Xformer: Hybrid X-Shaped Transformer for Image Denoising

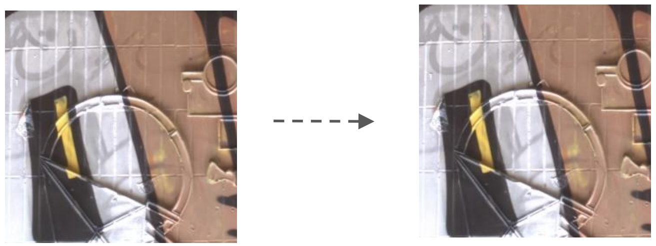
Jiale Zhang[1] , Yulun Zhang[1*] , Jinjin Gu[2,3] , Jiahua Dong[4] , Linghe Kong[1*] , Xiaokang Yang[1]

[1]Shanghai Jiao Tong University, [2]Shanghai AI Laboratory, [3]University of Sydney, [4]Shenyang Institute of Automation, Chinese Academy of Sciences

Introduction & Background

- Image Denoising

Synthetic Image Denoising



Real-world Image Denoising

- CNN-based networks
  - DnCNN, RNAN, RDN, DRUNet ……
- Transformer-based networks
  - SwinIR, IPT, Uformer, Restormer ……

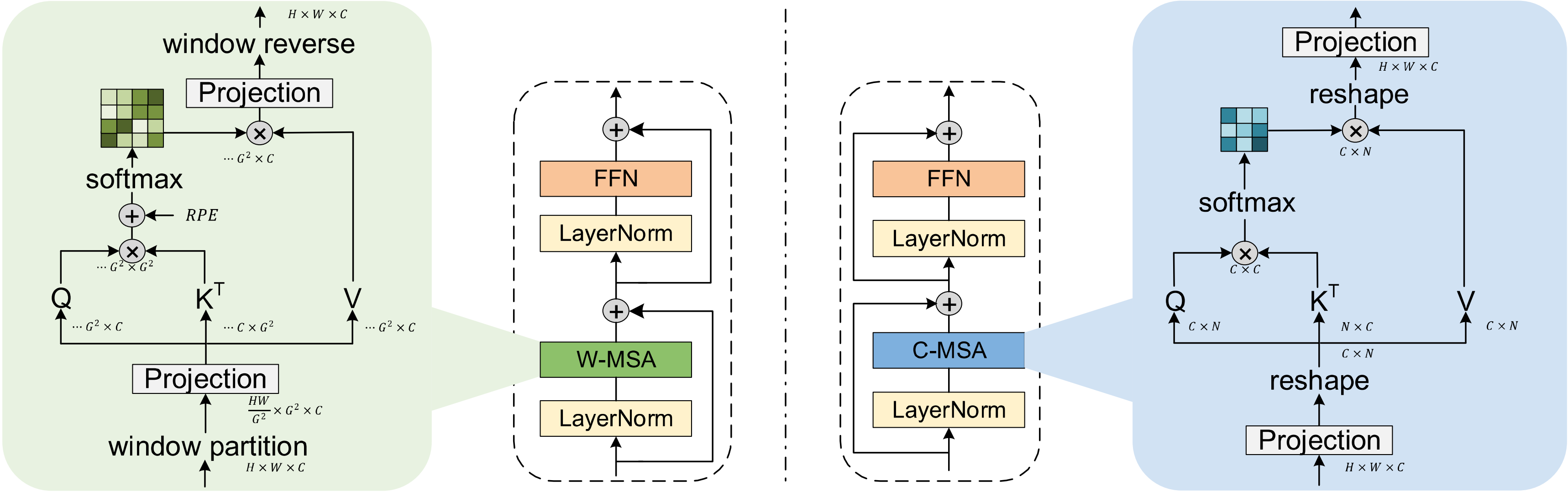
Efficient Self-attention Mechanism

- Spatial-wise Window-based Self-attention
  - Representative work — Swin Transformer
- Channel-wise Cross-covariance Self-attention
  - Representative work — XCiT

Analyses

- Spatial-wise Window-based Self-attention
  - Tokens are defined in spatial dimension.
  - Fine-grained interactions across local patches.
- Channel-wise Cross-covariance Self-attention
  - Tokens are defined in channel dimension
  - Direct interactions across global context patches.

