Overview

The supplementary material is organized as follows. In Sec. 1, we provide more discussions about AIM. Sec. 2 provides more versions of DAT. In Sec. 3, we provide further analyses, investigating the advantages of aggregating channel and spatial information. Sec. 4, we conduct experiments on the recent state-of-the-art method ELAN for fair comparisons with our method. Sec. 5 and Sec. 6 provide more quantitative and visual comparisons. Finally, in Sec. 7, we discuss the limitation and future work of our method.

1. More Discussions about AIM

Firstly, we further explain the motivation of the AIM. Then, we describe the design considerations of spatial-interaction (S-I) and channel-interaction (C-I) of AIM.

1.1. Motivation

In general, AIM is proposed to enhance the fusion of depth-wise convolution (DW-Conv) and self-attention (SA) branches, and to aggregate spatial and channel information in a single SA module. Firstly, considering the misalignment between local (DW-Conv) and global (SA) features [18, 24], the two branches cannot be fused effectively. Secondly, SA applies a weight-sharing mechanism, limiting its feature learning in shared dimensions [8, 3]. As shown in Fig. 1, SW-SA applies the same spatial attention map (dynamic weights, from the dot-product between query and key) to each channel, namely, sharing weights on channel dimensions. Similarly, CW-SA shares weights on spatial dimensions. Since weight sharing, a single SA module cannot effectively aggregate both spatial and channel dimensions. Finally, to alleviate above issues, we adaptively adjust features through dynamic weights. Meanwhile, considering the parallel structure, the dynamic weights are generated from interactions between two branches.

1.2. Design Considerations

To realize the above purpose, we propose AIM, which consists of spatial-interaction (S-I) and channel-interaction (C-I). Specifically, we first introduce interaction from DW-Conv to SA. For SW-SA, as the above analysis, we generate a channel attention map (C-Map) to adjust channel dimension. For CW-SA, we generate a spatial attention map (S-Map). For C-I, we follow the design of the SE layer [10]. For S-I, considering that the convolution branch already extracts spatial information, we only utilize 1×1 convolution to compress the channel dimension without explicitly modeling spatial information in S-I.

Furthermore, from the perspective of duality, we also introduce interaction from SA to DW-Conv. Since SW-SA extracts strong spatial information, we utilize S-I to transfer it to the corresponding DW-Conv branch. Similarly, we apply C-I in CW-SA to strengthen the channel expression of convolution. Synthesizing the above designs, we propose the AIM to enhance branch fusion and achieve feature aggregation. The ablation study in Table 1(b, c) in the main paper demonstrates the effectiveness of our AIM.

2. More DAT Variants

In this section, we provide more versions of DAT to demonstrate the effectiveness of our proposed method. Firstly, we provide DAT-2 with fewer Params (model parameters) and similar FLOPs (computational complexity) to SwinIR [13]. Secondly, we provide a light-weight model, DAT-light, for light-weight image SR. Finally, we provide DAT-3 with the same window size (8×8) as SwinIR.
2.2. Light-weight Model: DAT-light

Implementation details. We provide a light-weight model, DAT-light, for light-weight image SR. DAT-light only has 1 RG and 9 DATB pairs (9 DCTB and 9 DSTB). The channel number, attention head number, and channel expansion factor are set as 60, 6, and 2, respectively. The window size for DSTB is set as 8×32.

Training Settings. We train DAT-light on DIV2K [25] and Flickr2K [14], and test it on Set5 [1], Set14 [26], B100 [19], Urban100 [11], and Manga109 [20]. The training settings are consistent with DAT-S and DAT.

Quantitative Results. We compare our DAT-light with recent state-of-the-art lightweight methods: SwinIR [13] and ELAN [29], in Table 2. FLOPs are calculated when the output size is set as 3×1280×720 for three scale factors. Please note that we re-test the Params and FLOPs of SwinIR and ELAN with their official codes. Our DAT-light achieves better performance with fewer Params and FLOPs for all scale factors, compared with SwinIR and ELAN.

2.3. 8×8 Window Size Model: DAT-3

Implementation details. We provide DAT-3 with a window size of 8×8, the same as SwinIR [13]. Specifically, we set the RG number, DATB pair number, channel number, attention head number, and channel expansion factor as 6, 3, 180, 6, and 2. The window size for DSTB is set as 8×32.

