AdaFormer: Efficient Transformer with Adaptive Token Sparsification for Image Super-resolution

Xiaotong Luo[1*], Zekun Ai[1*], Qiuyuan Liang[1], Ding Liu[2], Yuan Xie[3†], Yanyun Qu[1†], Yun Fu[4]

[1]School of Informatics, Xiamen University, Fujian, China
[2]Bytedance Inc.
[3]School of Computer Science and Technology, East China Normal University, Shanghai, China
[4]Northeastern University
{xiaotluo, aizekun}@stu.xmu.edu.cn, yyqu@xmu.edu.cn

Abstract

Efficient transformer-based models have made remarkable progress in image super-resolution (SR). Most of these works mainly design elaborate structures to accelerate the inference of the transformer, where all feature tokens are propagated equally. However, they ignore the underlying characteristic of image content, i.e., various image regions have distinct restoration difficulties, especially for large images (2K-8K), failing to achieve adaptive inference. In this work, we propose an adaptive token sparsification transformer (AdaFormer) to speed up the model inference for image SR. Specifically, a texture-relevant sparse attention block with parallel global and local branches is introduced, aiming to integrate informative tokens from the global view instead of only in fixed local windows. Then, an early-exit strategy is designed to progressively halt tokens according to the token importance. To estimate the plausibility of each token, we adopt a lightweight confidence estimator, which is constrained by an uncertainty-guided loss to obtain a binary halting mask about the tokens. Experiments on large images have illustrated that our proposal reduces nearly 90% latency against SwinIR on Test8K, while maintaining a comparable performance.

Introduction

Single image super-resolution (SISR) aims to reconstruct high-resolution (HR) images from the degraded low-resolution (LR) counterparts. With the remarkable progress of transformer on high-level vision tasks (Yin et al. 2022b; Pang et al. 2023), a growing number of transformer-based methods (Liang et al. 2021a; Cai et al. 2023) have emerged for image SR, which significantly exceed the convolutional neural network (CNN) based methods in performance by mining the long-range pixel dependencies. However, most of these methods are time-consuming with substantial computational complexity. It is unbearable for intelligent devices with image resolution reached 4K (4096×2160) or even 8K (7680×4320). Therefore, how to achieve efficient SR for large images with lower computational complexity and inference time is an urgent problem to be solved.

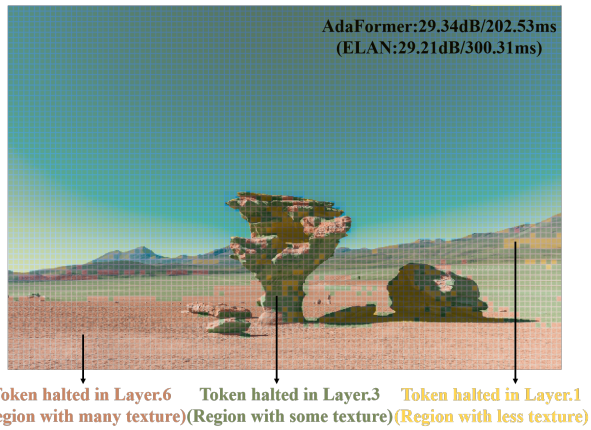

Figure 1: The SR result ($4\times$) of AdaFormer. It visualizes the token halted at different layers in yellow, green, and brown masks. Compared with ELAN (Zhang et al. 2022), our method achieves better PSNR with lower latency.

Recently, several efficient transformer-based SR methods have been proposed to reduce computational expenses. SwinIR (Liang et al. 2021a) is the first efficient transformer for image restoration, which introduces the local sliding window mechanism and performs self-attention (SA) within each window to speed up inference. ESRT (Lu et al. 2022) designs a lightweight hybrid backbone, which reduces the channel number of multi-head SA with a high-frequency filter module. Besides, ELAN (Zhang et al. 2022) employs group-wise multi-scale SA with different local window sizes and the shared-attention mechanism to save more computational costs. Although these methods have made some progress in inference speed, there still exist two underlying limitations. Firstly, they cannot achieve dynamic inference once the model is trained, since all the tokens are propagated without distinction and the relevance between recovery degree and computational resources for different tokens is ignored. Secondly, they mainly focus on designing the efficient window-based local SA, whereas the global edge and texture information cannot be well captured due to the limited receptive field of fixed window size.

According to the fact that different image patches have

*These authors contributed equally.
†Corresponding authors.

various reconstruction difficulties (Kong et al. 2021), some effective dynamic inference strategies for CNN-based models have been proposed (Xie et al. 2021; Wang et al. 2022), while it is still undeveloped for transformer-based SR models. Meanwhile, the dynamic token selection in high-level tasks (Fayyaz et al. 2021; Liang 2022; Yin et al. 2022b; Feng and Zhang 2023) motivates us to explore the token sparsification for efficient image SR. Note that these methods cannot be directly applied to low-level tasks. They either rely on the class token or introduce an extra scoring network to evaluate the token importance. However, image SR is a regression task, i.e., the output is pixel intensities, which lacks explicit semantic information and cannot provide direct guidance for token exiting. Thus, how to dynamically select the well-recovered tokens for SR is challenging.

To address the above issues, we propose an efficient transformer with adaptive token sparsification (AdaFormer) to speed up the inference on large images. Specifically, a texture-relevant sparse attention block (TSAB) is first designed, which includes a global cross-attention branch and a standard local SA branch. The global branch aims to retrieve global edge or texture information, which makes up for the limited receptive field of local windows. Then, a token early-exit strategy is introduced to dynamically filter out the trivial tokens in the two branches. To evaluate the token importance, a lightweight confidence estimator constrained with an uncertainty-guided loss is adopted to obtain a pixel-wise confidence map. Then, a binary halting mask is generated by accumulating the confidence score so as to adaptively halt the corresponding tokens. As shown in Fig. 1, we present the visualization diagram of adaptive token halting depth for our AdaFormer, where the image is divided into different windows (SA calculation units) to match the corresponding tokens. It is observed that different layers are allocated to tokens with various restoration difficulties. To the best of our knowledge, this is the first work to investigate the dynamic inference of the transformer with token-level sparsification for image SR.