Efficient Self-attention Mechanism



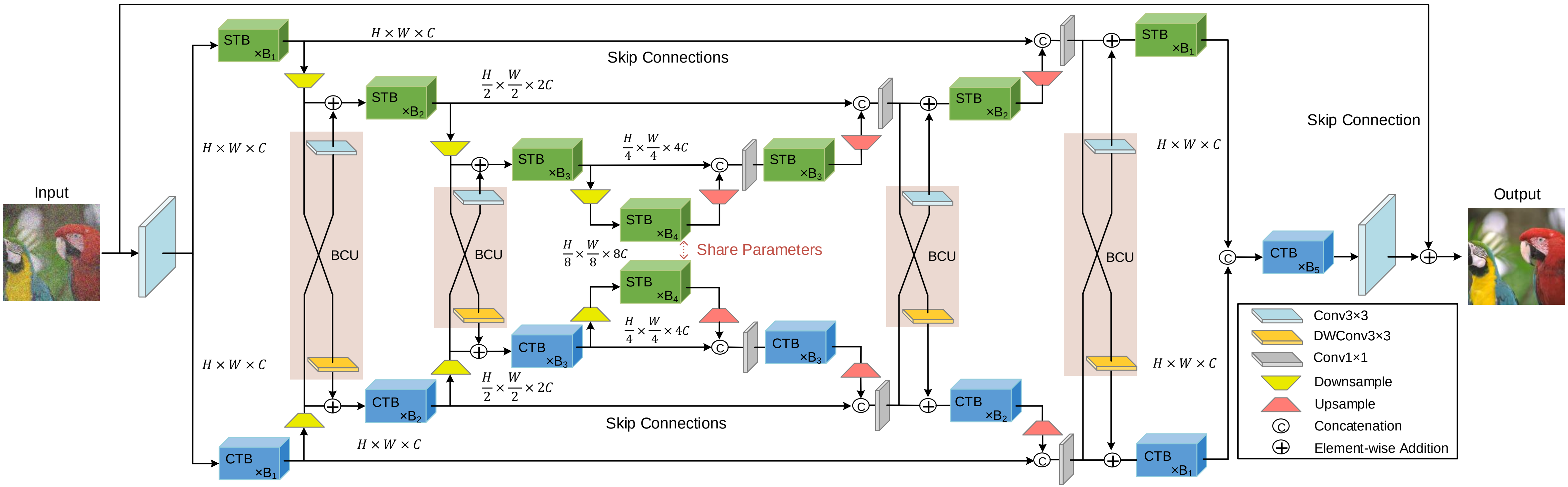
(a) STB (b) CTB

- Spatial-wise Transformer Block
  —capture patch-level information
- Channel-wise Transformer Block
  —capture channel-level information

Gaps exist between these two types of representation learning.

How to adopt them together?

Proposed Method - Xformer



- We propose a concurrent structure network for image denoising.
- The Bidirectional Connection Unit (BCU) bridges the two branches.

Spatial-wise Branch

- W-MSA: window-based multi-head self-attention
  - Q,K,V generated by three linear layers
- FFN: basic multi-layer perception (MLP)

Channel-wise Branch

- C-MSA: channel-wise multi-head self-attention
  - Q,K,V generated by 3×3 depth-wise convolution following 1×1 Conv
- FFN: gating mechanism with depth-wise convolutions[1]

[1] Syed Waqas Zamir, Aditya Arora, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In CVPR, 2022.

Ablation Study

Method	All STB	All CTB	STB+CTB
Params (M)	26.03	28.81	25.23
FLOPs (G)	38.1	42.3	42.2
PSNR (dB)	29.87	29.67	29.94
SSIM	0.8851	0.8830	0.8865

(a) Ablation study of Transformer blocks.

