CHANNEL-SPATIAL TRANSFORMER FOR EFFICIENT IMAGE SUPER-RESOLUTION

Jiuqiang Li, Shilei Zhu

School of Computing and Artificial Intelligence, Southwest Jiaotong University, China

Introduction

Motivation:

1. Most of existing methods overlook the mutual influence and facilitation between the channel and spatial aspects.
2. The feed-forward network (FFN) used in the Transformer architecture during the feature extraction process hinders the feature representation ability due to the presence of redundant information within the channels and ignores spatial information modeling.

Our work: We propose the Channel-Spatial Transformer (CST), which combines channel and spatial perspectives in self-attention to extract more reliable deep features.

Methodology

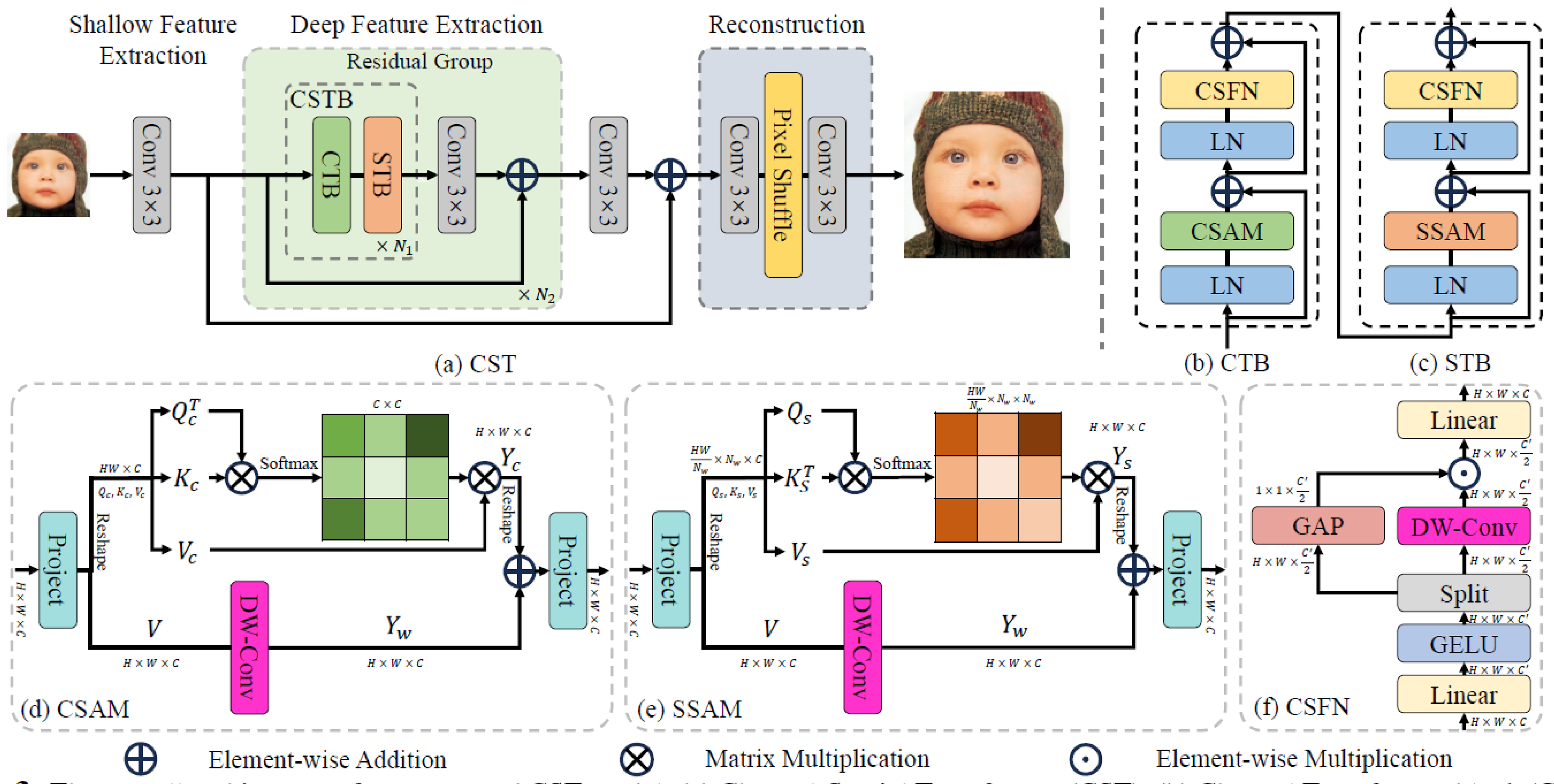