In summary, the main contributions are four-fold:

- An adaptive token sparsification transformer (AdaFormer) is proposed for efficient image SR, which introduces a texture-relevant sparse attention block (TSAB) with a token early-exit strategy.
- TSAB is designed for integrating the global and local texture information, which aims to eliminate the limited receptive field of the standard local sliding window.
- The early-exit strategy is adopted to achieve dynamic inference, where the token importance is evaluated via a confidence estimator constrained by an uncertainty loss.
- Extensive experiments demonstrate that our AdaFormer outperforms the state-of-the-art efficient transformer-based SR methods with less inference time.

Related Work

Efficient CNN-based SR Methods

The CNN-based SR methods have revealed remarkable progress, whereas most efficient models mainly rely on elaborate structure design. SRCNN (Dong et al. 2016) firstly designs a three-layer CNN to learn the mapping relation between the bicubic-upsampled LR image and the HR image. IMDN (Hui et al. 2019) designs information multi-distillation blocks to capture multi-level features by enlarging the receptive field. RFDN (Liu, Tang, and Wu 2020) and RLFN (Kong et al. 2022) propose the feature distillation connection and residual local feature learning for lightweight SR. LatticeNet (Luo et al. 2023) proposes a lattice block to assemble pair-wise residual blocks by learnable combination coefficients. EDTS (Chao et al. 2023) transforms time-consuming operations and speeds up the inference without damaging reconstruction accuracy. Though these models have obtained excellent results, the performance is still restricted by the local property of the convolution operation and the equal treatment of spatial features.

Efficient Transformer-based SR Methods

Transformer has emerged promising potential in computer vision (Cai et al. 2023; Hsu, Liao, and Huang 2023). IPT (Chen et al. 2021) is a backbone model based on the standard transformer for various low-level tasks and is pre-trained on large-scale datasets with abundant computational resources. SwinIR (Liang et al. 2021a) utilizes multiple swin transformer blocks with local attention and shifted-window interaction to generate excellent results. ESRT (Lu et al. 2022) designs a hybrid model including a lightweight CNN backbone with a high preserving block and a lightweight transformer backbone with a folding technique to reduce the channel numbers. ELAN (Zhang et al. 2022) excavates the long-range image dependency by calculating SA with different window sizes on non-overlapping feature groups. GRL-B (Li et al. 2023) models feature hierarchies within the regional, local and global range by the window-based SA, channel attention enhanced convolution operation and anchored stripe SA. N-Gram (Choi, Lee, and Yang 2023) introduces the N-Gram context to enlarge the receptive field for restoring the degraded pixel via the sliding-WSA. However, these methods only focus on how to design efficient SA structures. Here, we propose a token early-exit mechanism to accelerate inference by the recovery degree of each token.

Adaptive Inference in Visual Transformers

Dynamic inference in vision transformers for high-level tasks has been widely explored, which can be classified as hard pruning and soft pruning. The hard pruning approaches aim to filter out the trivial tokens by a predefined scoring mechanism. DynamicViT (Rao et al. 2021) and AdaViT (Meng et al. 2022) introduce additional predictive networks to score tokens. Evo-ViT (Xu et al. 2022), ATS (Fayyaz et al. 2021), and EViT (Liang 2022) use the class tokens to assess other token importance. However, it is difficult to achieve accurate scoring so as to suffer from a significant drop in accuracy. The soft pruning method generates new tokens from image tokens via introducing additional attention models (Ryoo et al. 2021). Various attempts have been made in high-level tasks, whereas few discuss them for low-level.

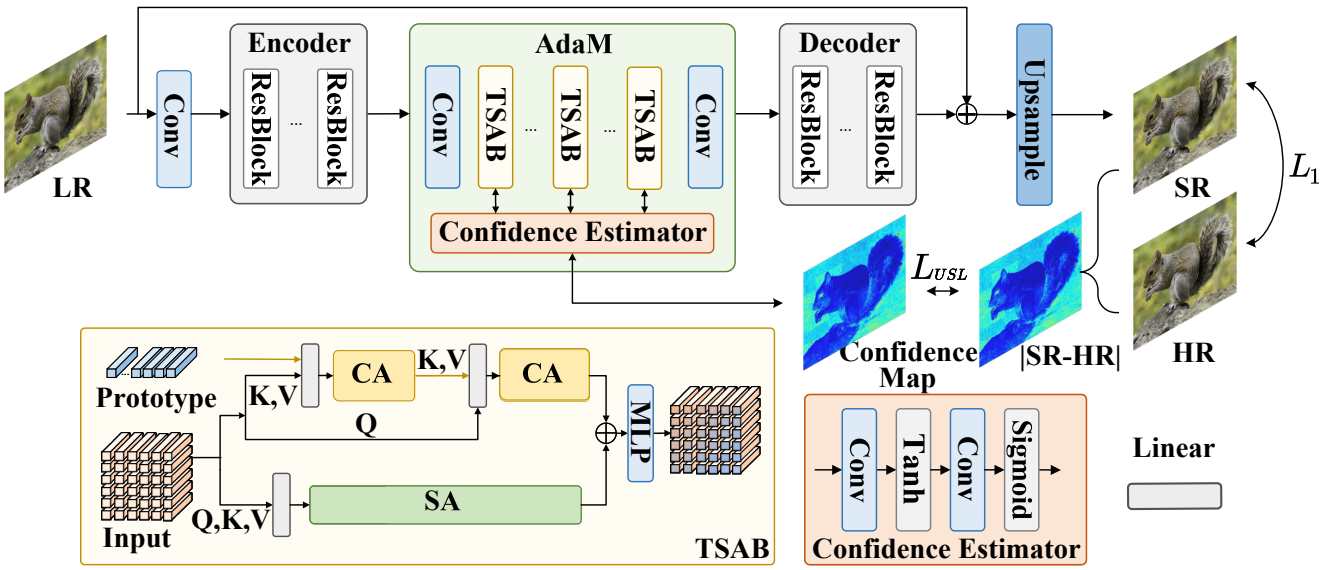

Figure 2: The overall framework of the proposed AdaFormer. It includes an Encoder, an **Ada**ptive token sparsification **M**odule (AdaM) with the token early-exit strategy, a Decoder, and an Upsampling module. AdaM mainly consists of L texture-relevant sparse attention blocks (TSAB), which combine a local self-attention (SA) branch with a global cross-attention (CA) branch. The confidence estimator aims to measure the token importance and provide the halting indication.