Conclusion: The joint usage of STB and CTB is necessary.

Method	w/o BCU	BCU-1	BCU-2	Complete BCU
Params (M)	24.70	24.71	25.22	25.23
FLOPs (G)	40.9	40.9	42.2	42.2
PSNR (dB)	29.82	29.84	29.92	29.94
SSIM	0.8842	0.8848	0.8859	0.8865

(b) Ablation study of BCU settings.

Conclusion: The BCU bridges the network in a interactive manner and greatly enhances the performance.

Method	w/o Shift	w/ Shift
Params (M)	25.23	25.23
FLOPs (G)	42.2	42.2
PSNR (dB)	29.88	29.94
SSIM	0.8852	0.8865

(c) Whether to use shift.

ID	STB	CTB	BCU	Structure	Params (M)	FLOPs (G)	PSNR (dB)	SSIM
1	✓			single-branch	26.48	40.6	29.84	0.8853
2		✓		single-branch	26.11	38.7	29.68	0.8829
3	✓	✓		two-branches	24.70	40.9	29.82	0.8842
4	✓	✓	✓	two-branches	25.23	42.2	29.94	0.8865

(d) Ablation study of designed models with different branches.

Conclusion: The shift operation can bring performance improvement.

Conclusion: The direct connection of dual branches brings limited performance. Equipped with BCU, the performance is greatly enhanced. Therefore, the effective information is very important.

Experiment

- Gaussian Image Denoising

Dataset	σ	BM3D	DnCNN	IRCNN	FFDNet	RNAN	RDN	DRUNet	P3AN	IPT	SwinIR	Restormer	Xformer (ours)
CBSD68	15	33.52	33.90	33.86	33.87	-	-	34.30	-	-	34.42	34.40	**34.43**
	25	30.71	31.24	31.16	31.21	-	-	31.69	-	-	31.78	31.79	**31.82**
	50	27.38	27.95	27.86	27.96	28.27	28.31	28.51	28.37	28.39	28.56	28.60	**28.63**
Kodak24	15	34.28	34.60	34.69	34.63	-	-	35.31	-	-	35.34	*35.35	**35.39**
	25	32.15	32.14	32.18	32.13	-	-	32.89	-	-	32.89	*32.93	**32.99**
	50	28.46	28.95	28.93	28.98	29.58	29.66	29.86	29.69	29.64	29.79	*29.87	**29.94**
McMaster	15	34.06	33.45	34.58	34.66	-	-	35.40	-	-	35.61	35.61	**35.68**
	25	31.66	31.52	32.18	32.35	-	-	33.14	-	-	33.20	33.34	**33.44**
	50	28.51	28.62	28.91	29.18	29.72	-	30.08	-	29.98	30.22	30.30	**30.38**
Urban100	15	33.93	32.98	33.78	33.83	-	-	34.81	-	-	35.13	35.13	**35.29**
	25	31.36	30.81	31.20	31.40	-	-	32.60	-	-	32.90	32.96	**33.21**
	50	27.93	27.59	27.70	28.05	29.08	29.38	29.61	29.51	29.71	29.82	30.02	**30.36**

