The Lighter the Better: Rethinking Transformers in Medical Image Segmentation Through Adaptive Pruning

Xian Lin®, Li Yu®, Senior Member, IEEE, Kwang-Ting Cheng®, Fellow, IEEE, and Zengqiang Yan®, Member, IEEE

Abstract—Vision transformers have recently set off a new wave in the field of medical image analysis due to their remarkable performance on various computer vision tasks. However, recent hybrid-transformer-based approaches mainly focus on the benefits of transformers in capturing long-range dependency while ignoring the issues of their daunting computational complexity, high training costs, and redundant dependency. In this paper, we propose to employ adaptive pruning to transformers for medical image segmentation and propose a lightweight and effective hybrid network APFormer. To our best knowledge, this is the first work on transformer pruning for medical image analysis tasks. The key features of APFormer are self-regularized self-attention (SSA) to improve the convergence of dependency establishment, Gaussian-prior relative position embedding (GRPE) to foster the learning of position information, and adaptive pruning to eliminate redundant computations and perception information. Specifically, SSA and GRPE consider the well-converted dependency distribution and the Gaussian heatmap distribution separately as the prior knowledge of self-attention and position embedding to ease the training of transformers and lay a solid foundation for the following pruning operation. Then, adaptive transformer pruning, both query-wise and dependency-wise, is performed by adjusting the gate control parameters for both complexity reduction and performance improvement. Extensive experiments on two widely-used datasets demonstrate the prominent segmentation performance of APFormer against the state-of-the-art methods with much fewer parameters and lower GFLOPs. More importantly, we prove, through ablation studies, that adaptive pruning can work as a plug-n-play module for performance improvement on other hybrid-transformer-based methods. Code is available at https://github.com/xianlin7/APFormer.

Index Terms—Transformer, adaptive pruning, medical image segmentation, self-regularized self-attention.

I. INTRODUCTION

Transformer, a new type of neural network, originally designed for natural language processing (NLP) [1], has achieved remarkable performance on various computer vision tasks [2], [3], [4], [5], breaking the longstanding dominance of convolutional neural networks (CNNs) [6], [7], [8], [40]. Compared to CNNs benefiting from translation equivariance, sparse interaction, and weight sharing while suffering from limited receptive fields, transformers reveal extraordinary ability in capturing long-range dependency with a multi-head self-attention module, which is more consistent with human vision in feature extraction [18], [43], [44], [45], [46]. Therefore, transformers have drawn explosive attention in medical image segmentation. Chen et al. [9] proposed TransUNet by introducing transformers into the encoder of U-Net, which is the first exploration of transformers in medical image segmentation. Zhang et al. [10] developed a novel parallel-in-branch architecture to combine CNNs and transformers to make full use of their respective advantages. Karimi et al. [11] proposed the first convolution-free model relying only on self-attention and multilayer perceptron for medical image segmentation and achieved competitive results. Wang et al. [12] designed TransBTS to make convolution and transformer responsible for local and global feature capturing respectively and complement each other. Chen et al. [13] proposed TransAttU net, a CNN-Transformer hybrid model with multi-level attention, to effectively address the information recession problem caused by the limited interaction range of CNNs. Though these transformer-based/hybrid methods have achieved encouraging results for medical image segmentation, they still suffer from poor convergence on small-scale datasets and daunting computational complexity, which hinders the deployment of transformers in clinical scenarios.
To alleviate severe performance degradation of transformers when deployed on small-scale datasets, several methods have been proposed to build a competitive transformer-based model without training/pre-training on large-scale datasets. Xie et al. [14] developed CoTr to speed up the convergence with a deformable self-attention mechanism to focus on partial areas. Gao et al. [15] constructed a U-shaped CNN-Transformer hybrid model to avoid pre-training by making a mass of convolutional operations responsible for capturing local features. Valanarasu et al. [16] devoted to deploying transformers on small-scale datasets by introducing learnable gate parameters to restrain the effect of poor position embedding. Wang et al. [17] proposed boundary-aware transformers for performance improvement by merging boundary-wise prior knowledge. Gao et al. [18] applied depth-wise separable convolution on the linear projection of self-attention and the feed-forward parts of the transformer to allow the model to be trained from scratch on small-scale datasets. Though these methods ease the deployment of transformers on small-scale medical image datasets by introducing inductive biases, the high computational complexity issue still has not been properly addressed.

Approaches for computation reduction of medical transformers can be classified into three categories: dependency establishment based on windows, deformable self-attention, and scale downsampling of queries and keys. Window-based models are mainly inspired by Swin Transformer [3], whose representations are extracted within shifted windows. Karimi et al. [11] built an UNet-like architecture only with Swin Transformer blocks to form a pure transformer-based model with acceptable computational cost. Similarly, Li et al. [19] presented a novel upsampling method by introducing Swin Transformer blocks to the decoder of U-Net. Though encouraging performance has been achieved by these methods, their local features are seriously destroyed by the rigid window partitioning scheme in Swin Transformer. To address this, Lin et al. [20] adopted a double-scale encoder to restore local information loss between different-scaled Swin Transformers. Deformable self-attention methods directly reduce computational complexity by only building dependency within a subset of patches [14], [21]. These two types of approaches ease the computation burden of transformers at the risk of losing important long-range dependency, which may result in sub-optimal feature expression. Comparatively, downsizing the scale of queries and keys by pooling [15], grouping [22], and dilating [23] can sufficiently preserve long-range dependency since they are uniformly downscaled at equal intervals. However, uniform scaling ignores the difference in importance of query-key pairs and cannot build optimal dependence relationships.

