Transformer for Single Image Super-Resolution

Zhisheng Lu[1†], Juncheng Li[2†], Hong Liu[1*], Chaoyan Huang[3], Linlin Zhang[1], Tieyong Zeng[2]

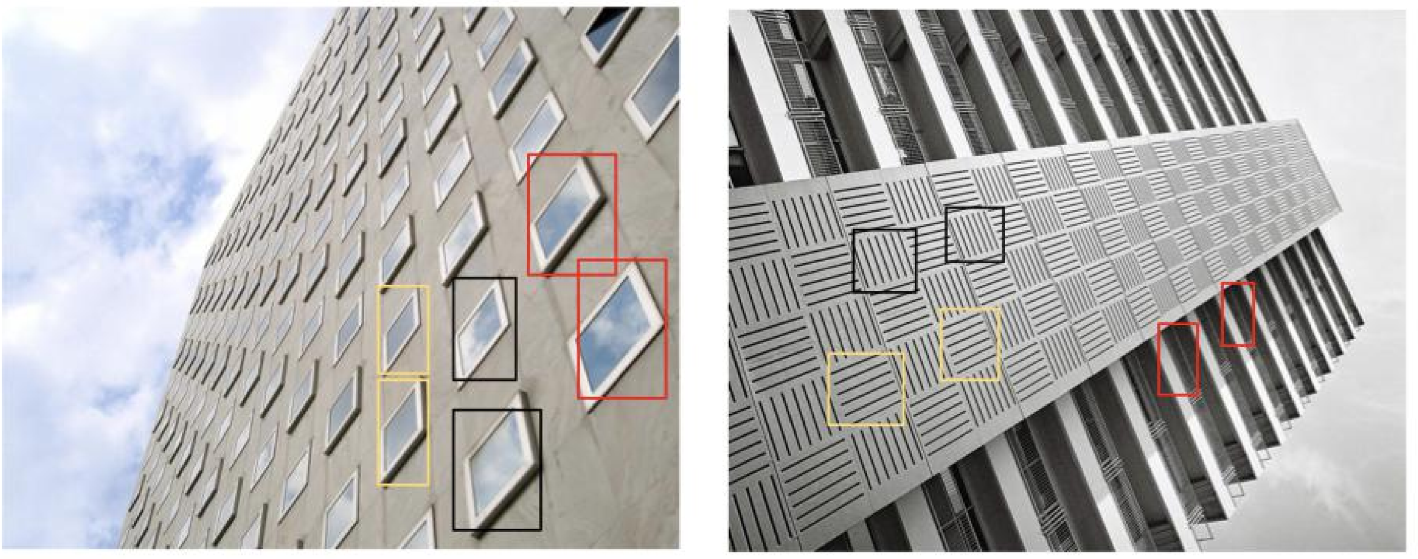
[1]Peking University Shenzhen Graduate School [2]The Chinese University of Hong Kong

[3]Nanjing University of Posts and Telecommunications

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Motivation

The inner areas of the boxes with the same color are similar to each other. Therefore, these similar image patches can be used as reference images for each other, so that the texture details of the certain patch can be restored with reference patches. Inspired by this, **we aim to introduce the Transformer into the SISR task since it has a strong feature expression ability to model such a long-term dependency in the image.**



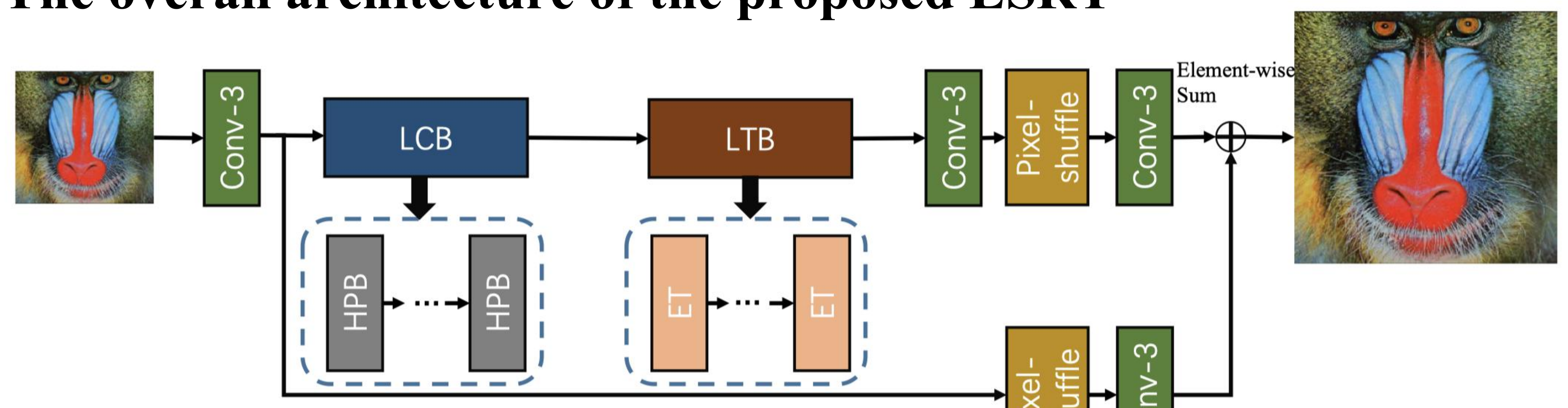
Recently, some Vision-Transformer have been proposed for computer vision tasks. However, these methods often occupy heavy GPU memory, which greatly limits their flexibility and application scenarios. Moreover, these methods cannot be directly transferred to SISR since the image restoration task often take a larger resolution image as input, which will take up huge memory. **Therefore, we aim to explore a more efficient Transformer.**

Contributions

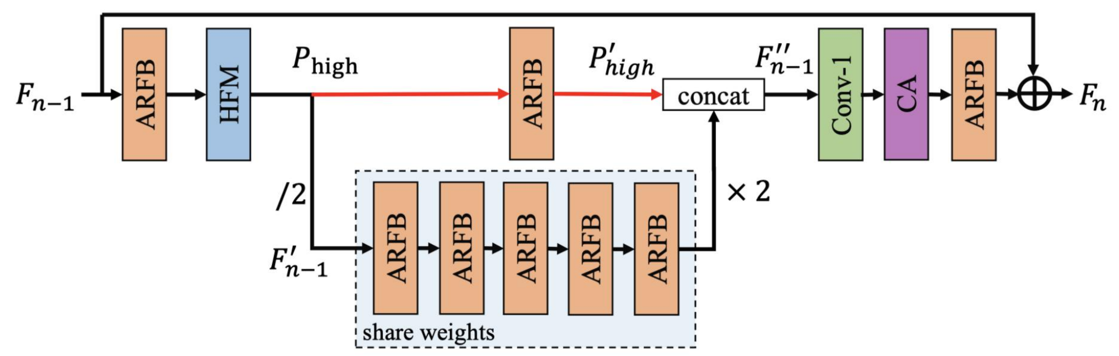
- We propose a **Lightweight CNN Backbone (LCB)**, which use High Preserving Blocks (HPBs) to dynamically adjust the size of the feature map to extract deep features with a low computational cost.
- We propose a **Lightweight Transformer Backbone (LTB)** to capture long-term dependencies between similar patches in an image with the help of the specially designed Efficient Transformer (ET) and Efficient Multi-Head Attention (EMHA) mechanism.
- A novel model called **Efficient SR Transformer (ESRT)** is proposed to effectively enhance the feature expression ability and the long-term dependence of similar patches in an image, so as to achieve better performance with low computational cost.