Fig. 3: The overall architecture of our proposed CST model. (a) Channel-Spatial Transformer (CST). (b) Channel Transformer block (CTB). (c) Spatial Transformer block (STB). (d) Channel Self-Attention Module (CSAM). (e) Spatial Self-Attention Module (SSAM). (f) Channel-Spatial Feed-Forward Network (CSFN).

Channel-Spatial Transformer Block

$$X_l' = CSAM\big(LN(X_{l-1})\big) + X_{l-1}$$

$$X_l = CSFN\big(LN(X_l')\big) + X_l'$$

$$X_{l+1}' = SSAM\big(LN(X_l)\big) + X_l$$

$$X_{l+1} = CSFN\big(LN(X_{l+1}')\big) + X_{l+1}'$$

Channel-Spatial Feed-Forward Network

$$\hat{X}' = \sigma\big(W_p^1 \hat{X}\big), \qquad [\hat{X}_1', \hat{X}_2'] = \hat{X}'$$

$$CSFN(\hat{X}') = W_p^2\big(H_{GP}(\hat{X}_1') \odot (W_d \hat{X}_2')$$

Experiments

Overall Performance Comparison

Table 1: Quantitative comparison (PSNR/SSIM) with state-of-the-art methods. The best and second-best results are coloured red and blue.