Proposed Method

Network Architecture

Overview. In Fig. 2, AdaFormer includes an Encoder, an **Ada**ptive token sparsification **M**odule (AdaM), a Decoder, and an Upsampling module. Let's denote the LR image as $I_{lr} \in \mathbb{R}^{H\times W\times 3}$, where H and W are the height and width. First, I_{lr} is fed into the Encoder H_{ef} consisting of several residual blocks to extract the local context feature F_{ef}:

$$F_{ef} = H_{ef}(I_{lr}) \in \mathbb{R}^{H\times W\times C}, \tag{1}$$

where C is the number of feature channels. Then, AdaM (H_{AdaM}) is employed to adaptively mine the global and local similarity dependence among F_{ef}:

$$F_{gf} = H_{AdaM}(F_{ef}) \in \mathbb{R}^{H\times W\times C}. \tag{2}$$

Next, F_{gf} is input to the Decoder H_{df}, which includes several residual blocks to further enhance local information:

$$F_{df} = H_{df}(F_{gf}) \in \mathbb{R}^{H\times W\times C}. \tag{3}$$

Finally, F_{df} is fed to the Upsampling module H_{up} with pixel shffule (Shi et al. 2016) to obtain the SR output:

$$I_{sr} = H_{up}(F_{df}). \tag{4}$$

By integrating the convolution-based encoder and decoder with the transformer, the advantages of local information extraction and global context modeling can be sufficiently exploited so as to enhance model representation.

Adaptive token sparsification module. AdaM consists of texture-relevant sparse attention blocks (TSAB) and convolutional layers. Given the input feature F_{ef} from the Encoder, we first tokenize it by a 1×1 convolutional layer:

$$T^0 = H_{conv}(F_{ef}) \in \mathbb{R}^{H\times W\times D}. \tag{5}$$

where D is the embedding dimension. Then, we excavate local and global texture dependencies by L TSABs H_{tsab}:

$$T^l = H_{tsab}(T^{l-1}), \quad l = 1, 2, ..., L, \tag{6}$$

Finally, a 1×1 convolutional layer is adopted to align the feature dimension and obtain the output F_{gf}:

$$F_{gf} = H_{conv}(T^L). \tag{7}$$

Texture-relevant sparse attention block. The existing transformer-based SR works adopt the window based self-attention (SA) as the basic component. However, they cannot effectively aggregate global information since SA is calculated within limited local range. To address this, we design TSAB, which introduces a global cross-attention branch parallel with the local SA branch. It aims to mine effective information within local windows and global dependencies.

Given the input tokens $T^l \in \mathbb{R}^{H\times W\times D}$ from the l-th TSAB, we feed them into the local and global branches. Note that the token early-exit strategy is performed on T^l to get the local tokens T^l_{local} and global tokens T^l_{global}.

(1) *Local self-attention branch*. Similar to SwinIR (Liang et al. 2021b), we adopt the standard window-based SA for the local branch. The local tokens $T^l_{local} \in \mathbb{R}^{\frac{HW}{S^2}\times S^2\times D}$ consist of $\frac{HW}{S^2}$ windows, which are obtained by partitioning T^l into non-overlapped $S\times S$ local windows. Then, SA is performed within each local window feature $X \in \mathbb{R}^{S^2\times D}$ to get the enhanced feature $\hat{T}^l_{local}$:

$$Q_l = W_{ql}X, K_l = W_{kl}X, V_l = W_{vl}X, \tag{8}$$

$$\hat{T}^l_{local} = \text{SoftMax}(Q_l K_l^{\mathrm{T}}/\sqrt{d})V_l, \tag{9}$$

where Q_l, K_l and V_l are generated by linear projections with matrix W_{ql}, W_{kl} and W_{vl}, respectively.

(2) *Global cross-attention branch*. The global branch is

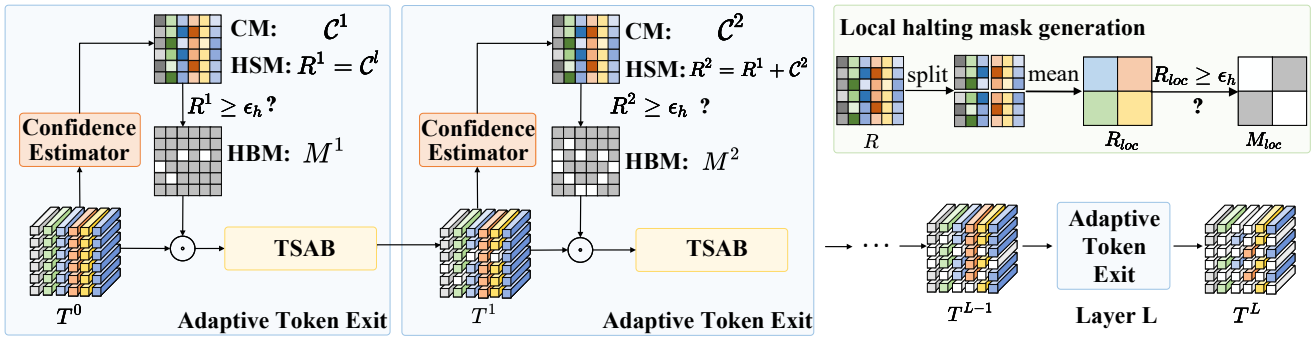

Figure 3: The pipeline of the token early-exit strategy. The confidence map (CM) $\mathcal{C}^l$ is measured for the input tokens T^l from the l-th layer. Then, a halting score map (HSM) R^l is calculated by accumulating the confidence map to generate a halting binary mask (HBM) M^l. Finally, with the guide of M^l, only the kept tokens will participate in the attention calculation.

adopted to remedy the limited receptive field of the local SA branch. Inspired by (Yin et al. 2022a), we introduce the texture-relevant prototype $P \in \mathbb{R}^{M \times D}$ as the query to retrieve the global edge and texture, which is initialized as 1 and learns M texture-shared prototypes as seed vectors. It acts as the role of class token in image classification, which represents the global attention of the image by weight allocation to other tokens to some extent. Therefore, the cross-attention (CA) between the texture-relevant prototype P and the global tokens T^l_{global} is calculated as:

$$Q_p = W_{qp}P, K_g = W_{kg}T^l_{global}, V_g = W_{vg}T^l_{global}, \quad (10)$$

$$\hat{P} = \text{SoftMax}(Q_p K_g^{\mathrm{T}}/\sqrt{d})V_g, \quad (11)$$

where $\hat{P}$ is the learned informative prototype, $Q_p \in \mathbb{R}^{M \times D}$ and $K_g, V_g \in \mathbb{R}^{HW \times D}$ are the query, key and value tensors, projected by learnable linear matrix W_{qp}, W_{kg} and W_{vg}. Next, we aggregate informative texture by performing cross-attention between the global tokens T^l_{global} with $\hat{P}$:

$$Q_g = W_{qg}T^l_{global}, K_p = W_{kp}\hat{P}, V_p = W_{vp}\hat{P}, \quad (12)$$

$$\hat{T}^l_{global} = \text{SoftMax}(\mathrm{Q_g K_p^T}/\sqrt{\mathrm{d}})\mathrm{V_p}, \quad (13)$$

where W_{qg}, W_{kp} and W_{vp} are learnable matrices. Finally, we project the element-wise sum of these extracted local and global features by a multi-layer perception (MLP) to obtain the output of the l-th TSAB as:

$$T^{l+1} = \text{MLP}(\hat{T}^l_{local} + \hat{T}^l_{global}) \in \mathbb{R}^{H \times W \times D}, \quad (14)$$

where MLP follows a standard structure composed of 1×1 convolution, GLUE, and 1×1 convolution layers.

Token Early-Exit Strategy

Inspired by the dynamic inference in high-level tasks (Yin et al. 2022b; Liang 2022), we propose a token early-exit strategy to speed up the model inference. The whole pipeline is depicted in Fig. 3. The core idea is to adaptively halt the SA calculation of well-recovered tokens as the network depth increases. Unlike image classification where the importance of each token depends on its contribution to the classified result, the main challenge for image SR is how to provide accurate halting signals.

Specifically, we adopt a confidence estimator to calculate the token importance, which indicates the recovery degree of tokens in the TSAB. The higher the confidence, the better the recovery effect. In order to halt tokens at different depths, the confidence map is progressively accumulated to obtain the halting score map, thus generating a binary halting mask to indicate whether the current token should exit or not.

Confidence estimation. Given the intermediate tokens $T^l \in \mathbb{R}^{H \times W \times D}$ from the l-th TSAB, the confidence map $\mathcal{C}^l \in \mathbb{R}^{H \times W \times 1}$ for T^l is measured by a weight-shared lightweight confidence estimator, which consists of Conv-Tanh-Conv-Sigmoid layers, *i.e.*,

$$\mathcal{C}^l = \text{Sigmoid}(\text{Conv}(\text{Tanh}(\text{Conv}(T^l)))). \quad (15)$$

where Conv denotes a 3×3 convolutional layer.

Inspired by (Ning et al. 2021), we adopt aleatoric uncertainty to perform confidence estimation, which aims to transform texture and edge pixels with high uncertainty into low-confidence representations, and flat regions with low uncertainty into high-confidence representations. Specifically, given the LR image I_{lr} and the corresponding HR image I_{hr} and the SR image I_{sr}, the aleatoric uncertainty can be modeled with an additional parameter term $\boldsymbol{\theta}$. In order to accurately estimate $\boldsymbol{\theta}$, Laplace distribution is used to model the Likelihood Function, which can be formulated as:

$$\ln p\left(I_{hr}, \boldsymbol{\theta} \mid I_{lr}\right) = -\frac{\left\|I_{hr} - I_{sr}\right\|_1}{\boldsymbol{\theta}} - \ln \boldsymbol{\theta} - \ln 2. \quad (16)$$

To transform the uncertainty estimation into the confidence estimation, we model $\boldsymbol{\theta} = \frac{1}{(\mathcal{C}^l)\uparrow}$, where the bilinear interpolation $(\cdot)\uparrow$ is adopted to upsample $\mathcal{C}^l$ to align the size with I_{hr}. Then, Eq. (16) can be reformulated as:

$$\mathcal{L}_{USL} = \sum_{l=1}^{L}(\mathcal{C}^l \left\|I_{hr} - I_{sr}\right\|_1 + \log \frac{1}{(\mathcal{C}^l + \epsilon)}), \quad (17)$$

where $\epsilon = 1e^{-8}$ is a small constant for stable training. By the confidence estimation about the recovery credibility of each token, the tokens with high enough confidence can be halted in the current layer to reduce inference time.

Sparse attention with halting mask. Here, we present how to calculate the sparse attention in the local and global branches of the TSAB with dynamic token selection. To achieve this, we first calculate the halting score map $R^l \in$

$\mathbb{R}^{H\times W\times 1}$ by accumulating $\mathcal{C}^l$ layer by layer:

$$R^l = R^{l-1} + \mathcal{C}^{l-1}. \tag{18}$$

When R^l exceeds some threshold, we can obtain a halting mask to indicate whether the tokens should halt.