Transformer pruning, a powerful approach to building lightweight transformers by dropping redundant structures or components, can effectively address the above issues. To date, there has not been any study focusing on exploring the potential benefit of transformer pruning for medical image segmentation. In fact, transformer pruning is under-explored even in the natural image domain. Zhu et al. [24] presented the first vision transformer pruning method by evaluating the impact of transformer dimensions and dropping unimportant dimensions accordingly. Pan et al. [25] developed an interpretable module to adaptively filter out the redundant patches. Rao et al. [26] introduced a prediction module to score each patch and then pruned redundant patches hierarchically. Yin et al. [27] reduced the inference cost by automatically minimizing the number of patches. Despite the great results achieved by these approaches, they only focused on the classification/recognition tasks and reduced computational complexity at the cost of minor performance degradation. Compared to classification/recognition tasks which focus on exacting global features of objects, segmentation tasks need to make pixel-wise classification decisions. Consequently, even slightly inaccurate patch pruning can cause dramatic performance degradation. It explains why the above transformer pruning approaches cannot be directly applied to segmentation with satisfactory performance.

In this paper, we, for the first time, propose a transformer pruning framework APFormer for lighter and better medical image segmentation. Specifically, a U-shaped CNN-Transformer hybrid framework is first constructed as the base framework for capturing sufficient local and global information. For the transformer module, three components are developed to make it feasible to train efficient transformers from scratch with fewer rounds: self-regularized self-attention (SSA), Gaussian-prior relative position embedding (GRPE), and adaptive pruning (AP). SSA introduces the well-converged dependency distribution as prior knowledge to establish high-quality dependency relationships. GRPE highlights the nearby patches for each query patch when introducing position information. These two components help transformers converge well without relying on expensive training data and time. Based on this transformer framework, transformer pruning follows a two-stage process. In the first stage, pruning is accomplished by dropping those query patches with high confidence of being background, which can significantly reduce redundant computation in useless queries. As for the second stage, redundant dependencies are dynamically eliminated for each query, which can decrease the perceptual redundancy of each crucial patch. Qualitative and quantitative results compared to the state-of-the-art approaches on two public datasets demonstrate the effectiveness of APFormer for medical image segmentation. More importantly, the proposed transformer pruning method can be extended to other hybrid-/transformer-based methods for computational complexity reduction and performance improvement. The contributions can be summarized as follows:

- APFormer, a lightweight CNN-Transformer hybrid segmentation framework for medical image segmentation, which is more feasible for deployment.
- A two-stage transformer pruning mechanism for simultaneous computational complexity reduction and performance improvement. To our best knowledge, this is the first work on transformer pruning for image segmentation. Furthermore, in contrast to existing transformer pruning methods which inevitably suffer from performance degradation, the proposed pruning mechanism can achieve performance improvement.
A self-regularized strategy for improving self-attention matrices in transformers to build truly-useful long-range dependency faster.

A novel position embedding approach utilizing prior knowledge to better introduce position information with lower training complexity.

### II. Problem Analysis

In this section, we revisit the structure of naive vision transformers and analyze their deployment challenges for medical image segmentation.

#### A. Structure of Naive Vision Transformer

A typical vision transformer layer is composed of a multi-head self-attention (MSA) block and a feed-forward network (FFN) in series as shown in Fig. 1. Given an input feature \( F \in \mathbb{R}^{D \times H \times W} \), it is tokenized and projected into a sequence of flattened patches \( F_p \in \mathbb{R}^{N \times d} \), where \( N \) is the number of patches. Since the position information of \( F \) is corrupted during tokenization, an input-independent learnable absolute position embedding \( P_a \in \mathbb{R}^{N \times d} \) is utilized to compensate for the lost position information in \( F_p \). Specifically, \( F_p + P_a \) is taken as the input of MSA and projected into query \( Q \in \mathbb{R}^{N \times d} \), key \( K \in \mathbb{R}^{N \times d} \), and value \( V \in \mathbb{R}^{N \times d_m} \) to build global dependency, which can be summarized as

\[
A = \text{softmax}(\frac{(F_p + P_a)E_q((F_p + P_a)E_k)^T}{\sqrt{d_m}})
\]

\[
F_{sa} = A(F_p + P_a)E_v
\]

\[
F_{msa} = (F_{sa}( \odot \cdots \odot F_{sa}))W_{msa}(F_p + P_a),
\]

where \( A \in \mathbb{R}^{N \times N} \) is the self-attention matrix, \( h \) is the number of self-attention heads, \( \odot \) and \( \oplus \) are the concatenation and residual connection operations, \( E_q \in \mathbb{R}^{d \times d_m}, E_k \in \mathbb{R}^{d \times d_m} \), and \( E_v \in \mathbb{R}^{d \times d_m} \) are the learnable weight matrices used for projecting the self-attention input \( (F_p + P_a) \) into \( