Transformer for Burst Image Super-Resolution

Zhilu Zhang[1], Rongjian Xu, Shuohao Zhang, Xiaohe Wu, Wangmeng Zuo
Harbin Institute of Technology
Harbin, China
[1] cszlzhang@outlook.com

1. Summary

We propose a transformer model for burst image super-resolution named TBSR, which utilizes the transformer module [7] as the reconstruction module in EBSR [4]. For track 1, we train TBSR with ℓ_1 loss end-to-end. For track 2, we take advantage of sliced Wasserstein (SW) loss introduced in SelfDZSR [9] to obtain the results with fine high-frequency details. Simultaneously, we handle the misalignment issue according to the alignment strategy proposed in RAW-to-sRGB [8].

2. Method Description

2.1. TBSR Model

We propose a transformer model for burst image super-resolution named TBSR, as shown in Fig. 1. TBSR borrows the alignment and fusion modules from EBSR [4], and takes the transformer module as the reconstruction module. The reconstruction module includes m transformer groups, and each transformer group includes n transformer blocks. We take the basic block proposed in Restormer [7] as our transformer block. The block implicitly captures long-range pixel interactions by applying self-attention across channels. Thus, the computational complexity of the blocks grows linearly with the spatial resolution, while that of the transformer-based methods that apply self-attention across the spatial dimension grows quadratically. The efficient blocks make TBSR applicable to large images. During the experiment, we set $m = n = 8$. The total number of parameters for TBSR model is $\sim$ 24 M.

2.2. Details for Track 1

During training, we utilize ℓ_1 loss to train TBSR end to end. The training burst data is synthesized from sRGB images in the Zurich RAW to RGB [2] training set. During testing, we use a self-ensemble strategy [3] for better quantitative performance.

2.3. Details for Track 2

We take the model pre-trained in track 1 as the initialization model for track 2. Then the model is trained with a combination of ℓ_1 loss, VGG-based perceptual loss [5] and sliced Wasserstein (SW) loss on BurstSR [1] training set. SW loss has been successfully applied in image super-resolution [9], and shows competitive results with adversarial loss in recovering details. The algorithm of SW loss can be seen in Alg. 1. First, the 2-dimensional VGG [5] features are mapped to 1-dimensional through random linear projection. Then the Wasserstein distance between the 1-dimensional representations of output and target is calculated as SW loss.

In order to handle the misalignment issue between the output and original ground-truth (GT), we take the strategy proposed in RAW-to-sRGB [8] for aligning GT with output. Specifically, we take the gamma-corrected output and GT into the optical flow estimation network [6]. Then the acquired optical flow can be used to warp GT towards aligning with output. Instead of the original GT, the warped GT can be used as the learning target of TBSR model. In another word, the loss is calculated between output and the warped GT.

References

[1] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *CVPR*, pages 9209–9218, 2021. 1

[2] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *CVPR Workshops*, pages 536–537, 2020. 1

[3] Jiaming Liu, Chi-Hao Wu, Yuzhi Wang, Qin Xu, Yuqian Zhou, Haibin Huang, Chuan Wang, Shaofan Cai, Yifan Ding, Haoqiang Fan, et al. Learning raw image denoising with bayer pattern unification and bayer preserving augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[4] Ziwei Luo, Lei Yu, Xuan Mo, Youwei Li, Lanpeng Jia, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. Ebsr: Feature enhanced burst super-resolution with deformable alignment. In *CVPR Workshops*, pages 471–478, 2021. 1

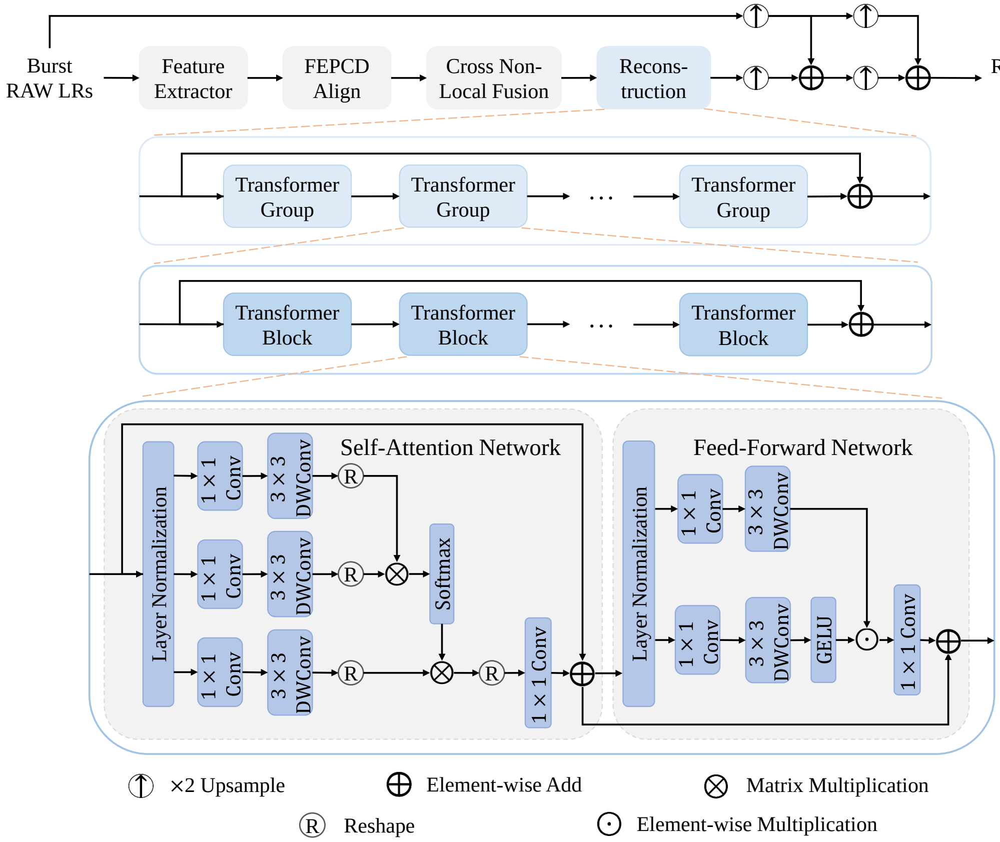

Figure 1. Overview of the network architecture for TBSR.

Algorithm 1 Pseudocode of SW loss

```
Require: $\mathbf{U} \in \mathbb{R}^{C \times H \times W}$: VGG features of output image; $\mathbf{V} \in \mathbb{R}^{C \times H \times W}$: VGG features of target image; $\mathbf{M} \in \mathbb{R}^{C' \times C}$: random projection matrix;
Ensure: $\mathcal{L}_{\text{SW}}(\mathbf{U}, \mathbf{V})$: the value of SW loss;
  flatten features $\mathbf{U}$ and $\mathbf{V}$ to $\mathbf{U_f} (\in \mathbb{R}^{C \times HW})$ and $\mathbf{V_f} (\in \mathbb{R}^{C \times HW})$, respectively;
  project the features onto $C'$ directions: $\mathbf{U_p} = \mathbf{M}\mathbf{U_f}$, $\mathbf{V_p} = \mathbf{M}\mathbf{V_f}$;
  sort projections for each direction: $\mathbf{U_s} = \mathbf{Sort}(\mathbf{U_p}, \text{dim=1})$, $\mathbf{V_s} = \mathbf{Sort}(\mathbf{V_p}, \text{dim=1})$;
  $\mathcal{L}_{\text{SW}}(\mathbf{U}, \mathbf{V}) = \|\mathbf{U_s} - \mathbf{V_s}\|_1$
```

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

[6] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, pages 8934–8943, 2018. 1

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

[8] Zhilu Zhang, Haolin Wang, Ming Liu, Ruohao Wang, Jiawei Zhang, and Wangmeng Zuo. Learning raw-to-srgb mappings with inaccurately aligned supervision. In *ICCV*, pages 4348–4358, 2021. 1

[9] Zhilu Zhang, Ruohao Wang, Hongzhi Zhang, Yunjin Chen, and Wangmeng Zuo. Self-supervised learning for real-world super-resolution from dual zoomed observations. *arXiv preprint arXiv:2203.01325*, 2022. 1