In the local SA branch, the local halting mask M_{loc} is generated to progressively reduce well-recovered patch windows. As depicted in the upper right corner of Fig. 3, we partition the halting score map $R^l \in \mathbb{R}^{H\times W\times 1}$ into non-overlapping windows with the size of $\frac{HW}{S^2} \times S^2 \times 1$, and calculate the mean value within each window to obtain R^l_{loc}. Then, M_{loc} can be obtained by:

$$M_{loc} = \begin{cases} 0 & \text{if } R^l_{loc} \le 1-\epsilon_h, \\ 1 & \text{if } R^l_{loc} \ge 1-\epsilon_h, \end{cases} \tag{19}$$

where ϵ_h is a small positive constant. The local window with $M_{loc} = 1$ will be halted for the next attention calculation. Therefore, the input token T^l_{local} for the local branch is:

$$T^l_{local} \leftarrow T^l \odot (1 - M_{loc}). \tag{20}$$

Next, the SA calculation within each kept window is performed as Eq. (8) and Eq. (9) to obtain $\hat{T}^l_{local}$ during the inference and the number of local tokens entered in TSAB is less than HW/S^2. And then we reshape $\hat{T}^l_{local}$ and M_{loc} back to the size of $H \times W \times D$ and $H \times W \times 1$, respectively. The original input value of halted windows are added, so the output of the local branch is represented as:

$$\hat{T}^l_{local} \leftarrow \hat{T}^l_{local} \odot (1 - M_{loc}) + T^l \odot M_{loc}. \tag{21}$$

In the global CA branch, the global halting mask M_{glo} is calculated for each spatial position as follows:

$$M_{glo} = \begin{cases} 0 & \text{if } R^l \le 1-\epsilon_h, \\ 1 & \text{if } R^l \ge 1-\epsilon_h, \end{cases} \in \mathbb{R}^{H\times W\times 1} \tag{22}$$

When $M^{ij}_{glo} = 1$, it means that the token in the position (i, j) is credible enough to be halted. Conversely, it will be fed to the next processing. With the guide of M_{glo}, the kept tokens T^l_{global} are selected from T^l with $M_{glo} = 0$ as:

$$T^l_{global} \leftarrow T^l \odot (1 - M_{glo}). \tag{23}$$

Then, T^l_{global} is flattened to $\mathbb{R}^{HW\times D}$ to calculate CA as Eq. (13). Finally, $\hat{T}_{global}$ and M_{glo} are reshaped to the original dimensions $\mathbb{R}^{H\times W\times D}$ and $\mathbb{R}^{H\times W\times 1}$, and add the original halt tokens to get the output of the global branch:

$$\hat{T}^l_{global} \leftarrow \hat{T}^l_{global} \odot (1 - M_{glo}) + T^l \odot M_{glo}. \tag{24}$$

Training Objective

Our AdaFormer adopts the commonly used L_1 loss and the uncertainty loss in Eq. (17) as the training objective, i.e.,

$$L = L_1 + \alpha L_{USL}, \tag{25}$$

where α is a regularization coefficient, empirically set as 1.

Experiments

Experimental Setup

Datasets. We use DIV2K (Timofte et al. 2017) (0001-0800) as the training dataset and evaluate our model on 100 images (0801-0900) of DIV2K. Following ClassSR (Kong et al. 2021), the model is tested on 300 images (1201-1500) from the DIV8K (Gu et al. 2019) for $4\times$ SR, which consists of Test2K, Test4K, and Test8K. The LR images are captured by bicubic downsampling to HR images. To further illustrate the effectiveness and robustness of our proposal, we also test on four SR benchmarks: Set5, Set14, B100 and Urban100.

Implementation details. The hyperparameters of AdaFormer are set as: the number of residual blocks in the Encoder and Decoder is 8; the number of TSABs L is 6; the number of feature channels C is 64, the embedding dimension D is 64, ϵ_h is 0.05 and M is 16. Following SwinIR, the window size S in the local branch of TSAB is 8. During training, we randomly crop 16 LR patches with the size of 48×48 as the input, which are further augmented by randomly rotated with $90°$, $180°$, $270°$ and flipped horizontally. ADAM optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$ is adopted to train the model. The learning rate is initialized as 2×10^{-4} and decreased by half every 200 epochs. We implement our model using PyTorch on 1 NVIDIA 2080Ti GPU and train for 1000 epochs in total. PSNR and SSIM are adopted as objective metrics, which are measured on the Y channel of YCbCr space. Besides, we present the inference time (Latency) and FLOPs to measure the model complexity, which are both measured by averaging each benchmark.

Comparisons with the State-of-arts

Quantitative results. To demonstrate the effectiveness of AdaFormer, we first compare with state-of-the-art efficient transformer-based SR models, including SwinIR-light (Liang et al. 2021a), ELAN-light (Zhang et al. 2022) and N-Gram (Choi, Lee, and Yang 2023) on large images for $4\times$ SR. In Tab. 1, it is observed that AdaFormer obtains comparable performance with less latency. Especially, it reduces 281ms latency and gains 0.05dB improvement in PSNR when compared with ELAN-light on Test8K. Meanwhile, we compare with several efficient CNN-based SR methods on four standard benchmarks in Tab. 2, i.e., IMDN (Hui et al. 2019), RFDN (Liu, Tang, and Wu 2020), RLFN (Kong et al. 2022), LAPAR-A (Li et al. 2020), and ETDS (Chao et al. 2023). Compared with CNN-based methods, transformer-based methods perform much better in accuracy, while inferior in latency. Compared with transformer-based methods, AdaFormer has obvious superiority in latency time, which is nearly $1.22\times$ faster than ELAN-light with comparable results on B100. Our method unites the advantages of transformer and CNN, balancing latency and accuracy well.

