SwinIR: Image Restoration Using Swin Transformer

Le Van The

20110746

J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image Restoration Using Swin Transformer." *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021.

1. Basic knowledge

❖ Self-attention matrices

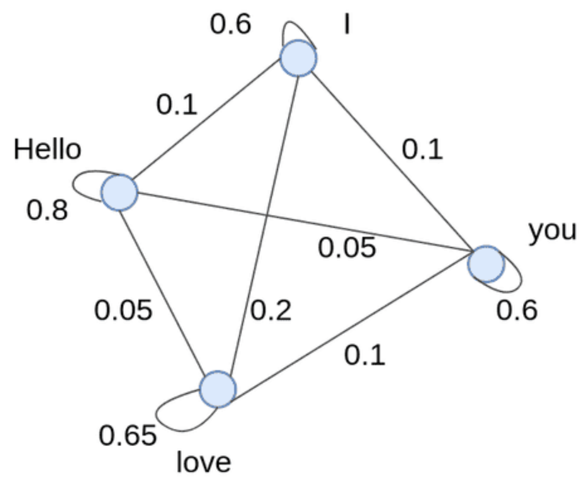
- Self-attention matric shows the self-correlation in the sequence.

Self-attention
Probability score matrix

	Hello	I	love	you
Hello	0.8	0.1	0.05	0.05
I	0.1	0.6	0.2	0.1
love	0.05	0.2	0.65	0.1
you	0.2	0.1	0.1	0.6

Softmax(Attention)
equation

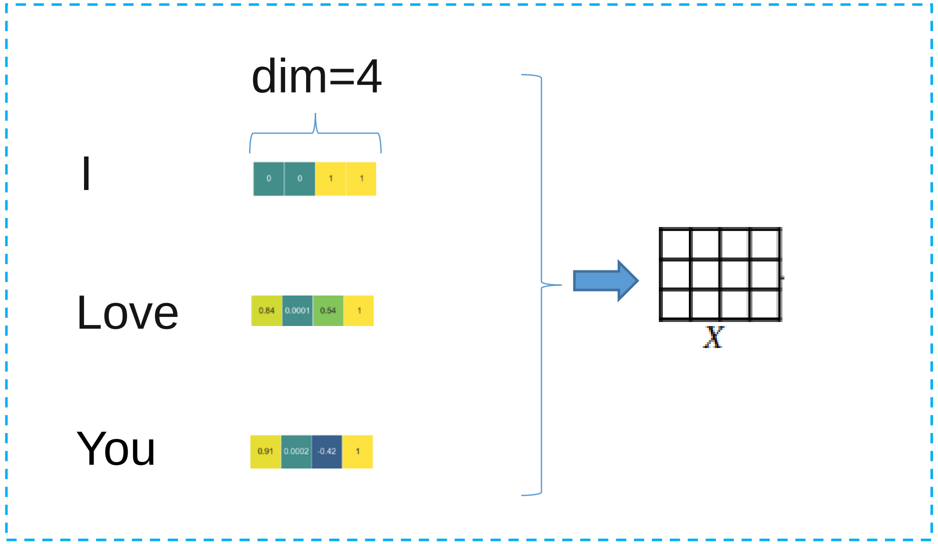
Self-attention as a
undirected weighted graph

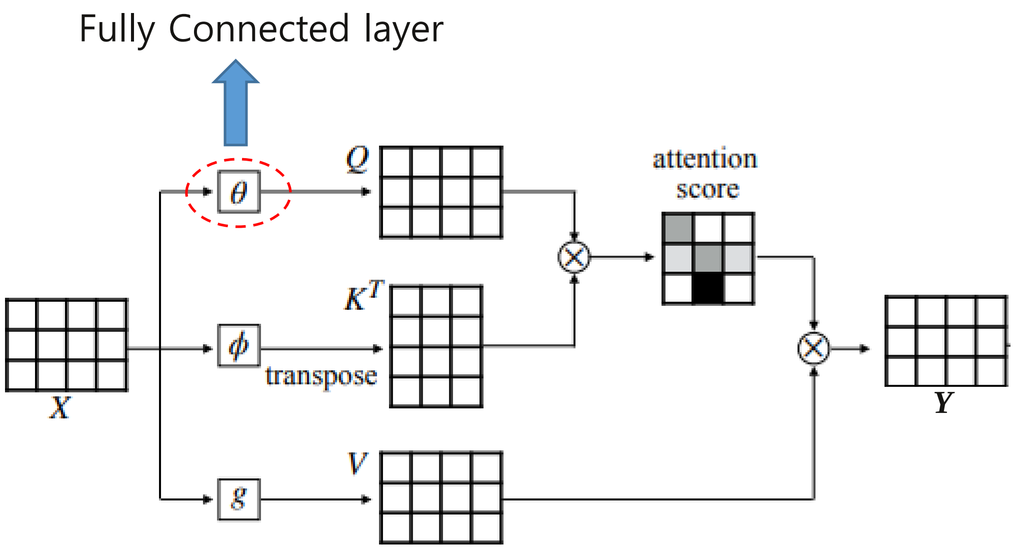


1. Basic knowledge

❖ Self-attention

Positional Embeddings





$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

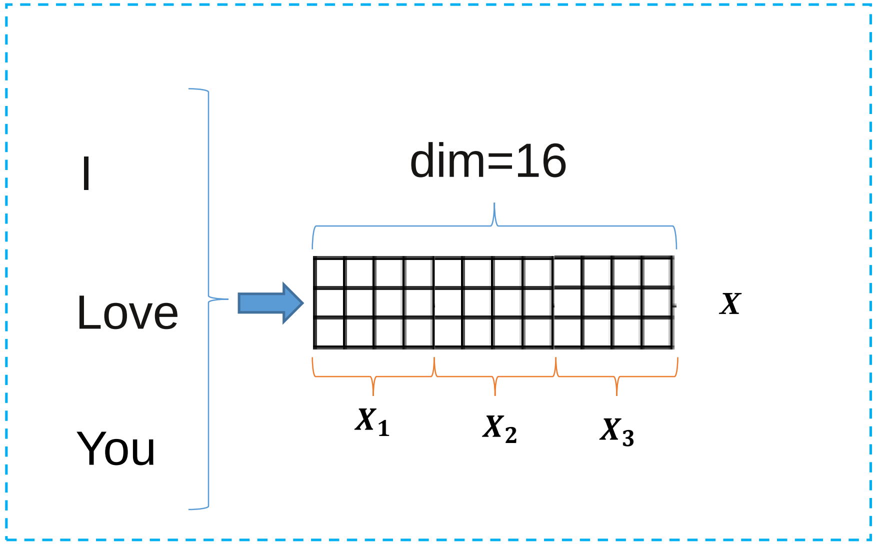
Positional Embeddings present the words under numerical number array