Method	Scale	Bicubic -	EDSR [12]	RCAN [13]	SAN [14]	IGNN [10]	RFANet [15]	HAN [16]	CSNLN [17]	DGSM-Swin [18]	NLSA [19]	ELAN [8]	DFSA [20]	SwinIR [7]	CAT-A [9]	CST (ours)
Set5 [21]	×2	33.66/0.9299	38.11/0.9602	38.27/0.9614	38.31/0.9620	38.24/0.9613	38.26/0.9615	38.27/0.9614	38.28/0.9616	38.30/0.9618	38.34/0.9618	38.36/0.9620	38.38/0.9620	38.42/0.9623	38.51/0.9626	38.57/0.9630
	×3	30.39/0.8682	34.65/0.9280	34.74/0.9299	34.75/0.9300	34.72/0.9298	34.79/0.9300	34.75/0.9299	34.74/0.9300	34.83/0.9307	34.85/0.9306	34.90/0.9313	34.92/0.9312	34.97/0.9318	35.06/0.9326	35.15/0.9332
	×4	28.42/0.8104	32.46/0.8968	32.63/0.9002	32.64/0.9003	32.57/0.8998	32.66/0.9004	32.64/0.9002	32.68/0.9004	32.70/0.9014	32.59/0.9000	32.75/0.9022	32.79/0.9019	32.92/0.9044	33.08/0.9052	33.11/0.9055
Set14 [22]	×2	30.24/0.8688	33.92/0.9195	34.12/0.9216	34.07/0.9213	34.07/0.9217	34.16/0.9220	34.16/0.9217	34.12/0.9223	34.19/0.9230	34.08/0.9231	34.20/0.9228	34.33/0.9232	34.46/0.9250	34.78/0.9265	34.81/0.9270
	×3	27.55/0.7742	30.52/0.8462	30.65/0.8482	30.59/0.8476	30.66/0.8484	30.67/0.8487	30.67/0.8483	30.66/0.8482	30.69/0.8504	30.70/0.8485	30.80/0.8504	30.83/0.8507	30.93/0.8534	31.04/0.8538	31.09/0.8549
	×4	26.00/0.7027	28.80/0.7876	28.87/0.7889	28.92/0.7888	28.85/0.7891	28.88/0.7894	28.90/0.7890	28.95/0.7888	28.97/0.7917	28.87/0.7891	28.96/0.7914	29.06/0.7922	29.09/0.7950	29.18/0.7960	29.23/0.7972
BSD100 [23]	×2	29.56/0.8431	32.32/0.9013	32.41/0.9027	32.42/0.9028	32.41/0.9025	32.41/0.9026	32.41/0.9027	32.40/0.9024	32.43/0.9029	32.43/0.9027	32.45/0.9030	32.50/0.9036	32.53/0.9041	32.59/0.9047	32.61/0.9050
	×3	27.21/0.7385	29.25/0.8093	29.32/0.8111	29.33/0.8112	29.31/0.8105	29.34/0.8115	29.32/0.8110	29.33/0.8105	29.35/0.8127	29.34/0.8117	29.38/0.8124	29.42/0.8128	29.46/0.8145	29.52/0.8160	29.55/0.8168
	×4	25.96/0.6675	27.71/0.7420	27.77/0.7436	27.78/0.7436	27.77/0.7434	27.79/0.7442	27.80/0.7442	27.80/0.7439	27.83/0.7452	27.78/0.7444	27.83/0.7459	27.87/0.7458	27.92/0.7489	27.99/0.7510	28.01/0.7515
Urban100 [24]	×2	26.88/0.8403	32.93/0.9351	33.34/0.9384	33.10/0.9370	33.23/0.9383	33.33/0.9389	33.35/0.9385	33.25/0.9386	33.18/0.9462	33.42/0.9394	33.44/0.9391	33.66/0.9412	33.81/0.9427	34.26/0.9440	34.36/0.9458
	×3	24.46/0.7349	28.80/0.8653	29.09/0.8702	28.93/0.8671	29.03/0.8696	29.15/0.8720	29.10/0.8705	29.13/0.8712	29.15/0.8725	29.25/0.8726	29.32/0.8745	29.44/0.8761	29.75/0.8826	30.12/0.8862	30.18/0.8884
	×4	23.14/0.6577	26.64/0.8033	26.82/0.8087	26.79/0.8068	26.84/0.8090	26.92/0.8112	26.85/0.8094	27.22/0.8168	27.06/0.8142	26.96/0.8109	27.13/0.8167	27.17/0.8163	27.45/0.8254	27.89/0.8339	27.92/0.8343
Manga109 [25]	×2	30.80/0.9339	39.10/0.9773	39.44/0.9786	39.32/0.9792	39.35/0.9786	39.44/0.9783	39.46/0.9785	39.37/0.9785	39.60/0.9790	39.59/0.9789	39.62/0.9793	39.98/0.9798	39.92/0.9797	40.10/0.9805	40.32/0.9808
	×3	26.95/0.8556	34.17/0.9476	34.44/0.9499	34.30/0.9494	34.39/0.9496	34.59/0.9506	34.48/0.9500	34.45/0.9502	34.59/0.9511	34.57/0.9508	34.73/0.9517	35.07/0.9525	35.12/0.9537	35.38/0.9546	35.57/0.9553
	×4	24.89/0.7866	31.02/0.9148	31.22/0.9173	31.18/0.9169	31.28/0.9182	31.41/0.9187	31.42/0.9177	31.43/0.9201	31.58/0.9216	31.27/0.9184	31.68/0.9226	31.88/0.9266	32.03/0.9260	32.39/0.9285	32.51/0.9292