Qualitative results. The visual comparisons on Test2K and Test4K datasets are presented in Fig. 5. For "1215" in Test2K, our AdaFormer recovers more clear edge and texture details than ELAN-light and SwinIR-light. Especially, our AdaFormer obtains a clear boundary between leaves and branches. For "1318" in Test4K, our AdaFormer also obtains more favorable results on the building texture than other methods. Therefore, the proposed AdaFormer has the superiority of capturing better structural information.

The sparsity of halting mask. As shown in Fig. 4, we analyze the sparsity (the percentage of halted tokens to all tokens) of the global and local halting masks on Test4K and Set14 datasets. It shows that the sparsity gradually increases

Model	DIV2K			Test2K			Test4K			Test8K		
	PSNR	SSIM	Latency(ms)	PSNR	SSIM	Latency(ms)	PSNR	SSIM	Latency(ms)	PSNR	SSIM	Latency(ms)
SwinIR-light	30.60	**0.8427**	934.96	27.69	**0.7786**	466.48	29.11	0.8249	2103.71	35.04	0.8969	11766.19
ELAN-light	30.60	0.8417	147.26	27.69	0.7778	82.61	29.12	0.8246	300.31	35.08	0.8970	1636.38
N-Gram	30.60	0.8418	345.	27.70	0.7780	508.15	29.10	0.8245	1908.14	35.08	0.8971	3441.76
AdaFormer (ours)	**30.63**	0.8424	**112.35**	**27.70**	0.7782	**71.99**	**29.15**	**0.8252**	**257.94**	**35.13**	**0.8974**	**1355.34**

Table 1: The quantitative (PSNR(dB)/SSIM) and latency (ms) comparisons with different efficient transformer-based SR models on DIV2K, Test2K, Test4K, and Test8K for $4\times$ SR. The best and second best results are highlighted in bold and underline.

Method	Set5			Set14			B100			Urban100		
	PSNR	SSIM	Latency(ms)	PSNR	SSIM	Latency(ms)	PSNR	SSIM	Latency(ms)	PSNR	SSIM	Latency(ms)
IMDN	32.21	0.9605	8.40	28.58	0.7811	8.78	27.56	0.7353	6.92	26.04	0.7838	12.22
LAPAR-A	32.15	0.8944	12.28	28.61	0.7818	17.75	27.61	0.766	12.89	26.14	0.7871	37.14
RFDN	32.28	0.8957	44.22	28.61	0.7818	19.55	27.58	0.7363	17.96	26.20	0.7883	29.55
RLFN	32.24	0.8952	-	28.62	0.7813	-	27.60	0.7364	-	26.17	0.7877	-
ETDS	31.69	0.8889	11.19	28.31	0.7751	11.39	27.37	0.7302	13.22	25.47	0.7643	8.49
SwinIR-light	**32.44**	**0.8976**	54.41	28.77	0.7858	79.83	27.69	0.7406	64.02	26.47	0.7980	202.21
ESRT	32.19	0.8947	22.75	28.69	0.7833	26.60	27.69	0.7379	21.54	26.39	0.7962	88.22
ELAN-light	32.43	0.8975	36.06	28.78	0.7858	29.52	27.69	0.7406	27.06	**26.54**	0.7982	59.11
N-Gram	32.33	0.8963	82.76	28.78	**0.7859**	116.05	27.66	0.7396	96.26	26.45	0.7963	238.51
AdaFormer (ours)	32.43	0.8974	**22.28**	**28.80**	0.7858	**25.28**	**27.70**	**0.7407**	**20.87**	26.48	**0.7982**	**55.05**

Table 2: The quantitative (PSNR(dB)/SSIM) and latency (ms) comparisons with different efficient SR models on benchmark datasets for $4\times$ SR. The best and second best results are highlighted in bold and underline.

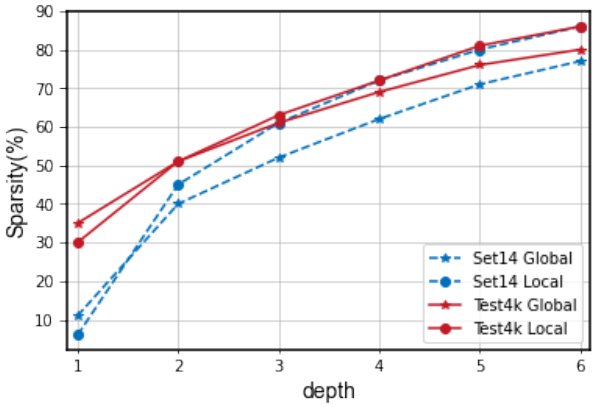

Figure 4: The sparsity of the global and local halting masks of TSAB for different depths on Test4K and Set14.

as the number of TSAB increases. Especially, the sparsity on Test4K is almost 90% in the last TSAB. Therefore, it demonstrates that our adaptive token-exit strategy reduces latency and computational costs substantially.

Visualization of confidence map and halting mask. We visualize the learned confidence map and the halting mask on Test4K in Fig. 6. Remarkably, our method adaptively halts tokens according to different restoration difficulties. For flat regions, plenty of the tokens are halted in the early stage to speed up the inference, while for regions with more complicated textures and edges, the model tends to keep the tokens restored until they reach the required confidence score. Therefore, it demonstrates the reliability of the exit signals guided by the uncertainty constraint.

Ablation Study

The ablation analysis includes the local and global branches and the early-exit strategy. Note that all SR models in the ablation are trained for 400 epochs and tested on Test4K.