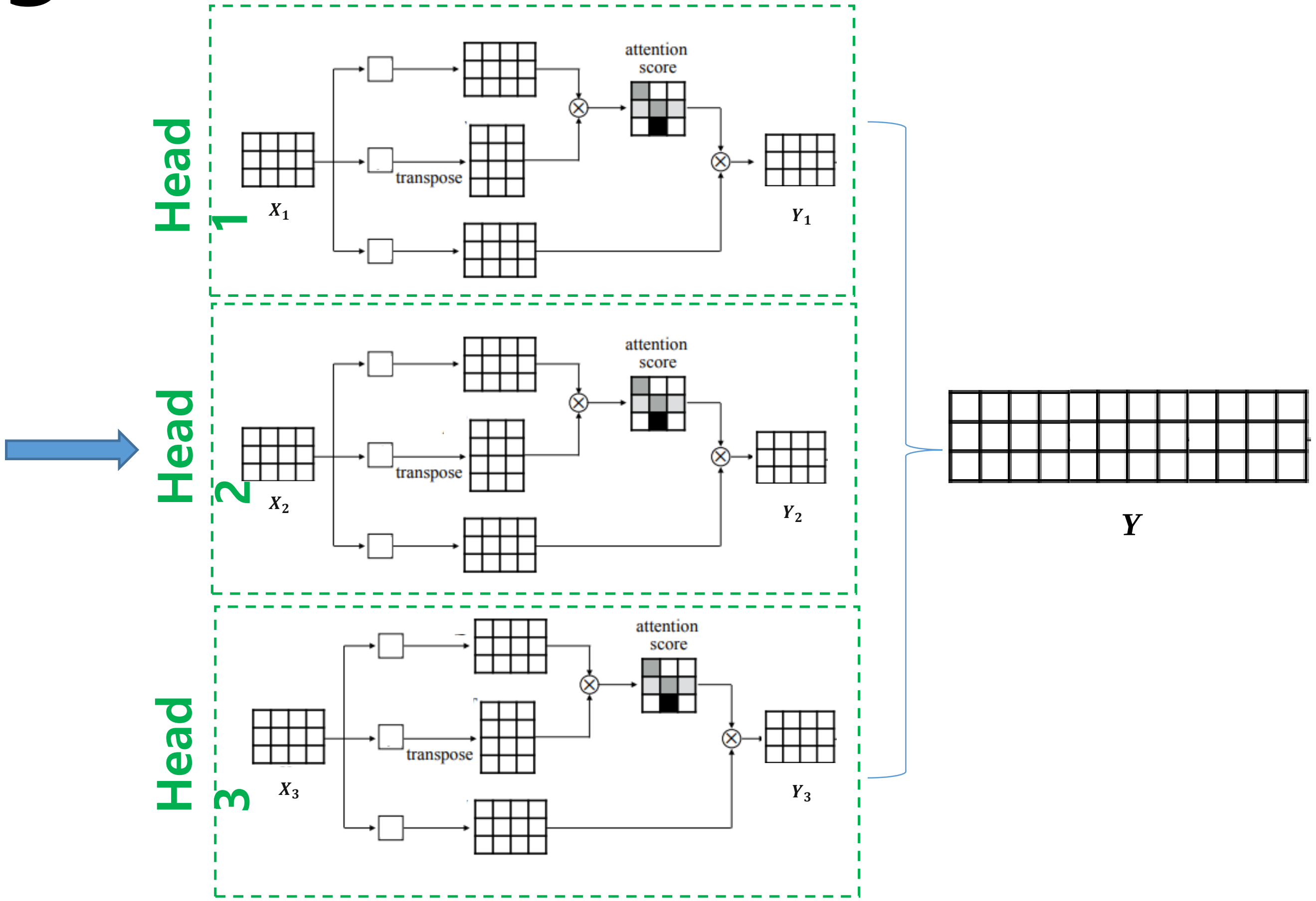
X is the input word sequence, and we calculate three values from that which is **Q(Query)**, **K(Key)** and **V(Value)**.

1. Basic knowledge

❖ Multi-head self-attention

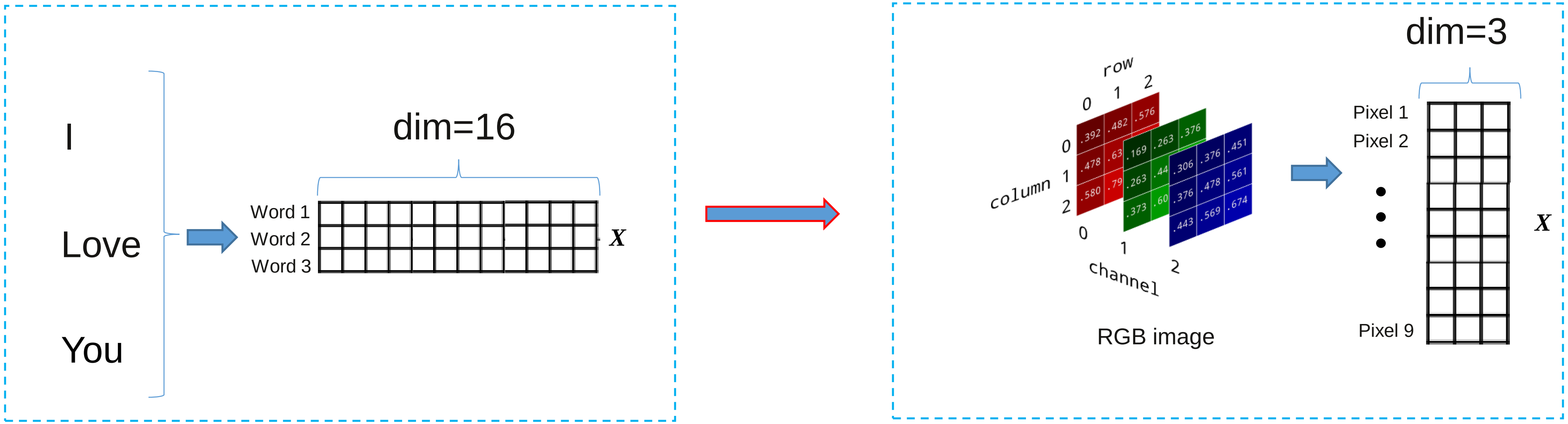
Positional Embeddings





2. Image consideration

❖ Multi-head self-attention for image

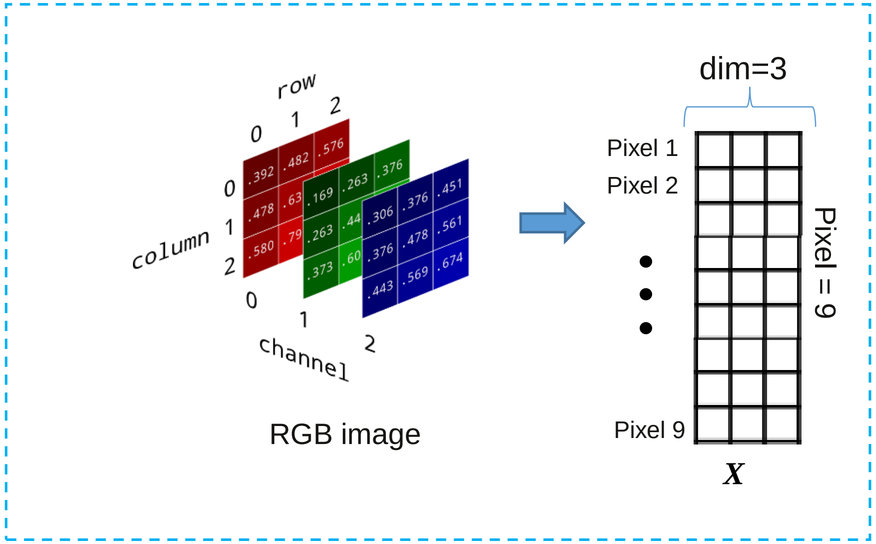


Positional Embeddings for wo rds

“Embeddings” for pixel value

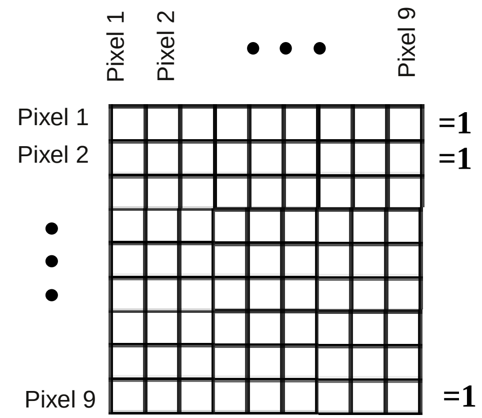
2. Image consideration

❖ Multi-head self-attention for image



“Embeddings” for pixel value

Self-attention
Probability score matrix



Attention matrices for pixels

Self-attention
Probability score matrix

	Hello	I	love	you	
Hello	0.8	0.1	0.05	0.05	
I	0.1	0.6	0.2	0.1	=1
love	0.05	0.2	0.65	0.1	=1
you	0.2	0.1	0.1	0.6	=1

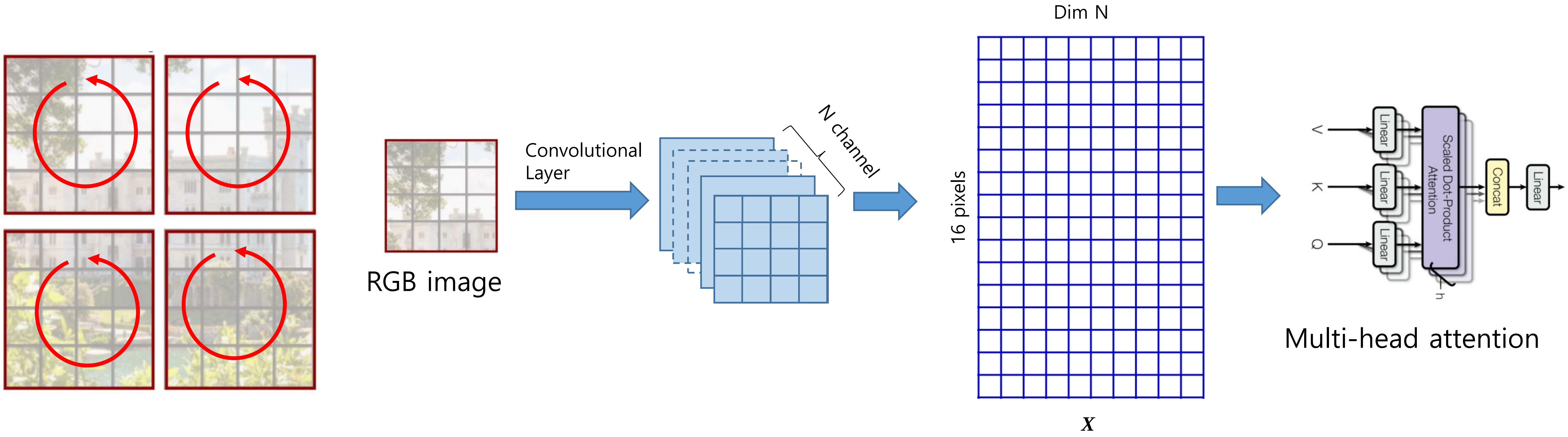
Softmax(Attention)
equation

Attention matrices for sequences

2. Image consideration

❖ Local windows

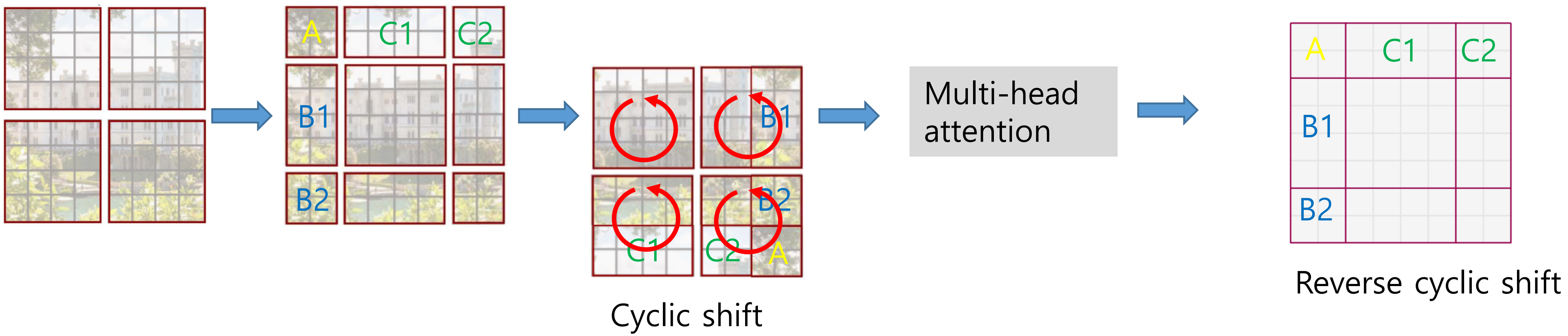
- Swin Transformer will split image into many patches with fixed size.
- → it just learn local correlation



•**W-MSA:** window multi-head self attention module

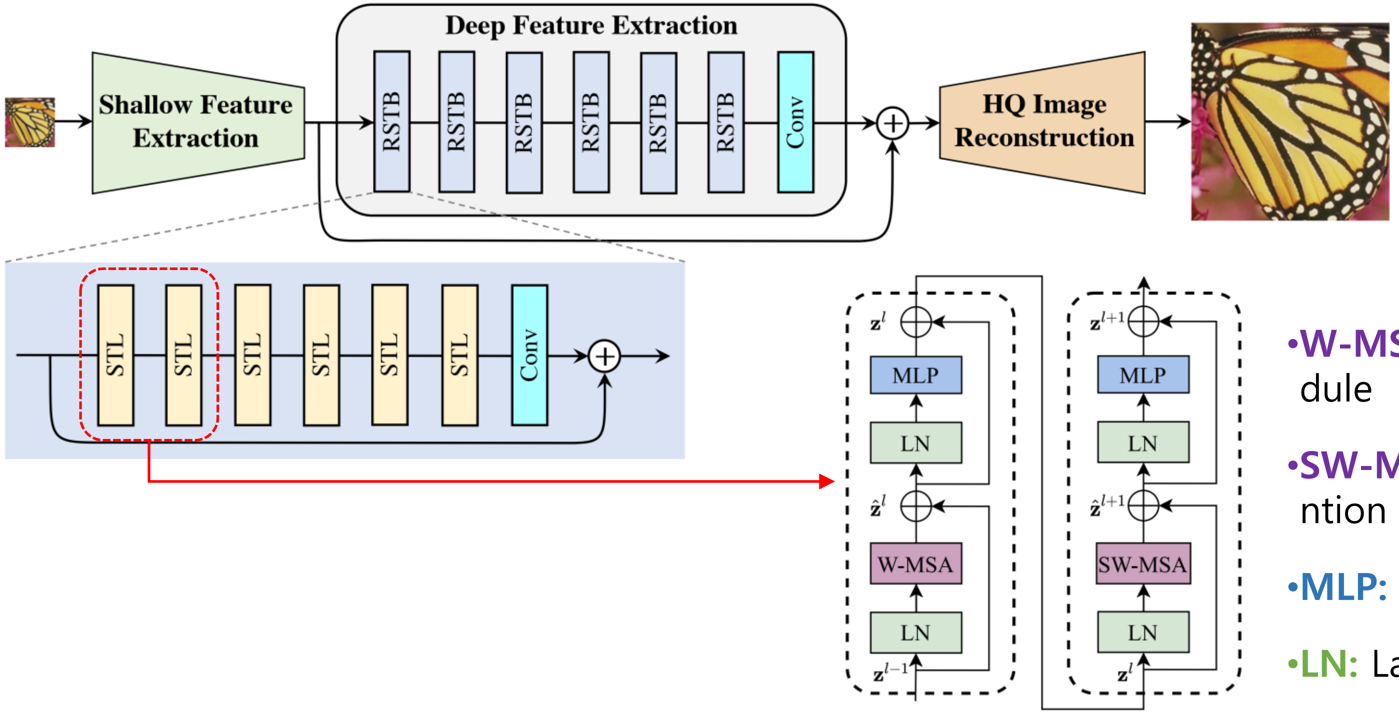
2. Image consideration

❖ Shifted window partitioning



•**SW-MSA:** shifted windowing multi-head self-attention mod ule

2. SwinIR: Image Restoration Using Swin Transformer



- **W-MSA:** window multi-head self attention mo dule
- **SW-MSA:** shifted window multi-head self atte ntion module
- **MLP:** Multi-Layer Perceptron Layer
- **LN:** LayerNorm

The architecture of the proposed SwinIR for image resto ration.

2. SwinIR: Image Restoration Using Swin Transformer

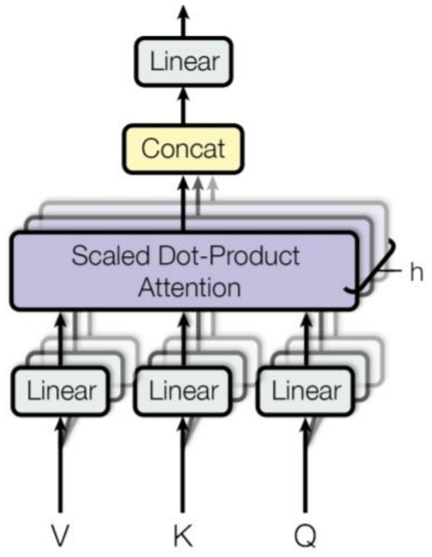
❖ Before

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

❖ Swin Transformer

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T/\sqrt{d} + \boxed{B})V,$$

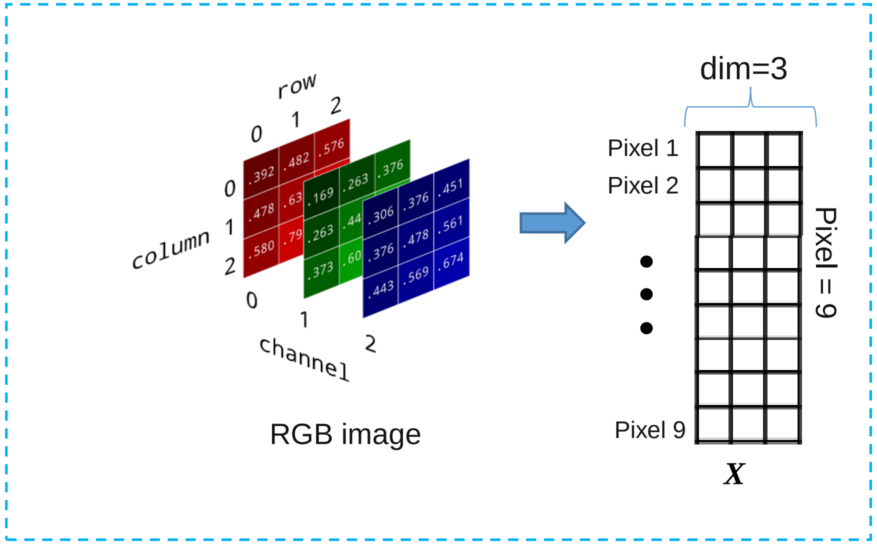
where B is the learnable relative positional bias



2. SwinIR: Image Restoration Using Swin Transformer

❖ Relative positional bias

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T/\sqrt{d} + \boxed{B})V,$$



Loss position information

"Embeddings" for pixel value

2. SwinIR: Image Restoration Using Swin Transformer

❖ **Relative positional bias**

$$\mathrm{Attention}(Q,K,V)=\mathrm{SoftMax}(QK^T/\sqrt{d}+\boxed{B})V$$

Window size (M) = 3

x axis

0	1	2	3
1	4	5	6
2	7	8	9

x axis matrix

	1	2	3	4	5	6	7	8	9
1	0	0	0	-1	-1	-1	-2	-2	-2
2	0	0	0	-1	-1	-1	-2	-2	-2
3	0	0	0	-1	-1	-1	-2	-2	-2
4	1	1	1	0	0	0	-1	-1	-1
5	1	1	1	0	0	0	-1	-1	-1
6	1	1	1	0	0	0	-1	-1	-1
7	2	2	2	1	1	1	0	0	0
8	2	2	2	1	1	1	0	0	0
9	2	2	2	1	1	1	0	0	0

y axis

1	2	3
4	5	6
7	8	9
0	1	2

y axis matrix

	1	2	3	4	5	6	7	8	9
1	0	-1	-2	0	-1	-2	0	-1	-2
2	1	0	-1	1	0	-1	1	0	-1
3	2	1	0	2	1	0	2	1	0
4	0	-1	-2	0	-1	-2	0	-1	-2
5	1	0	-1	1	0	-1	1	0	-1
6	2	1	0	2	1	0	2	1	0
7	0	-1	-2	0	-1	-2	0	-1	-2
8	1	0	-1	1	0	-1	1	0	-1
9	2	1	0	2	1	0	2	1	0

2. SwinIR: Image Restoration Using Swin Transformer

❖ **Relative positional bias**

$$\text{Attention}(Q,K,V) = \text{SoftMax}(QK^T/\sqrt{d} + \boxed{B})V,$$

x axis matrix

	1	2	3	4	5	6	7	8	9
1	0	0	0	-1	-1	-1	-2	-2	-2
2	0	0	0	-1	-1	-1	-2	-2	-2
3	0	0	0	-1	-1	-1	-2	-2	-2
4	1	1	1	0	0	0	-1	-1	-1
5	1	1	1	0	0	0	-1	-1	-1
6	1	1	1	0	0	0	-1	-1	-1
7	2	2	2	1	1	1	0	0	0
8	2	2	2	1	1	1	0	0	0
9	2	2	2	1	1	1	0	0	0

y axis matrix

	1	2	3	4	5	6	7	8	9
1	0	-1	-2	0	-1	-2	0	-1	-2
2	1	0	-1	1	0	-1	1	0	-1
3	2	1	0	2	1	0	2	1	0
4	0	-1	-2	0	-1	-2	0	-1	-2
5	1	0	-1	1	0	-1	1	0	-1
6	2	1	0	2	1	0	2	1	0
7	0	-1	-2	0	-1	-2	0	-1	-2
8	1	0	-1	1	0	-1	1	0	-1
9	2	1	0	2	1	0	2	1	0

```
# Step 1
x_axis_matrix += window_size - 1
y_axis_matrix += window_size - 1

# Step 2
x_axis_matrix *= 2 * window_size - 1
relative_position_M = x_axis_matrix + y_axis_matrix
```

Relative Position index

12	11	10	7	6	5	2	1	0
13	12	11	8	7	6	3	2	1
14	13	12	9	8	7	4	3	2
17	16	15	12	11	10	7	6	5
18	17	16	13	12	11	8	7	6
19	18	17	14	13	12	9	8	7
22	21	20	17	16	15	12	11	10
23	22	21	18	17	16	13	12	11
24	23	22	19	18	17	14	13	12

2. SwinIR: Image Restoration Using Swin Transformer

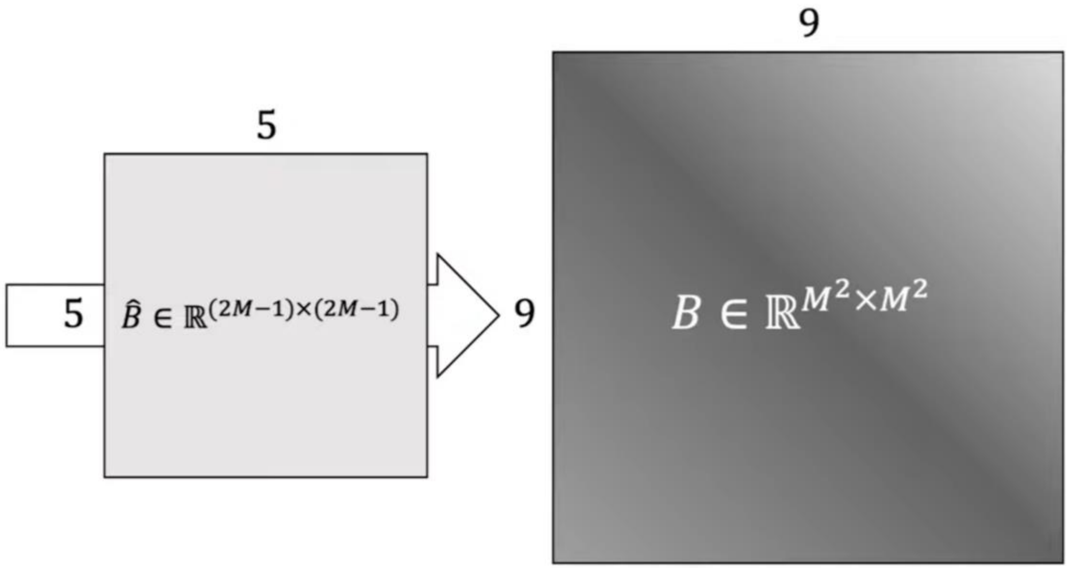
❖ **Relative positional bias**

$$\text{Attention}(Q,K,V) = \text{SoftMax}(QK^T/\sqrt{d} + \boxed{B})V,$$

Window size (M) = 3

Relative Position index

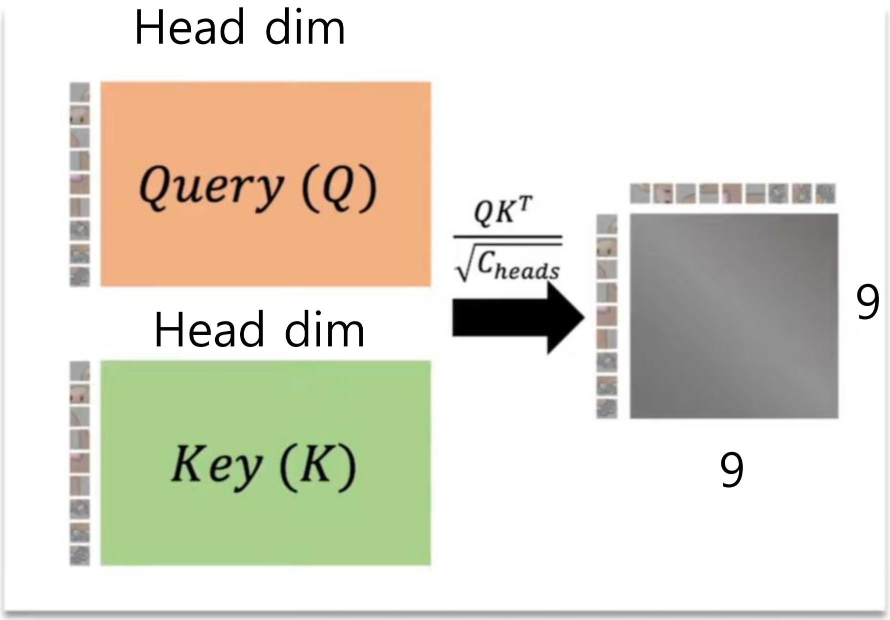
12	11	10	7	6	5	2	1	0
13	12	11	8	7	6	3	2	1
14	13	12	9	8	7	4	3	2
17	16	15	12	11	10	7	6	5
18	17	16	13	12	11	8	7	6
19	18	17	14	13	12	9	8	7
22	21	20	17	16	15	12	11	10
23	22	21	18	17	16	13	12	11
24	23	22	19	18	17	14	13	12

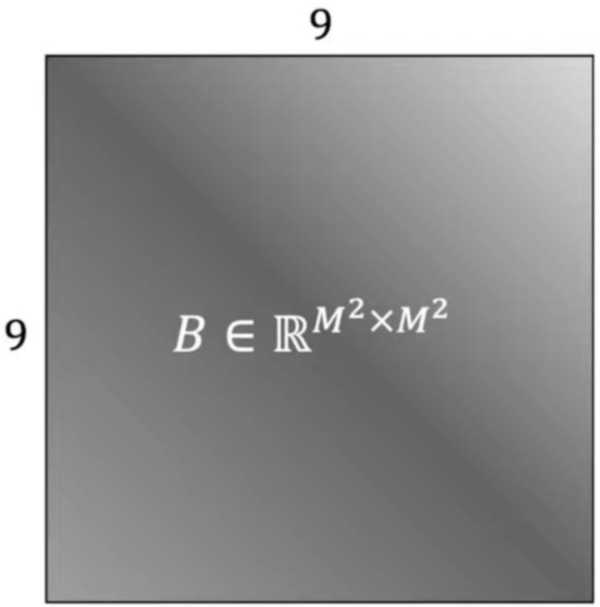


- Save number of parameters
- Show Relative position

2. SwinIR: Image Restoration Using Swin Transformer

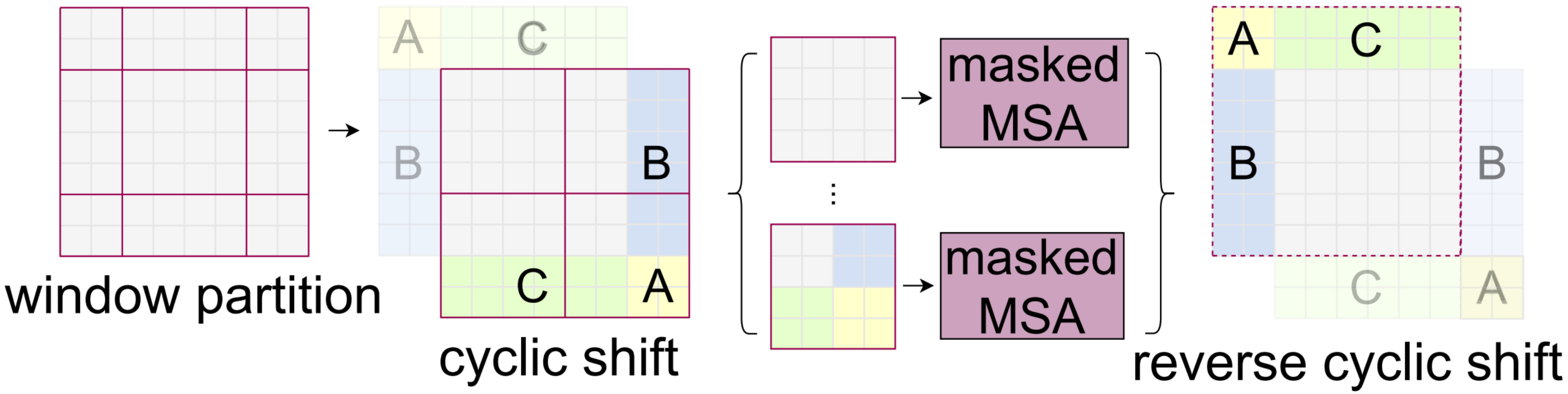
❖ Relative positional bias $\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T/\sqrt{d} + \boxed{B})V,$





2. SwinIR: Image Restoration Using Swin Transformer

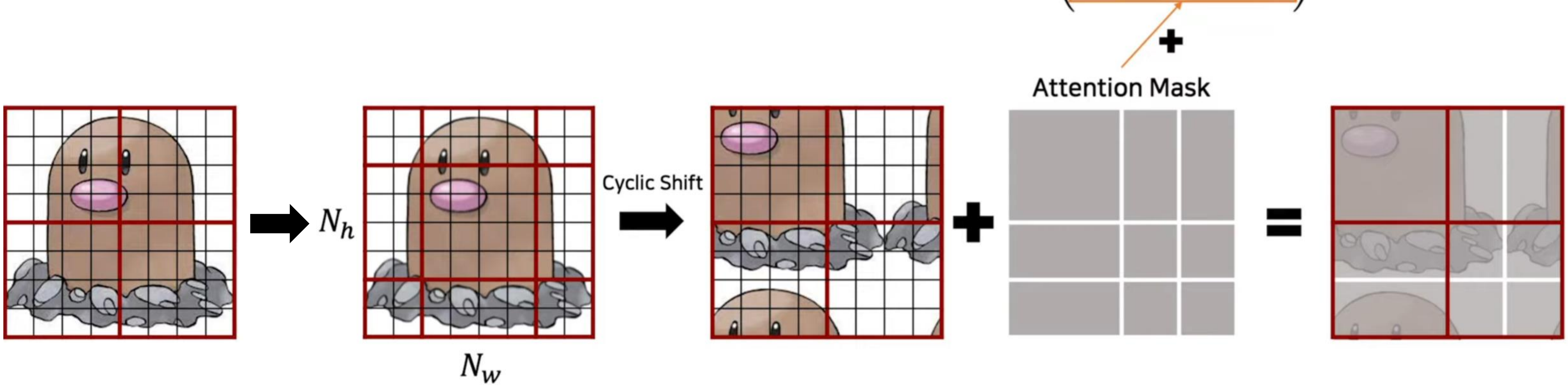
❖ SW-MSA: shifted windowing multi-head self attention module



3. SwinIR: Image Restoration Using Swin Transformer

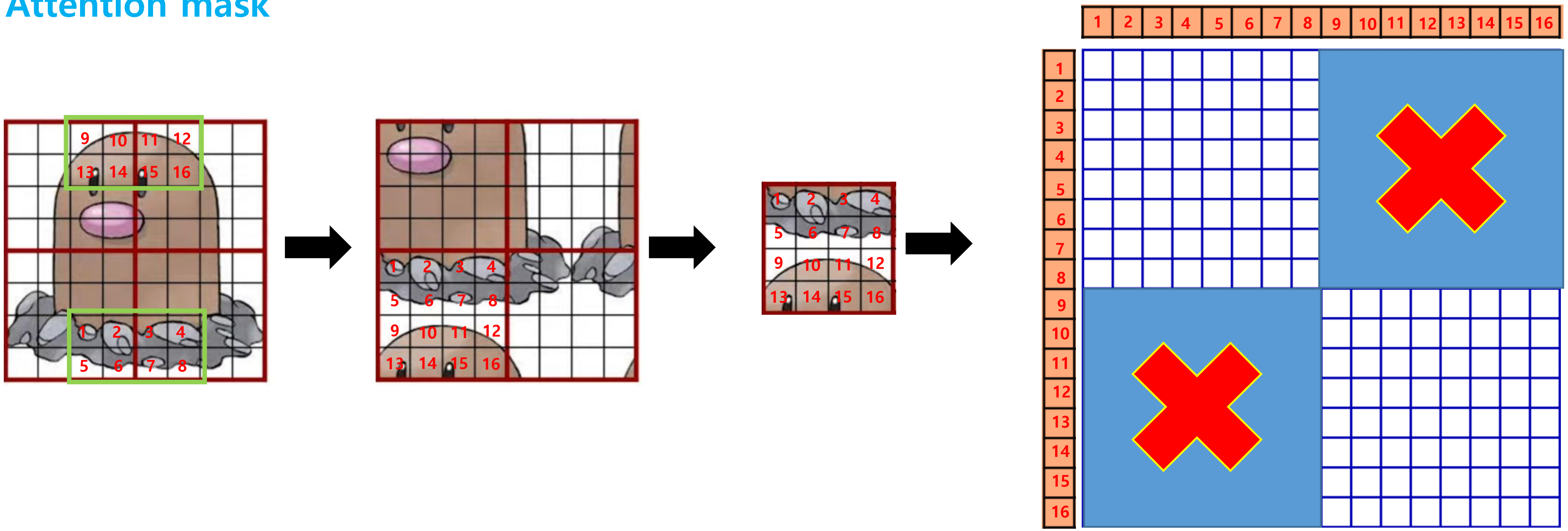
❖ SW-MSA: shifted windowing multi-head self attention module

$$\mathrm{Attention}(Q,K,V) = \mathrm{SoftMax}\left(\boxed{QK^T/\sqrt{d}+B}\right)V$$



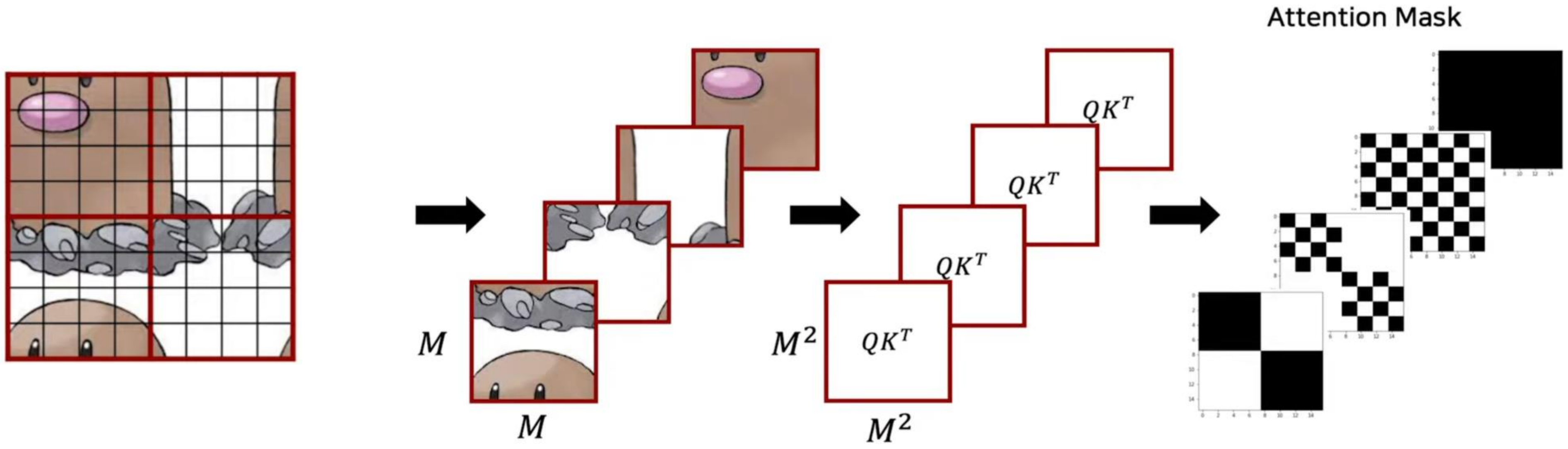
3. SwinIR: Image Restoration Using Swin Transformer

❖ Attention mask



3. SwinIR: Image Restoration Using Swin Transformer

❖ SW-MSA: shifted windowing multi-head self attention module

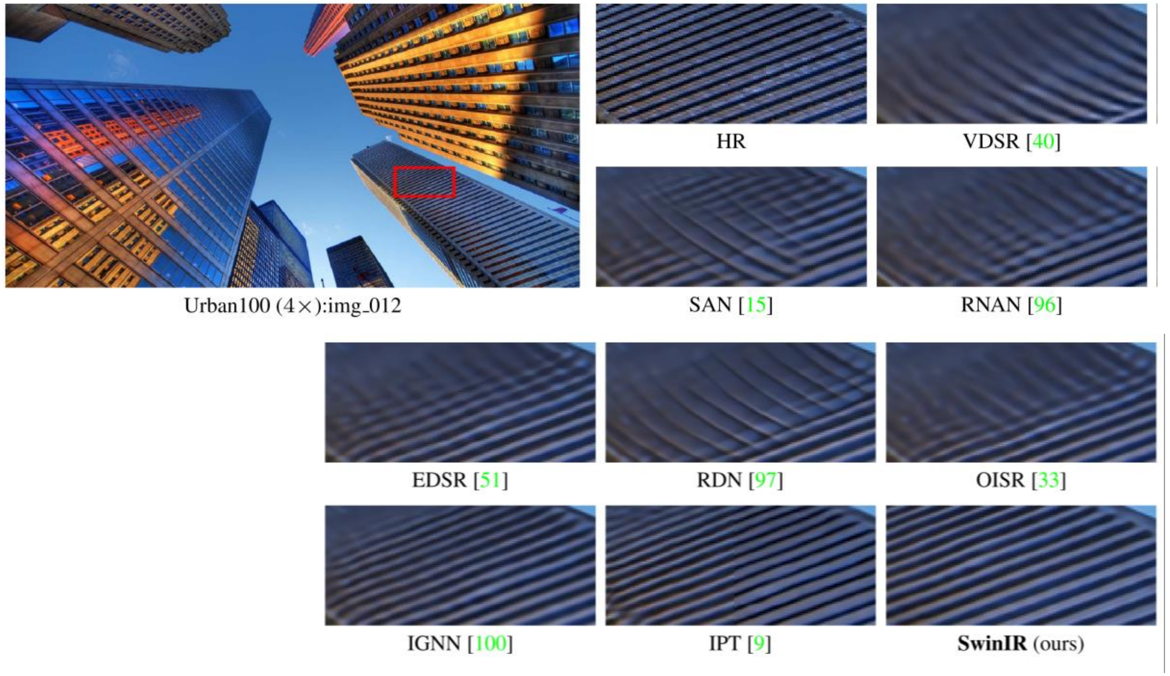


3. SwinIR: Image Restoration Using Swin Transformer

❖ Result

Table 2: Quantitative comparison (average PSNR/SSIM) with state-of-the-art methods for **classical image SR** on benchmark datasets. Best and second best performance are in red and blue colors, respectively. Results on ×8 are provided in supplementary.

Method	Scale	Training Dataset	Set5 [3] PSNR	Set5 [3] SSIM	Set14 [87] PSNR	Set14 [87] SSIM	BSD100 [58] PSNR	BSD100 [58] SSIM	Urban100 [34] PSNR	Urban100 [34] SSIM	Manga109 [60] PSNR	Manga109 [60] SSIM
RCAN [95]	×2	DIV2K	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
SAN [15]	×2	DIV2K	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
IGNN [100]	×2	DIV2K	38.24	0.9613	34.07	0.9217	32.41	0.9025	33.23	0.9383	39.35	0.9786
HAN [63]	×2	DIV2K	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
NLSA [61]	×2	DIV2K	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789
SwinIR (Ours)	×2	DIV2K	38.35	0.9620	34.14	0.9227	32.44	0.9030	33.40	0.9393	39.60	0.9792
SwinIR+ (Ours)	×2	DIV2K	38.38	0.9621	34.24	0.9233	32.47	0.9032	33.51	0.9401	39.70	0.9794
DBPN [31]	×2	DIV2K+Flickr2K	38.09	0.9600	33.85	0.9190	32.27	0.9000	32.55	0.9324	38.89	0.9775
IPT [9]	×2	ImageNet	38.37	-	34.43	-	32.48	-	33.76	-	-	-
SwinIR (Ours)	×2	DIV2K+Flickr2K	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
SwinIR+ (Ours)	×2	DIV2K+Flickr2K	38.46	0.9624	34.61	0.9260	32.55	0.9043	33.95	0.9433	40.02	0.9800
RCAN [95]	×3	DIV2K	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
SAN [15]	×3	DIV2K	34.75	0.9300	30.59	0.8476	29.33	0.8112	28.93	0.8671	34.30	0.9494
IGNN [100]	×3	DIV2K	34.72	0.9298	30.66	0.8484	29.31	0.8105	29.03	0.8696	34.39	0.9496
HAN [63]	×3	DIV2K	34.75	0.9299	30.67	0.8483	29.32	0.8110	29.10	0.8705	34.48	0.9500
NLSA [61]	×3	DIV2K	34.85	0.9306	30.70	0.8485	29.34	0.8117	29.25	0.8726	34.57	0.9508
SwinIR (Ours)	×3	DIV2K	34.89	0.9312	30.77	0.8503	29.37	0.8124	29.29	0.8744	34.74	0.9518
SwinIR+ (Ours)	×3	DIV2K	34.95	0.9316	30.83	0.8511	29.41	0.8130	29.42	0.8761	34.92	0.9526
IPT [9]	×3	ImageNet	34.81	-	30.85	-	29.38	-	29.49	-	-	-
SwinIR (Ours)	×3	DIV2K+Flickr2K	34.97	0.9318	30.93	0.8534	29.46	0.8145	29.75	0.8826	35.12	0.9537
SwinIR+ (Ours)	×3	DIV2K+Flickr2K	35.04	0.9322	31.00	0.8542	29.49	0.8150	29.90	0.8841	35.28	0.9543
RCAN [95]	×4	DIV2K	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
SAN [15]	×4	DIV2K	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
IGNN [100]	×4	DIV2K	32.57	0.8998	28.85	0.7891	27.77	0.7434	26.84	0.8090	31.28	0.9182
HAN [63]	×4	DIV2K	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
NLSA [61]	×4	DIV2K	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
SwinIR (Ours)	×4	DIV2K	32.72	0.9021	28.94	0.7914	27.83	0.7459	27.07	0.8164	31.67	0.9226
SwinIR+ (Ours)	×4	DIV2K	32.81	0.9029	29.02	0.7928	27.87	0.7466	27.21	0.8187	31.88	0.9423
DBPN [31]	×4	DIV2K+Flickr2K	32.47	0.8980	28.82	0.7860	27.72	0.7400	26.38	0.7946	30.91	0.9137
IPT [9]	×4	ImageNet	32.64	-	29.01	-	27.82	-	27.26	-	-	-
RRDB [81]	×4	DIV2K+Flickr2K	32.73	0.9011	28.99	0.7917	27.85	0.7455	27.03	0.8153	31.66	0.9196
SwinIR (Ours)	×4	DIV2K+Flickr2K	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
SwinIR+ (Ours)	×4	DIV2K+Flickr2K	32.93	0.9043	29.15	0.7958	27.95	0.7494	27.56	0.8273	32.22	0.9273



3. SwinIR: Image Restoration Using Swin Transformer

❖ Result

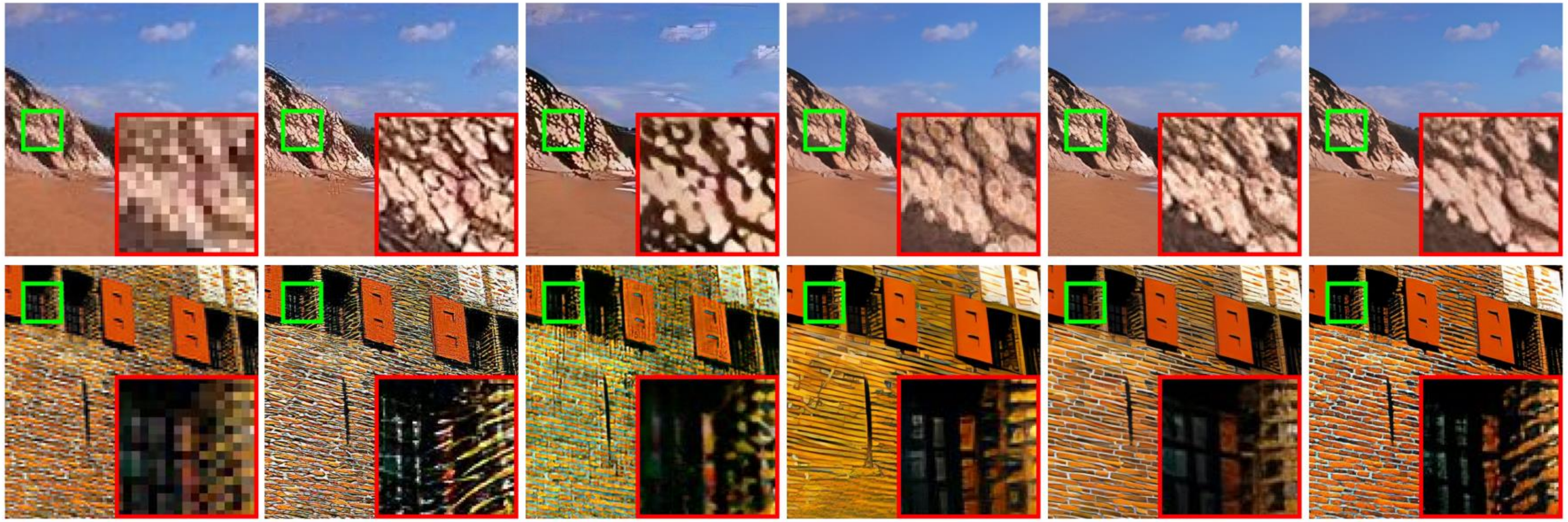
Table 3: Quantitative comparison (average PSNR/SSIM) with state-of-the-art methods for **lightweight image SR** on benchmark datasets. Best and second best performance are in red and blue colors, respectively.

Method	Scale	#Params	#Mult-Adds	Set5 [3]		Set14 [87]		BSD100 [58]		Urban100 [34]		Manga109 [60]	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CARN [2]	×2	1,592K	222.8G	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256	38.36	0.9765
FALSR-A [12]	×2	1,021K	234.7G	37.82	0.959	33.55	0.9168	32.1	0.8987	31.93	0.9256	-	-
IMDN [35]	×2	694K	158.8G	38.00	0.9605	33.63	0.9177	32.19	0.8996	32.17	0.9283	38.88	0.9774
LAPAR-A [44]	×2	548K	171.0G	38.01	0.9605	33.62	0.9183	32.19	0.8999	32.10	0.9283	38.67	0.9772
LatticeNet [57]	×2	756K	169.5G	38.15	0.9610	33.78	0.9193	32.25	0.9005	32.43	0.9302	-	-
SwinIR (Ours)	×2	878K	195.6G	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.12	0.9783
CARN [2]	×3	1,592K	118.8G	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493	33.50	0.9440
IMDN [35]	×3	703K	71.5G	34.36	0.9270	30.32	0.8417	29.09	0.8046	28.17	0.8519	33.61	0.9445
LAPAR-A [44]	×3	544K	114.0G	34.36	0.9267	30.34	0.8421	29.11	0.8054	28.15	0.8523	33.51	0.9441
LatticeNet [57]	×3	765K	76.3G	34.53	0.9281	30.39	0.8424	29.15	0.8059	28.33	0.8538	-	-
SwinIR (Ours)	×3	886K	87.2G	34.62	0.9289	30.54	0.8463	29.20	0.8082	28.66	0.8624	33.98	0.9478
CARN [2]	×4	1,592K	90.9G	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.07	0.7837	30.47	0.9084
IMDN [35]	×4	715K	40.9G	32.21	0.8948	28.58	0.7811	27.56	0.7353	26.04	0.7838	30.45	0.9075
LAPAR-A [44]	×4	659K	94.0G	32.15	0.8944	28.61	0.7818	27.61	0.7366	26.14	0.7871	30.42	0.9074
LatticeNet [57]	×4	777K	43.6G	32.30	0.8962	28.68	0.7830	27.62	0.7367	26.25	0.7873	-	-
SwinIR (Ours)	×4	897K	49.6G	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151

3. SwinIR: Image Restoration Using Swin Transformer

❖ **Result**

Quantitative comparison (average PSNR/SSIM) with state-of-the-art methods for **lightweight image SR**



LR (×4) ESRGAN [81] RealSR [37] BSRGAN [89] Real-ESRGAN [80] **SwinIR** (ours)

3. SwinIR: Image Restoration Using Swin Transformer

❖ Result

Table 4: Quantitative comparison (average PSNR/SSIM/PSNR-B) with state-of-the-art methods for **JPEG compression artifact reduction** on benchmark datasets. Best and second best performance are in red and blue colors, respectively.

Dataset	q	ARCNN [17]	DnCNN-3 [90]	QGAC [20]	RNAN [96]	RDN [98]	DRUNet [88]	**SwinIR** (ours)
Classic5 [22]	10	29.03/0.7929/28.76	29.40/0.8026/29.13	29.84/0.8370/29.43	29.96/0.8178/29.62	30.00/0.8188/-	30.16/0.8234/29.81	30.27/0.8249/29.95
	20	31.15/0.8517/30.59	31.63/0.8610/31.19	31.98/0.8850/31.37	32.11/0.8693/31.57	32.15/0.8699/-	32.39/0.8734/31.80	32.52/0.8748/31.99
	30	32.51/0.8806/31.98	32.91/0.8861/32.38	33.22/0.9070/32.42	33.38/0.8924/32.68	33.43/0.8930/-	33.59/0.8949/32.82	33.73/0.8961/33.03
	40	33.32/0.8953/32.79	33.77/0.9003/33.20	-	34.27/0.9061/33.4	34.27/0.9061/-	34.41/0.9075/33.51	34.52/0.9082/33.66
LIVE1 [67]	10	28.96/0.8076/28.77	29.19/0.8123/28.90	29.53/0.8400/29.15	29.63/0.8239/29.25	29.67/0.8247/-	29.79/0.8278/29.48	29.86/0.8287/29.50
	20	31.29/0.8733/30.79	31.59/0.8802/31.07	31.86/0.9010/31.27	32.03/0.8877/31.44	32.07/0.8882/-	32.17/0.8899/31.69	32.25/0.8909/31.70
	30	32.67/0.9043/32.22	32.98/0.9090/32.34	33.23/0.9250/32.50	33.45/0.9149/32.71	33.51/0.9153/-	33.59/0.9166/32.99	33.69/0.9174/33.01
	40	33.63/0.9198/33.14	33.96/0.9247/33.28	-	34.47/0.9299/33.66	34.51/0.9302/-	34.58/0.9312/33.93	34.67/0.9317/33.88

3. SwinIR: Image Restoration Using Swin Transformer

❖ Result

Table 5: Quantitative comparison (average PSNR) with state-of-the-art methods for **grayscale image denoising** on benchmark datasets. Best and second best performance are in red and blue colors, respectively.

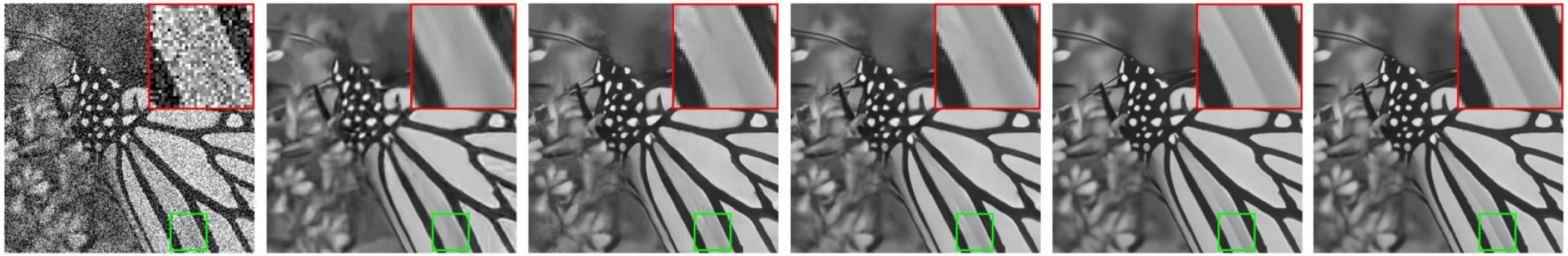
Dataset	σ	BM3D [14]	WNNM [29]	DnCNN [90]	IRCNN [91]	FFDNet [92]	N3Net [65]	NLRN [52]	FOCNet [38]	RNAN [96]	MWCNN [54]	DRUNet [88]	**SwinIR** (ours)
Set12 [90]	15	32.37	32.70	32.86	32.76	32.75	-	33.16	33.07	-	33.15	33.25	33.36
	25	29.97	30.28	30.44	30.37	30.43	30.55	30.80	30.73	-	30.79	30.94	31.01
	50	26.72	27.05	27.18	27.12	27.32	27.43	27.64	27.68	27.70	27.74	27.90	27.91
BSD68 [59]	15	31.08	31.37	31.73	31.63	31.63	-	31.88	31.83	-	31.86	31.91	31.97
	25	28.57	28.83	29.23	29.15	29.19	29.30	29.41	29.38	-	29.41	29.48	29.50
	50	25.60	25.87	26.23	26.19	26.29	26.39	26.47	26.50	26.48	26.53	26.59	26.58
Urban100 [34]	15	32.35	32.97	32.64	32.46	32.40	-	33.45	33.15	-	33.17	33.44	33.70
	25	29.70	30.39	29.95	29.80	29.90	30.19	30.94	30.64	-	30.66	31.11	31.30
	50	25.95	26.83	26.26	26.22	26.50	26.82	27.49	27.40	27.65	27.42	27.96	27.98

Table 6: Quantitative comparison (average PSNR) with state-of-the-art methods for **color image denoising** on benchmark datasets. Best and second best performance are in red and blue colors, respectively.

Dataset	σ	BM3D [14]	DnCNN [90]	IRCNN [91]	FFDNet [92]	DSNet [64]	RPCNN [85]	BRDNet [71]	RNAN [96]	RDN [98]	IPT [9]	DRUNet [88]	**SwinIR** (ours)
CBSD68 [59]	15	33.52	33.90	33.86	33.87	33.91	-	34.10	-	-	-	34.30	34.42
	25	30.71	31.24	31.16	31.21	31.28	31.24	31.43	-	-	-	31.69	31.78
	50	27.38	27.95	27.86	27.96	28.05	28.06	28.16	28.27	28.31	28.39	28.51	28.56
Kodak24 [23]	15	34.28	34.60	34.69	34.63	34.63	-	34.88	-	-	-	35.31	35.34
	25	32.15	32.14	32.18	32.13	32.16	32.34	32.41	-	-	-	32.89	32.89
	50	28.46	28.95	28.93	28.98	29.05	29.25	29.22	29.58	29.66	29.64	29.86	29.79
McMaster [94]	15	34.06	33.45	34.58	34.66	34.67	-	35.08	-	-	-	35.40	35.61
	25	31.66	31.52	32.18	32.35	32.40	32.33	32.75	-	-	-	33.14	33.20
	50	28.51	28.62	28.91	29.18	29.28	29.33	29.52	29.72	-	29.98	30.08	30.22
Urban100 [34]	15	33.93	32.98	33.78	33.83	-	-	34.42	-	-	-	34.81	35.13
	25	31.36	30.81	31.20	31.40	-	31.81	31.99	-	-	-	32.60	32.90
	50	27.93	27.59	27.70	28.05	-	28.62	28.56	29.08	29.38	29.71	29.61	29.82

3. SwinIR: Image Restoration Using Swin Transformer

❖ **Result**



Noisy | BM3D [14] | DnCNN [90] | FFDNet [92] | DRUNet [88] | **SwinIR** (ours)

Figure 6: Visual comparison of **grayscale image denoising** (noise level 50) methods on image "*Monarch*" from Set12 [90]. Compared images are derived from [88].

Noisy | DnCNN [90] | FFDNet [92] | IPT [9] | DRUNet [88] | **SwinIR** (ours)

Figure 7: Visual comparison of **color image denoising** (noise level 50) methods on image "*163085*" from CBSD68 [59]. Compared images are derived from [88].

4. Conclusion

- Main contribution:
  - Apply Swin transformer structure for image restoration task.
  - Get highest performance:
    - ✓ classic image SR
    - ✓ lightweight image SR
    - ✓ real-world image SR
    - ✓ grayscale image denoising
    - ✓ color image denoising
    - ✓ JPEG compression artifact reduction