Method

The overall architecture of the proposed ESRT

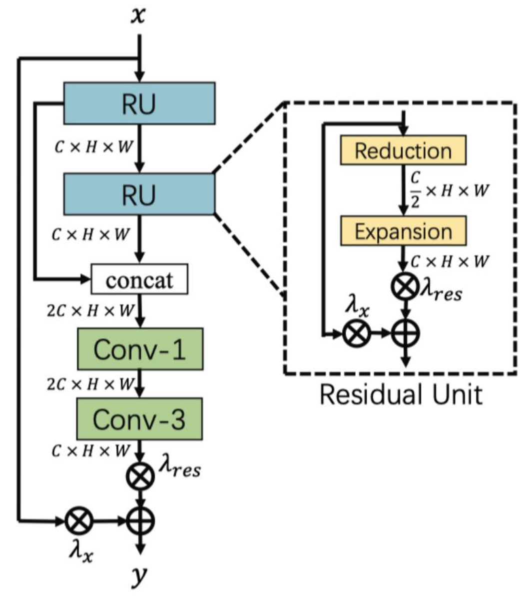


The architecture of HPB



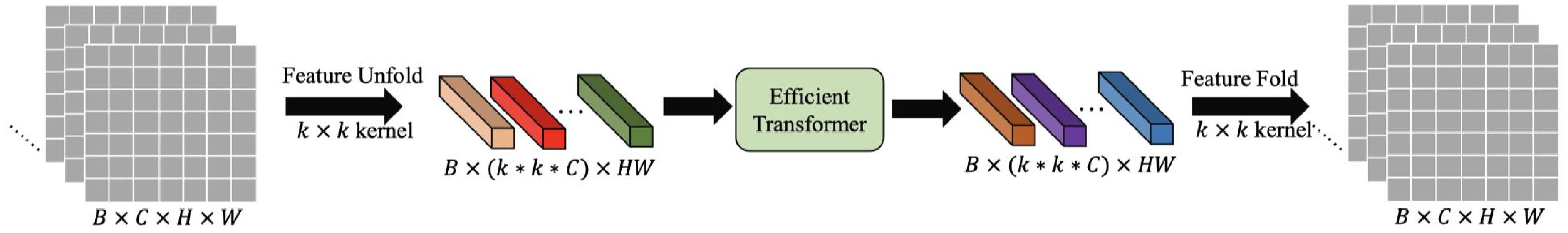
$$F''_{n-1} = [f_a(P_{high}), \uparrow f_a^{\circlearrowleft 5}(\downarrow F'_{n-1})]$$

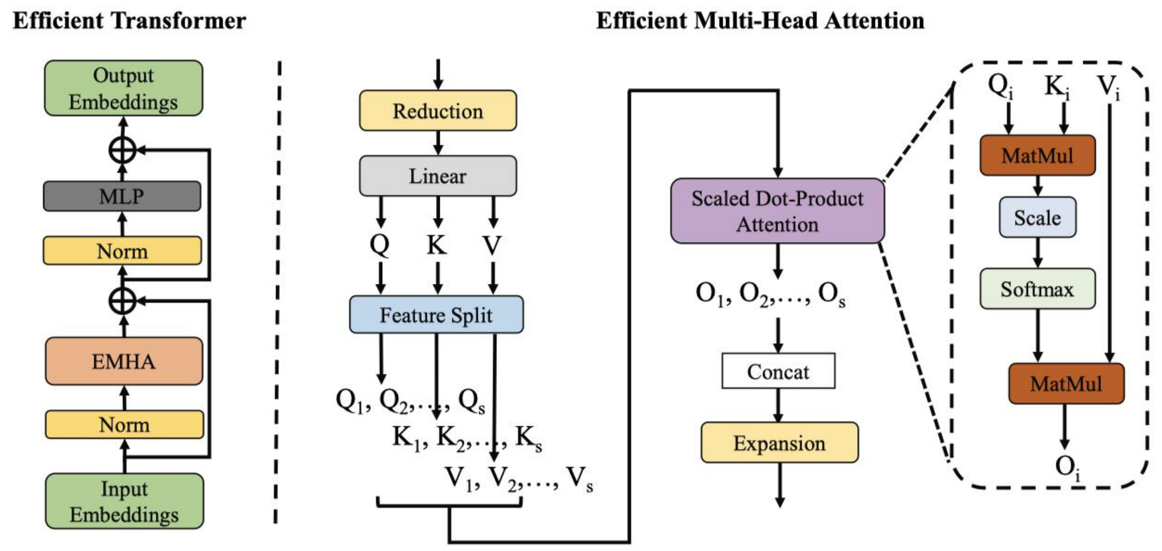
The architecture of ARFB



$$y_{ru} = \lambda_{res} \cdot f_{ex}(f_{re}(x_{ru})) + \lambda_x \cdot x$$

Pre- and Post-processing for ET





Efficient Transformer (ET) & Efficient Multi-Head Attention (EMHA)

$$E_{m1} = EMHA(Norm(E_i)) + E_i,$$
$$E_o = MLP(Norm(E_{m1})) + E_{m1},$$

where E_o is the output of the ET, $EMHA(\cdot)$ and $MLP(\cdot)$ represent the EMHA and MLP operations, respectively.

