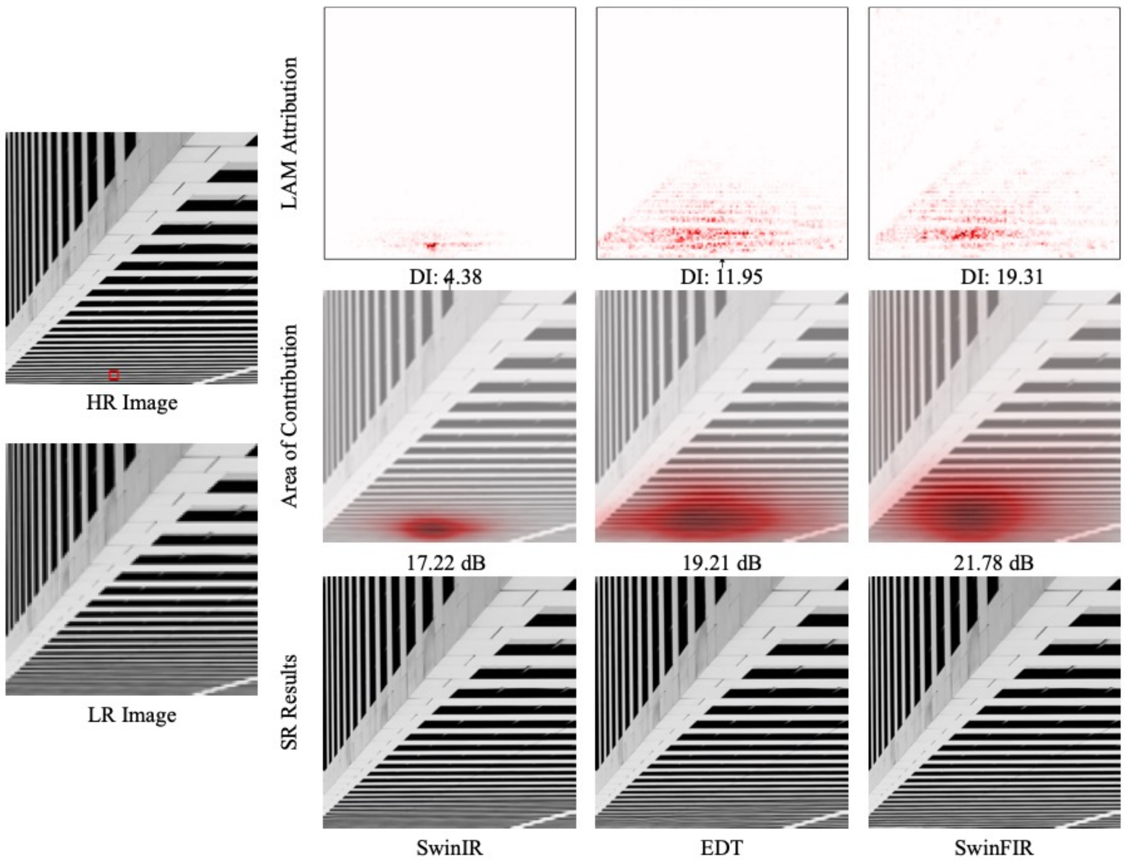


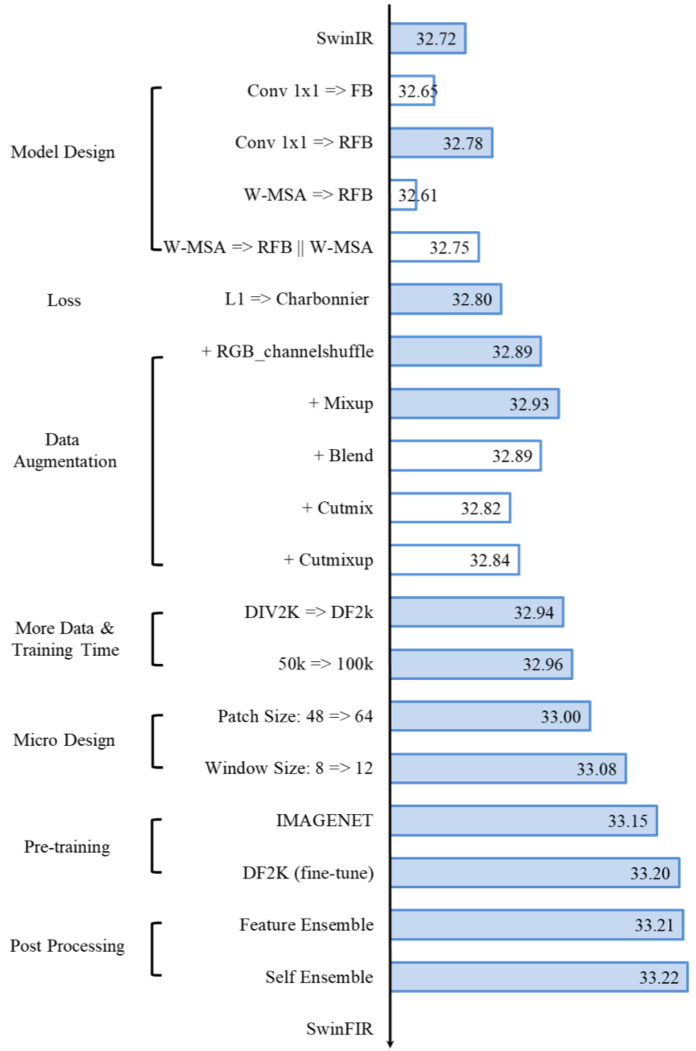
✓ Same-Task Pre-Training Strategy

For example, when we want to train a model for ×4 SR, we first train a ×4 SR model on ImageNet, then fine-tune it on the specific dataset, such as DF2K. It is worth mentioning that **sufficient training iterations for pre-training and an appropriate small learning rate for fine-tuning are very important** for the effectiveness of the pre-training strategy. We think that it is because Transformer requires more data and iterations to learn general knowledge for the task, but needs a small learning rate for fine-tuning to avoid overfitting to the specific dataset.

Data Augmentation

- Rotation and Flip
- RGB channel shuffle, Mixup, Blend, CutMix and CutMixup

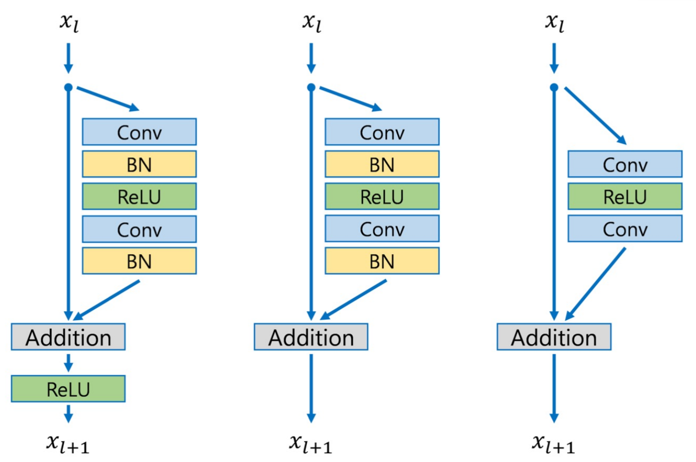




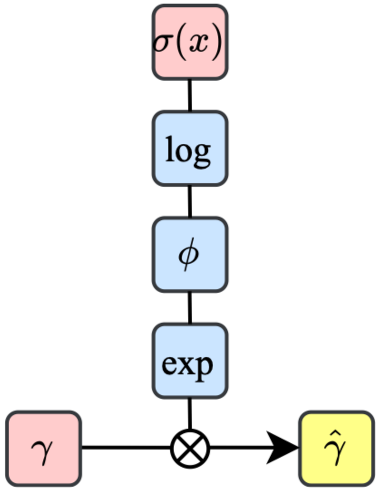
Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin. Swinfir: Revisiting the Swinir with Fast Fourier Convolution and Improved Training for Image Super-Resolution. arXiv preprint arXiv:2208.11247, 2022.

