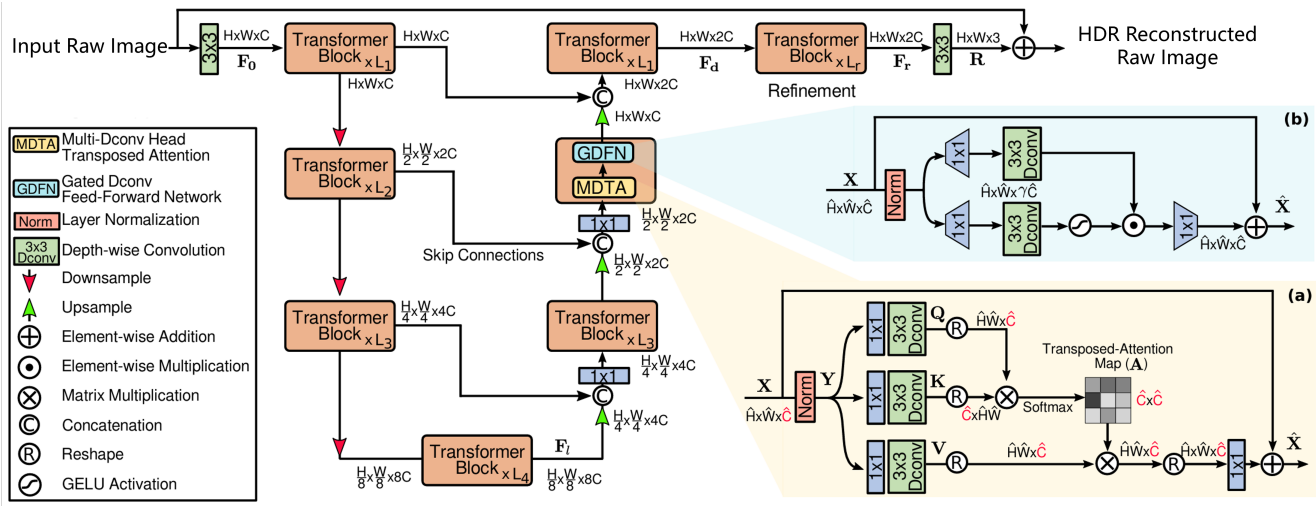

Figure 1. Our Restormer Framework for HDR Reconstruction.

1. Team details

- **Challenge name:**
  HDR Reconstruction from a Single Raw Image
- **User name:**
  USTCX
- **Team leader information:**
  Name: Senyan Xu
  Email: syxu@mail.ustc.edu.cn
- **Team member information:**
  Member 1: Yunkang Zhang, ykzhang@mail.ustc.edu.cn
  Member 2: Siyuan Jiang, syjiang@mail.ustc.edu.cn
- **Team institute:**
  University of Science and Technology of China
- **Code Link:**
  Google Drive

2. Network Architecture

Overall Pipeline Inspired by Restormer[2], our network architecture is shown in Fig.1. It overcomes the limitations of traditional Convolutional Neural Networks (CNNs) by utilizing Transformers' ability to capture long-range pixel interactions, which is crucial for High Dynamic Range (HDR) image reconstruction. We introduce the two modules of the Transformer block: **(a)** multi-Dconv head transposed attention (MDTA) and **(b)** gated-Dconv feed-forward network (GDFN).

Multi-Dconv Head Transposed Attention This module replaces the standard multi-head self-attention mechanism. It operates across feature dimensions instead of spatial dimensions, reducing complexity. It uses 1×1 convolutions for pixel-wise cross-channel context aggregation and 3×3 depth-wise convolutions for channel-wise spatial context encoding.

Gated-Dconv Feed-Forward Network This module includes a gating mechanism and depth-wise convolutions to control information flow and encode local image structures. The gating mechanism is an element-wise product of two linear transformation paths, one activated with GELU nonlinearity.

Loss Functions In our work, we use the Charbonnier loss[1] to optimize our network. This loss function is particularly effective for handling outliers and robust to noise. Its formulation is as follows:

$$\mathcal{L}_{\text{content}} = \sqrt{\left\| \hat{I} - I \right\|_2 + \epsilon^2} \tag{1}$$

where $\hat{I}$ is the predicted HDR raw image, I is the ground truth, and ϵ is set to 0.0001 as default.

In addition to the content loss, we leverage frequency domain information to introduce auxiliary loss to our network, which is defined as follows:

$$\mathcal{L}_{\text{frequency}} = \left\| \mathcal{F}\left(\hat{I}\right) - \mathcal{F}\left(I\right) \right\|_1 \tag{2}$$

where $\mathcal{F}(\cdot)$ indicates the Fast Fourier Transform (FFT). Finally, the total loss could be defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{content}} + \lambda \mathcal{L}_{\text{frequency}} \tag{3}$$

where λ denotes the balanced weight, and we empirically set λ to 0.5 as default.

3. Implementation details

We utilized PyTorch 1.8 within an NVIDIA 3090 GPU environment, equipped with 24GB of memory, to train our model on official datasets with a batch size of 4. The input images were standardized to an 80×80 resolution. The training spanned approximately 23 hours, with a learning rate that started at 3×10^{-4}, reduced to 1×10^{-7} over 75,000 iterations using a Cosine Annealing schedule. This was followed by a second phase with a learning rate of 6×10^{-5}, also reduced to 1×10^{-7} over an additional 60,000 iterations. Notably, no special efficiency optimization strategies were applied during this process.

References

[1] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1954–1963, 2019. 1

[2] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 1