DTNSR: Deep Transformer Network for Single Image Super-Resolution

Electronic Supplementary Material

Jagrati Talreja, *Graduate Student Member, IEEE*, Supavadee Aramvith, *Senior Member, IEEE* and Takao Onoye, *Senior Member, IEEE*

A. Initial Feature Extraction and Patch embedding

As seen in Figure 1, the initial features of the input are extracted using a normal 3 × 3 convolution, and then patch embedding is applied to the convoluted input. Equations S1 and S2 demonstrate the initial feature extraction stage.

$$H_0 = H_{Conv}(H_{LR}), \tag{S1}$$

Here, $H_{Conv}(.)$ represents 3 × 3 convolution operation. H_{LR} is the input Low-Resolution (LR) image, fed to the normal convolution for extracting initial features, and H_0 is the output of the convolution layer.

$$H_P = P_{embed}(H_0), \tag{S2}$$

$P_{embed}(.)$ represents Patch Embedding, and H_P results from embeddings obtained after applying Patch embedding. After obtaining the initial features, patch embedding is applied on H_0, as depicted in Equation 2. *L* and *P* are the embedding dimensions of the input token of subsequent Transformer.

B. Local Feature Window Transformer Block (LFWT)

$$M_{SWMSA} = H_{SWMSA}(H_{LN}(H_{I1})) + H_{I1}, \tag{S3}$$

Here, H_{I1} is the input to the Local Feature Window Transformer Block, $H_{LN}(.)$ is the Layer Norm function, $(H_{SWMSA})(.)$ is the Shifted Window Multi-head Self-Attention function, M_{SWMSA} is the output of the SW-MSA [19] module in the LFWT Block.

$$H_{LFWT} = H_{MLP}(H_{LN}(M_{SWMSA})) + M_{SWMSA}, \tag{S4}$$

In Equation 4, $H_{MLP}(.)$ is the output function of the Multi-layer Proton, and H_{LFWT} represents the output of the Local Feature Window Transformer (LFWT) Block.

C. Xception Block

$$\begin{aligned} H_X = {} & \text{ReLU}\big(H_{\text{Conv}}(H_{\text{DWConv}}(H_{\text{LFWT}}))\big) \\ & + \text{ReLU}\big(H_{\text{Conv}}(H_{\text{DWConv}}(H_{\text{LFWT}}))\big), \end{aligned} \tag{S5}$$

Here, $H_{\text{DWConv}}(.)$ is the Depthwise Separable convolution operation function, $H_{\text{Conv}}(.)$ is the Pointwise convolution operation function, $ReLU(.)$ represents the non-linearity, and H_X represents the output of the Xception Block.

D. Multi-Layer Feature Fusion Block

As the features from multiple paths enter the MLFF block, they are concatenated, as shown in Equation S6.

$$MLFF_{Cat} = cat(H_{S1}, H_{S2}, H_{S3}, H_{D1}, H_{D2}, H_{D3}), \tag{S6}$$

In Equation 6, $MLFF_{Cat}$ denotes the output of the concatenation in the MLFF Block, $cat(.)$ is the concatenation operation, $H_{S1}, H_{S2}, H_{S3}, H_{D1}, H_{D2}, H_{D3}$ are the shallow and dense features from paths S1, S2, S3, D1, D2, and D3.

After this, these concatenated features are passed through a Depthwise Separable convolution layer and a Fully Connected layer to reduce the computational burden.

$$MLFF_{DW} = H_{DWConv}(H_{FC}(MLFF_{Cat})), \tag{S7}$$

In Equation 7, $H_{FC}(.)$ is the Fully Connected layer function, $MLFF_{DW}$ is the result obtained in the MLFF Block by the Depthwise Separable convolution.

After applying the non-linearity and residual connection, the generated features are then concatenated together.

$$MLFF_{FC} = cat(ReLU(MLFF_{DW}), MLFF_{DW}), \tag{S8}$$

$MLFF_{FC}$, as seen in Equation 8, is the input of the last Fully Connected layer in the MLFF Block.

The concatenated output is then passed to the Fully Connected layer to better generalize the data.

$$MLFF_O = H_{FC}(MLFF_{FC}), \tag{S9}$$

$MLFF_O$ in Equation 9 is the output of the MLFF Block. Finally, the Multi-Layer Feature Fusion Block (MLFF) output is fed to the deconvolution layer to reconstruct the high-resolution image.

$$HR_O = H_{DConv}(MLFF_O), \tag{S10}$$

Lastly, in Equation 10, $H_{DConv}(.)$ is the deconvolution operation, and HR_O is the generated high-resolution output.

TABLE S1

IMPACT OF VARYING WINDOW SIZES ON PERFORMANCE FOR $\times 4$ UPSCALING FACTOR. **RED** INDICATES THE BEST QUANTITATIVE VALUE, WHEREAS THE BLUE INDICATES THE SECOND-BEST QUANTITATIVE VALUE.

Method	Window Size	Set5 [38]		Set14 [39]		BSD100 [40]		Urban100 [41]		Manga109 [42]	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SwinIR [17]	8×8	38.24	0.9615	33.94	0.9212	32.39	0.9023	33.09	0.9373	39.34	0.9784
	16×16	38.32	0.9618	34.00	0.9212	32.44	0.9030	33.40	0.9394	39.53	0.9791
	24×24	38.35	0.9620	34.04	0.9214	**32.48**	**0.9034**	33.54	0.9402	39.71	**0.9798**
SRFormer [19]	8×8	38.20	0.9611	34.06	0.9214	32.36	0.9021	32.92	0.9361	39.10	0.9777
	16×16	38.31	0.9617	34.10	0.9217	32.43	0.9026	33.26	0.9385	39.36	0.9785
	24×24	38.38	0.9621	34.13	**0.9228**	32.44	0.9030	33.51	0.9405	39.49	0.9788
DTNSR (Ours)	8×8	38.26	0.9616	34.08	0.9215	32.38	0.9020	33.06	0.9368	39.22	0.9782
	16×16	38.34	0.9620	34.11	0.9217	32.44	0.9028	33.44	0.9398	39.51	0.9793
	24×24	**38.40**	**0.9622**	**34.14**	0.9222	**32.48**	**0.9034**	**33.68**	**0.9409**	**39.86**	0.9796

TABLE S2

PERFORMANCE EVALUATION FOR NOISE DEGRADATION OF IMAGES ON URBAN [41] FOR SCALE FACTOR $\times 2$. THE BEST QUANTITATIVE VALUE HAS BEEN RECORDED AS BOLD WITH **RED** COLOR. THE SECOND BEST QUANTITATIVE VALUE IS SHOWN IN BLUE COLOR WITH AN UNDERLINE.

Methods / Noise Level	BM3D[46]	FFDNet [47]	NCSR [48]	DnCNN [23]	DTNSR (Our)
$\sigma = 5$	31.18	31.34	31.56	31.67	**31.78**
$\sigma = 10$	29.61	29.76	29.44	29.82	29.89
$\sigma = 15$	28.12	28.48	28.64	28.58	28.77