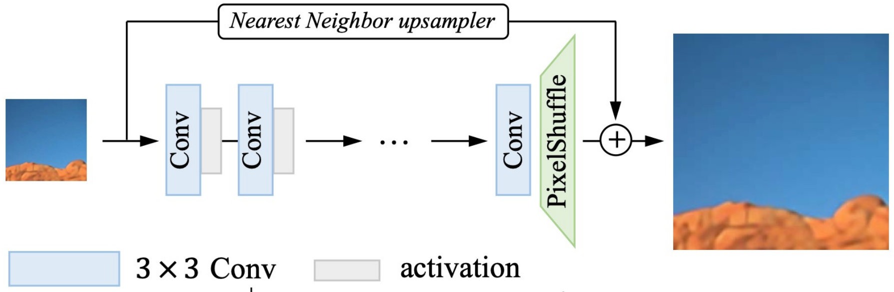
Batch Normalization

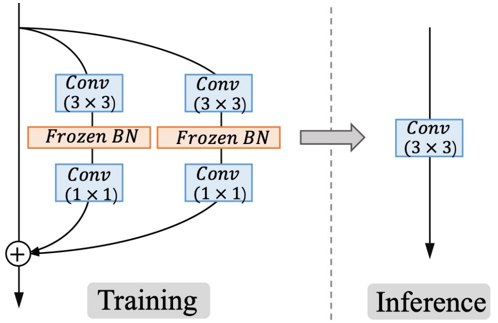
We remove the batch normalization layers from our network as Nah et al.[19] presented in their image deblurring work. Since batch normalization layers normalize the features, they get rid of range flexibility from networks by normalizing the features, it is better to remove them.



(a) Original (b) SRResNet (c) Proposed







Jie Liu, Jie Tang, and Gangshan Wu. Adadm: Enabling Normalization for Image Super-Resolution. arXiv preprint arXiv:2111.13905, 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced Deep Residual Networks for Single Image Super-Resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 136–144, 2017.

Xintao Wang, Chao Dong, and Ying Shan. Repsr: Training Efficient Vgg-Style Super-Resolution Networks with Structural Re-Parameterization and Batch Normalization. In Proceedings of the 30th ACM International Conference on Multimedia, pages 2556–2564, 2022.

Thank You