Training Settings. We train DAT-3 on DIV2K [25] and Flickr2K [14]. The training settings are consistent with DAT-S and DAT. The main paper has more details.
Quantitative Results. To demonstrate the effectiveness of our method, we build DAT-3 with the same window size as SwinIR [13]. Due to time issues, we only train DAT-3 on image SR (×2), and compare it with SwinIR. The results are shown in Table 3. FLOPs are calculated when the output size is set as 3×128×128. As we can see, with fewer Params and FLOPs, our DAT-3 outperforms SwinIR, except for the SSIM value on Urban100. Especially, our DAT-3 obtains a 0.2 dB PSNR gains on Manga109. All these results demonstrate the effectiveness of our methods.

3. Further Analyses

In this section, we provide more quantitative and visual analyses. Firstly, we apply the LAM [7] to visualize the range of information utilization. Then, we introduce several perceptual similarity metrics to evaluate our method. Finally, we plot the convergence curves during training for SwinIR and our models.

3.1. LAM Analyses

We apply the LAM [7] to analyze the performance of our DAT. LAM is a diagnostic tool designed for image super-resolution (SR). It can show the pixels that contribute most to the reconstruction of the selected region. The corresponding pixels are marked as red in related images. More marked pixels mean the model can utilize more information, thus resulting in better performance. Fig. 2 shows the LAM comparisons between CAT-A [4] and DAT. Comparing the second and fourth columns, we can find that the number of red marker points of our DAT is more than CAT-A. It indicates that our DAT has larger receptive fields and utilizes more global information to restore images. This is because our method has a stronger representation ability through aggregating spatial and channel features.

3.2. Perceptual Similarity Analyses

In the main paper, we quantitatively compare our method with current methods using metrics: PSNR/SSIM. However, the literature [2] reveals that the superiority of PSNR values does not always accord with better visual quality. Moreover, we also found that compared to CAT-A, our DAT has a lower PSNR value on Urban100 (×4), but has better visual results. To further evaluate our methods, we introduce two metrics: LPIPS [28] and DISTS [6]. Compared with PSNR, LPIPS and DISTS align more with human perception. The lower the value of LPIPS and DISTS, the more similar the two images. We compare our DAT with SwinIR and CAT-A on Urban100 and Manga109 with a scale factor of ×4. The results are listed in Table 4. Our DAT achieves the best performance (lowest value) on both datasets. This result demonstrates the superiority of our method. It is also consistent with the visual comparison in Figs. 4 and 5.

3.3. Convergence Analyses

The convergence curves for SwinIR, DAT-S, and DAT are shown in Fig. 3. PSNR values are tested on Set5 [1], Set14 [26], B100 [19], Urban100 [11], and Manga109 [20] (×2). For fair comparisons, we train SwinIR under the official code with the same training settings as our methods. The total training iterations are 5×10^5. We sample every 5×10^3 iterations on Set5, while every 5×10^4 iterations on other datasets. We can observe that both DAT-S and DAT converge faster than SwinIR on all datasets. These results are in accord with quantitative comparisons in Table 6, further demonstrating the effectiveness of our method.

4. Recent Method: ELAN

Recently, Zhang et al. [29] proposed a new image SR Transformer model, named efficient long-range attention network (ELAN). ELAN improves the efficiency of Transformer in SR tasks and outperforms SwinIR [13] in some cases. However, ELAN is trained on DIV2K [25], while our DAT-S and DAT are on DIV2K and Flickr2K [14]. Meanwhile, the model size and computational complexity of ELAN are smaller than our model. For fair comparisons, we increase the efficient long-range attention block (ELAB) number in ELAN, thus constructing a new variant of ELAN, denoted as ELAN-2. Then we re-train ELAN and ELAN-2 on DIV2K and Flickr2K.

4.1. Experimental Settings

Implementation Details. For ELAN, we adopt the settings in the official paper [29]. Specifically, the ELAB number is 36, and the channel number is 180. The GMSA module contains three window sizes: 4×4, 8×8, and 16×16. For ELAN-2, we only increase the ELAB number from 36 to 48, while other settings are the same as ELAN.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scale</th>
<th>Urban100 LPIPS ↓</th>
<th>DISTS ↓</th>
<th>Manga109 LPIPS ↓</th>
<th>DISTS ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>SwinIR [13]</td>
<td>×4</td>
<td>0.1840</td>
<td>0.1533</td>
<td>0.0926</td>
<td>0.0766</td>
</tr>
<tr>
<td>CAT-A [4]</td>
<td>×4</td>
<td>0.1801</td>
<td>0.1502</td>
<td>0.0906</td>
<td>0.0753</td>
</tr>
<tr>
<td>DAT (ours)</td>
<td>×4</td>
<td>0.1765</td>
<td>0.1487</td>
<td>0.0896</td>
<td>0.0735</td>
</tr>
</tbody>
</table>

Table 4: Perceptual Similarity (LPIPS/DISTS) Comparison.
We compare re-trained ELAN and ELAN-2 on DIV2K [25] and Flickr2K [14] with the official code. We train two models with patch size 64×64 and batch size 32. Following the training details in the official paper [29], we use Adam optimizer with β1=0.9, β2=0.999, and ϵ=10−8. The total training epochs are 500. The initial learning rate is set as 2×10−4 and reduced by half at epochs [250,400,425,450,475]. Moreover, we adopt random rotation and flips for data augmentation.

### 4.2. Model Comparisons

We compare re-trained ELAN and ELAN-2 with official ELAN [29], SwinIR [13], and our DAT-S on five benchmark datasets: Set5 [1], Set14 [26], B100 [19], Urban100 [11], and Manga109 [20] with scale factor ×2. The performance (PSNR/SSIM), parameters, and FLOPs are reported. FLOPs are calculated when the input size is 3×128×128. The results are listed in Table 5. We can observe that using DIV2K and Flickr2K to train ELAN can improve the performance of the model. Compared with training on DIV2K, ELAN obtains 0.1 dB and 0.12 dB gains on Urban100 and Manga109. Meanwhile, increasing the number of efficient long-range attention blocks in ELAN can also advance performance. Furthermore, comparing ELAN-2, SwinIR, and our DAT-S, we can find that our DAT-S outperforms the other two models with fewer parameters and FLOPs. Especially the three models are trained on the same dataset.

### 5. More Quantitative Results

We compare our models: DAT-S, DAT-2, and DAT, with state-of-the-art methods: EDSR [14], SRMDNF [27], RDN [32], OISR [9], RCAN [30], NRNLX [22], RNAN [31], SRFBN [12], SAN [5], RFANet [16], IGNN [34], HAN [23], CSNLLN [22], NLSA [21], CRAN [33], ELAN [29], DFSA [17], SwinIR [13], and CAT-A [4]. We use self-ensemble strategy in testing and mark the model with a symbol “+”. Quantitative results are shown in Table 6. Our DAT outperforms other methods in all cases, except for the PSNR value (CAT-A) on Urban100 (×4). Additionally, our DAT-S and DAT-2 achieve comparable or better performance than previous methods.

### 6. More Visual Results

We provide more visual comparisons in Figs. 4 and 5 as the supplement of the visualization in the main paper. As we can see, most compared methods suffer from blurring artifacts and cannot recover some details in some challenging
In contrast, our DAT can alleviate the blurring artifact to some degree and recover sharp textures. For instance, in img.021, our DAT recovers more textures and patterns than other methods. Similar observations are shown in other images. These visual comparisons further demonstrate that our method has powerful modeling capability by aggregating spatial and channel features.

7. Limitations and Future Work

In this work, we propose the dual aggregation Transformer (DAT), for image SR. Our DAT can aggregate spatial and channel information via the inter-block and intra-block manner, thus obtaining powerful representation ability. Our DAT outperforms recent state-of-the-art image SR methods. Nevertheless, we for more types of image SR tasks (e.g., blind and real-world image SR), we have not explored. We will apply our DAT to more kinds of image SR tasks in the future to further demonstrate the effectiveness of our proposed method. In addition, we mainly focus on designing the Transformer block to aggregate spatial and channel information. For the network architecture, we have not investigated it. In future work, we will explore other network structures, such as parallel or multi-scale architectures.

References


[34] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. In NeurIPS, 2020. 4, 7
Table 6: Quantitative comparison with state-of-the-art methods. The best and second-best results are coloured red and blue.
Figure 4: Visual comparison for image SR (×4) in some challenging cases.
Figure 5: Visual comparison for image SR (×4) in some challenging cases.