Visual Results

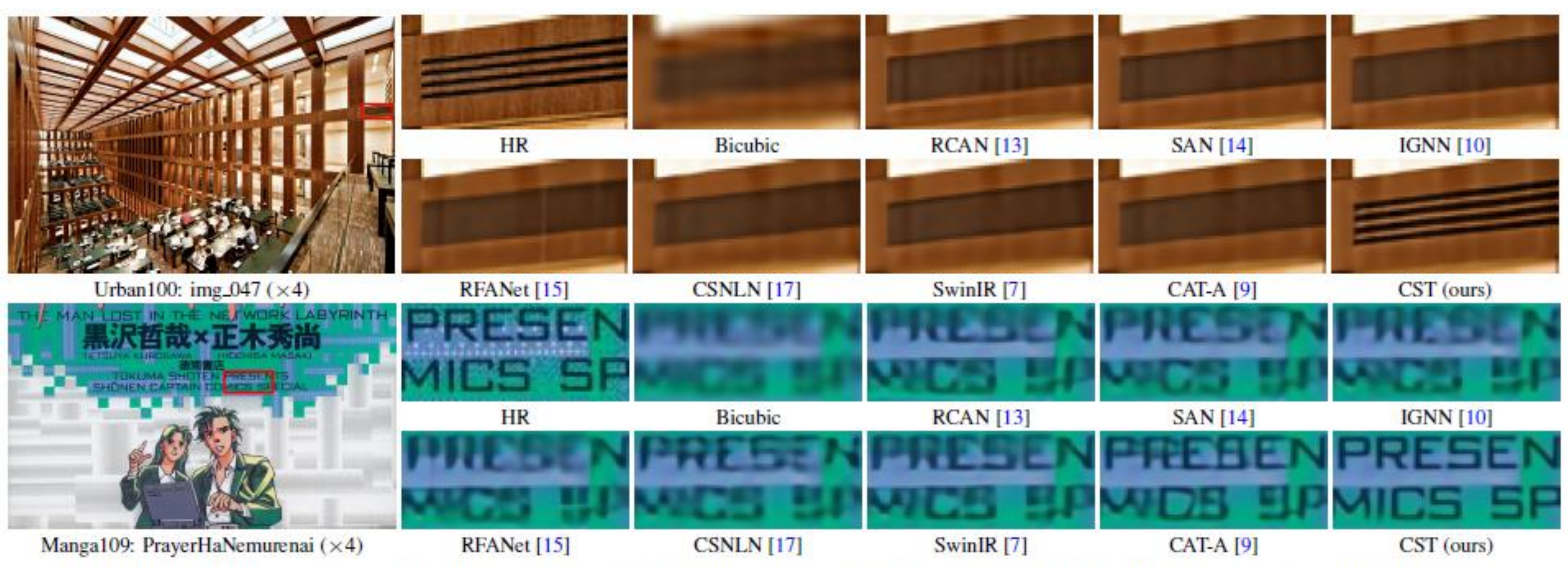

Fig. 4: Qualitative comparison for image SR (×4). The patches for comparison are marked with red boxes in the original images.

Ablation Study on our CSFN

Model	Params (M)	FLOPs (G)	PSNR (dB)	SSIM
FFN	14.95	250.49	33.83	0.9419
CSFN w/o Conv	12.56	215.91	33.74	0.9408
CSFN w/o GAP	12.68	216.55	33.81	0.9415
CSFN w/o Split	16.14	256.48	33.96	0.9424
CSFN	12.80	218.37	34.36	0.9458

Table 3: Ablation study of proposed CSFN on Urban100 (×2).

Model complexity comparisons

Method	EDSR [12]	RCAN [13]	HAN [16]	CSNLN [17]	SwinIR [7]	CAT-A [9]	CST (ours)
Params (M)	43.09	15.59	16.07	6.57	11.90	16.60	12.80
FLOPs (G)	823.34	261.01	269.13	84,155.24	215.32	360.67	218.37
Urban100	26.64	26.82	26.85	27.22	27.45	27.89	27.92
Manga109	31.02	31.22	31.42	31.43	32.03	32.39	32.51

Table 4: Model complexity comparisons (×4). PSNR (dB) on Urban100 and Manga109, FLOPs, and Params are reported.

Ablation Study on our CSTB

CSAM	SSAM	Params (M)	FLOPs (G)	PSNR (dB)	SSIM
✓		12.79	215.94	33.91	0.9415
	✓	12.81	221.87	34.11	0.9442
✓	✓	12.80	218.37	34.36	0.9458

Table 2: Ablation study of our self-attention module on Urban100 (×2).

Conclusion

In this paper, we propose a channel-spatial Transformer (CST) for efficient image super-Resolution. Our CST extracts channel and spatial features through the cross-interaction of channel attention information and spatial attention information, enabling powerful representation capabilities. Specifically, the alternating channel self-attention and spatial selfattention form a sequence of consecutive Channel-Spatial Transformer Blocks (CSTBs). CST models global dependencies and extracts alternatively fused features from the channel and spatial dimensions through these CSTBs. Additionally, CST includes the Channel-Spatial Feed-Forward Network (CSFN) to enhance each CSTB and facilitate more effective interaction between channel and spatial information. Extensive experiments demonstrate that CST outperforms previous methods in terms of computational cost and performance.