Break-down ablation. As shown in Tab. 3, we perform an ablation to investigate the effect of different components in AdaFormer, which includes the following variants: ***Case 1***: following SwinIR (Liang et al. 2021a), we adopt the local branch as the baseline model, which calculates the SA in a local sliding window. It obtains 29.03dB on Test4K. ***Case 2***: introducing the global CA branch based on the baseline model. It gains by 0.03dB with 42G FLOPs and 92ms latency increase. This means that the global CA branch can lead to an increase in performance, FLOPs, and latency. ***Case 3***: adopting the early-exit strategy on Case 2 with the local halting mask. The FLOPs and latency are reduced by 10G and 72ms with comparable performance. ***Case 4***: applying the early-exit strategy on both branches. The Flops and latency are further reduced by 63G and 77ms, while the performance is slightly improved by 0.06dB. It also shows that the token early-exit strategy reduces FLOPs and inference time. Meanwhile, it alleviates the overfitting problem and unnecessary noise by performing attention calculations on the informative tokens Therefore, our method strikes an excellent tradeoff between latency and accuracy.

Early-exit strategy comparison. To validate the effectiveness of our token early-exit strategy, we compare it with several other halting strategies as shown in Tab. 4. ***1) Comparison to uniform-exit and random-exit.*** We first compare with uniform-exit and random-exit strategies. It is observed that our strategy obtains 0.32dB and 0.25dB improvement in PSNR with lower FLOPs and inference time. ***2) Comparison to A-Vit.*** Following the exit strategy of A-Vit (Yin

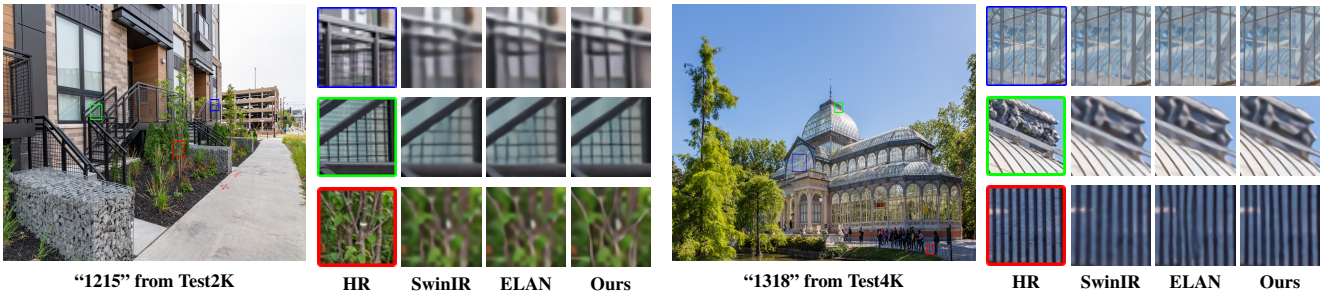

Figure 5: Visual comparisons with the state-of-the-art methods on Test2K and Test4K for $4\times$ SR. Zoom in for a better view.

Figure 6: Visualization of the confidence map (left bottom) and binary halting mask (right) for the original image (left upper). In the confidence map, light/dark color represents a higher/lower confidence score requiring less/more computation. In the halting mask, the halting pixel/patch is represented by the white color. Please zoom in for a better view.

Case	TSAB		Early-exit		Test4K		
	LB	GB	LM	GM	FLOPs(G)	Latency(ms)	PSNR(dB)
1	✓				654.25	314.98	29.03
2	✓	✓			696.08	406.83	29.06
3	✓	✓	✓		686.83	334.69	29.04
4	✓	✓	✓	✓	623.51	257.94	29.10

Table 3: Ablation study of the proposed AdaFormer on Test4K dataset for $4\times$ SR.

Early-Exit Strategy	PSNR (dB)	Latency(ms)	FLOPs (G)
Uniform Exit	28.78	310.01	671.78
Random Exit	28.85	238.10	653.67
A-Vit	28.81	374.87	648.52
APE	29.10	499.45	664.92
WU estimator	29.09	279.92	666.44
WS estimator	**29.10**	**267.24**	**623.51**

Table 4: Quantitative comparisons of different early-exit strategies on Test4K for $4\times$ SR.

et al. 2022b), we adopt the ponder loss to encourage early exit and regularize the halting distribution towards Gaussian distribution using KL divergence. It suffers a 0.29dB drop in PSNR with higher latency and FLOPs compared to our method. The reason is that A-Vit is designed for image classification task, which adopts predefined prior knowledge to constrain the token halting distribution to Gaussian distribution. Since objects in classification tasks are mainly concentrated in the image center, it is consistent with the fact that most samples in ImageNet are centered. However, image SR is more concerned with the image texture than the central object, so it is unreasonable to use predefined priors for the constraints. ***3) Comparison to APE.*** APE (Wang et al. 2022) introduces an incremental capacity measured by PSNR for CNN-based methods to judge whether the patch should exit or not. We apply the incremental capacity as an exit signal to train the transformer-based SR model. It shows that our proposal reduces the inference time by nearly 50% with similar performance. Besides, we compare weight-shared (WS) and weight-unshared (WU) confidence estimators , showing that the shared-weight estimator performs better with less latency. Therefore, it demonstrates that our adaptive token early-exit strategy is superior in performance and efficiency.

Conclusion

In this paper, we propose an adaptive token sparsification module with an early-exit strategy to accelerate the inference of the transformer for image SR. The key idea is using a confidence estimator constrained by an uncertainty-driven loss to obtain the binary halting mask, which provides a halting signal for each token to indicate its recovery importance. Besides, a texture-relevant sparse attention block is designed for local and global texture information interaction. Extensive experimental results show that our proposal outperforms the mainstream efficient transformer-based methods in less latency with comparable performance.

Acknowledgments

The authors, except Yun Fu, were supported by National Natural Science Foundation of China under Grant No.62176224, No.62222602, No.62176092; Natural Science Foundation of Chongqing under No.CSTB2023NSCOJOX0007, CCF-Lenovo Blue Ocean Research Fund.

References

Cai, Q.; Qian, Y.; Li, J.; Lyu, J.; Yang, Y.; Wu, F.; and Zhang, D. 2023. HIPA: Hierarchical Patch Transformer for Single Image Super Resolution. *TIP*.