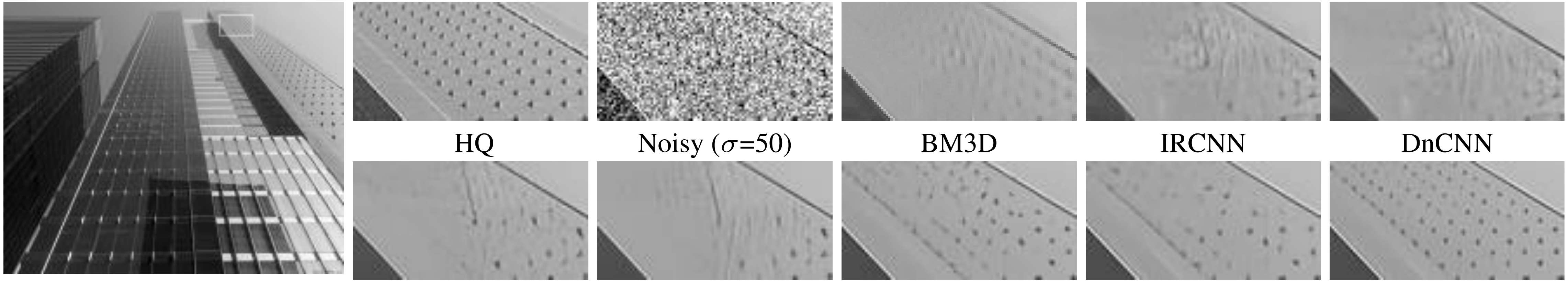
Dataset	σ	BM3D	DnCNN	IRCNN	FFDNet	NLRN	MWCNN	RNAN	RDN	DRUNet	SwinIR	Restormer	Xformer (ours)
Set12	15	32.37	32.86	32.76	32.75	33.16	33.15	-	-	33.25	33.36	33.42	**33.46**
	25	29.97	30.44	30.37	30.43	30.80	30.79	-	-	30.94	31.01	31.08	**31.16**
	50	26.72	27.18	27.12	27.32	27.64	27.74	27.70	27.60	27.90	27.91	28.00	**28.10**
BSD68	15	31.08	31.73	31.63	31.63	31.88	31.86	-	-	31.91	31.97	31.96	**31.98**
	25	28.57	29.23	29.15	29.19	29.41	29.41	-	-	29.48	29.50	29.52	**29.55**
	50	25.60	26.23	26.19	26.29	26.47	26.53	26.48	26.41	26.59	26.58	26.62	**26.65**
Urban100	15	32.35	32.64	32.46	32.40	33.45	33.17	-	-	33.44	33.70	33.79	**33.98**
	25	29.70	29.95	29.80	29.90	30.94	30.66	-	-	31.11	31.30	31.46	**31.78**
	50	25.95	26.26	26.22	26.50	27.49	27.42	27.65	27.40	27.96	27.98	28.29	**28.71**

Experiment

- Real Image Denoising

Dataset	Method	BM3D	DnCNN	CBDNet	RIDNet	AINDNet	VDN	SADNet	DANet	CycleISP	MIRNet	DeamNet	DAGL	MAXIM	Uformer	Restormer	Xformer
SIDD	PSNR	25.65	23.66	30.78	38.71	39.08	39.28	39.46	39.47	39.52	39.72	39.47	38.94	39.96	39.89	**40.02**	<u>39.98</u>
	SSIM	0.685	0.583	0.801	0.951	0.954	0.956	0.957	0.957	0.957	0.959	0.957	0.953	0.960	0.960	<u>0.960</u>	**0.960**
DND	PSNR	34.51	32.43	38.06	39.26	39.37	39.38	39.59	39.58	39.56	39.88	39.63	39.77	39.84	<u>40.04</u>	40.03	**40.19**
	SSIM	0.851	0.790	0.942	0.953	0.951	0.952	0.952	0.955	0.956	0.956	0.953	0.956	0.954	0.956	<u>0.956</u>	**0.957**

- Visual Comparison



Urban100: img_033

HQ | Noisy (σ=50) | BM3D | IRCNN | DnCNN

RNAN | RDN | SwinIR | Restormer | **Xformer (ours)**

Contributions

(1) We propose Xformer, an X-Shaped Transformer with hybrid implementation of spatial-wise and channel-wise Transformer blocks, thereby exploiting the stronger global representation of tokens.

(2) We propose the Bidirectional Connection Unit (BCU) that is able to effectively couple the learned representations from two branches of Xformer. This simple design significantly enhances the global information modeling of our method.

(3) We employ Xformer to train an efficient and effective Transformer-based network for image denoising. We conduct extensive experiments on the synthetic and real-world denoising tasks. Our method achieves state-of-the-art performance.

Thanks!