Results

Method	Scale	Params	Set5 PSNR / SSIM	Set14 PSNR / SSIM	BSD100 PSNR / SSIM	Urban100 PSNR / SSIM	Manga109 PSNR / SSIM
VDSR [18]	×3	666K	33.66 / 0.9213	29.77 / 0.8314	28.82 / 0.7976	27.14 / 0.8279	32.01 / 0.9340
MemNet [34]		678K	34.09 / 0.9248	30.00 / 0.8350	28.96 / 0.8001	27.56 / 0.8376	32.51 / 0.9369
EDSR-baseline [26]		1,555K	34.37 / 0.9270	30.28 / 0.8417	29.09 / 0.8052	28.15 / 0.8527	33.45 / 0.9439
SRMDNF [43]		1,528K	34.12 / 0.9254	30.04 / 0.8382	28.97 / 0.8025	27.57 / 0.8398	33.00 / 0.9403
CARN [2]		1,592K	34.29 / 0.9255	30.29 / 0.8407	29.06 / 0.8034	28.06 / 0.8493	33.50 / 0.9440
IMDN [16]		703K	34.36 / 0.9270	30.32 / 0.8417	29.09 / 0.8046	28.17 / 0.8519	33.61 / 0.9445
RFDN-L [27]		633K	34.47 / 0.9280	30.35 / 0.8421	29.11 / 0.8053	28.32 / 0.8547	33.78 / 0.9458
MAFFSRN [31]		807K	34.45 / 0.9277	30.40 / 0.8432	29.13 / 0.8061	28.26 / 0.8552	- / -
LatticeNet [29]		765K	**34.53 / 0.9281**	30.39 / 0.8424	**29.15** / 0.8059	28.33 / 0.8538	- / -
ESRT(ours)		770K	34.42 / 0.9268	**30.43 / 0.8433**	**29.15 / 0.8063**	**28.46 / 0.8574**	**33.95 / 0.9455**
VDSR [18]	×4	666K	31.35 / 0.8838	28.01 / 0.7674	27.29 / 0.7251	25.18 / 0.7524	28.83 / 0.8870
MemNet [34]		678K	31.74 / 0.8893	28.26 / 0.7723	27.40 / 0.7281	25.50 / 0.7630	29.42 / 0.8942
EDSR-baseline [26]		1,518K	32.09 / 0.8938	28.58 / 0.7813	27.57 / 0.7357	26.04 / 0.7849	30.35 / 0.9067
SRMDNF [43]		1,552K	31.96 / 0.8925	28.35 / 0.7787	27.49 / 0.7337	25.68 / 0.7731	30.09 / 0.9024
CARN [2]		1,592K	32.13 / 0.8937	28.60 / 0.7806	27.58 / 0.7349	26.07 / 0.7837	30.47 / 0.9084
IMDN [16]		715K	32.21 / 0.8948	28.58 / 0.7811	27.56 / 0.7353	26.04 / 0.7838	30.45 / 0.9075
RFDN-L [27]		643K	32.28 / 0.8957	28.61 / 0.7818	27.58 / 0.7363	26.20 / 0.7883	30.61 / 0.9096
MAFFSRN [31]		830K	32.20 / 0.8953	26.62 / 0.7822	27.59 / 0.7370	26.16 / 0.7887	- / -
LatticeNet [29]		777K	**32.30 / 0.8962**	28.68 / 0.7830	27.62 / 0.7367	26.25 / 0.7873	- / -
ESRT (ours)		751K	32.19 / 0.8947	**28.69 / 0.7833**	**27.69 / 0.7379**	**26.39 / 0.7962**	**30.75 / 0.9100**

Method	Layers	RL	Param.	FLOPs (x4)	Running time
VDSR [18]	20	Yes	0.67M	612.6G	0.00597s
LapSRN [20]	27	Yes	0.25M	149.4G	0.00330s
DRRN [33]	52	No	0.30M	6796.9G	0.08387s
CARN [2]	34	Yes	1.6M	90.9G	0.00278s
IMDN [16]	34	Yes	0.7M	40.9G	0.00258s
ESRT	163	Yes	0.68M	67.7G	0.01085s

Scale	Model	Param	Set5	Set14	Urban100
×3	RCAN [44]	16M	**34.74dB**	**30.65dB**	29.09dB
	RCAN/2+ET	8.7M	34.69dB	30.63dB	**29.16dB**
×4	RCAN [44]	16M	**32.63dB**	28.87dB	26.82dB
	RCAN/2+ET	8.7M	32.60dB	**28.90dB**	**26.87dB**

Visual Comparison