Chao, J.; Zhou, Z.; Gao, H.; Gong, J.; Yang, Z.; Zeng, Z.; and Dehbi, L. 2023. Equivalent Transformation and Dual Stream Network Construction for Mobile Image Super-Resolution. In *CVPR*.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021. Pre-trained image processing transformer. In *CVPR*.

Choi, H.; Lee, J.; and Yang, J. 2023. N-Gram in Swin Transformers for Efficient Lightweight Image Super-Resolution.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2016. Image Super-Resolution Using Deep Convolutional Networks. *TPAMI*.

Fayyaz, M.; Koohpayegani, S. A.; Jafari, F. R.; Sommerlade, E.; Joze, H. R. V.; Pirsiavash, H.; and Gall, J. 2021. ATS: Adaptive Token Sampling For Efficient Vision Transformers. abs/2111.15667.

Feng, Z.; and Zhang, S. 2023. Efficient Vision Transformer via Token Merger. *TIP*.

Gu, S.; Lugmayr, A.; Danelljan, M.; Fritsche, M.; Lamour, J.; and Timofte, R. 2019. Div8k: Diverse 8k resolution image dataset. In *ICCVW*.

Hsu, T.; Liao, Y.; and Huang, C. 2023. Video Summarization With Spatiotemporal Vision Transformer. *TIP*.

Hui, Z.; Gao, X.; Yang, Y.; and Wang, X. 2019. Lightweight Image Super-Resolution with Information Multi-distillation Network. In *ACM MM*.

Kong, F.; Li, M.; Liu, S.; Liu, D.; He, J.; Bai, Y.; Chen, F.; and Fu, L. 2022. Residual Local Feature Network for Efficient Super-Resolution. In *CVPR*.

Kong, X.; Zhao, H.; Qiao, Y.; and Dong, C. 2021. ClassSR: A General Framework to Accelerate Super-Resolution Networks by Data Characteristic. In *CVPR*.

Li, W.; Zhou, K.; Qi, L.; Jiang, N.; Lu, J.; and Jia, J. 2020. LAPAR: Linearly-Assembled Pixel-Adaptive Regression Network for Single Image Super-resolution and Beyond. In *NeurIPS*.

Li, Y.; Fan, Y.; Xiang, X.; Demandolx, D.; Ranjan, R.; Timofte, R.; and Van Gool, L. 2023. Efficient and explicit modelling of image hierarchies for image restoration. In *CVPR*.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021a. Swinir: Image restoration using swin transformer. In *ICCV*.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021b. Swinir: Image restoration using swin transformer. In *ICCVW*.

Liang, Y. 2022. EViT: Expediting Vision Transformers via Token Reorganizations. In *ICLR*.

Liu, J.; Tang, J.; and Wu, G. 2020. Residual Feature Distillation Network for Lightweight Image Super-Resolution. In Bartoli, A.; and Fusiello, A., eds., *ECCVW*.

Lu, Z.; Li, J.; Liu, H.; Huang, C.; Zhang, L.; and Zeng, T. 2022. Transformer for single image super-resolution. In *CVPRW*.

Luo, X.; Qu, Y.; Xie, Y.; Zhang, Y.; Li, C.; and Fu, Y. 2023. Lattice Network for Lightweight Image Restoration. *TPAMI*.

Meng, L.; Li, H.; Chen, B.; Lan, S.; Wu, Z.; Jiang, Y.; and Lim, S. 2022. AdaViT: Adaptive Vision Transformers for Efficient Image Recognition. In *CVPR*.

Ning, Q.; Dong, W.; Li, X.; Wu, J.; and Shi, G. 2021. Uncertainty-Driven Loss for Single Image Super-Resolution. In *NeurIPS*.

Pang, Y.; Zhao, X.; Zhang, L.; and Lu, H. 2023. CAVER: Cross-Modal View-Mixed Transformer for Bi-Modal Salient Object Detection. *TIP*.

Rao, Y.; Zhao, W.; Liu, B.; Lu, J.; Zhou, J.; and Hsieh, C.-J. 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *NeurIPS*.

Ryoo, M. S.; Piergiovanni, A. J.; Arnab, A.; Dehghani, M.; and Angelova, A. 2021. TokenLearner: What Can 8 Learned Tokens Do for Images and Videos? abs/2106.11297.

Shi, W.; Caballero, J.; Huszar, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *CVPRW*.

Timofte, R.; Agustsson, E.; Van Gool, L.; Yang, M.-H.; and Zhang, L. 2017. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*.

Wang, S.; Liu, J.; Chen, K.; Li, X.; Lu, M.; and Guo, Y. 2022. Adaptive Patch Exiting for Scalable Single Image Super-Resolution. In *ECCV*.

Xie, W.; Song, D.; Xu, C.; Xu, C.; Zhang, H.; and Wang, Y. 2021. Learning Frequency-aware Dynamic Network for Efficient Super-Resolution. In *ICCV*.

Xu, Y.; Zhang, Z.; Zhang, M.; Sheng, K.; Li, K.; Dong, W.; Zhang, L.; Xu, C.; and Sun, X. 2022. Evo-ViT: Slow-Fast Token Evolution for Dynamic Vision Transformer. In *AAAI*.

Yin, D.; Ren, X.; Luo, C.; Wang, Y.; Xiong, Z.; and Zeng, W. 2022a. Retriever: Learning Content-Style Representation as a Token-Level Bipartite Graph. In *ICLR*.

Yin, H.; Vahdat, A.; Alvarez, J. M.; Mallya, A.; Kautz, J.; and Molchanov, P. 2022b. A-ViT: Adaptive Tokens for Efficient Vision Transformer. In *CVPR*.

Zhang, X.; Zeng, H.; Guo, S.; and Zhang, L. 2022. Efficient Long-Range Attention Network for Image Super-Resolution. In *ECCV*.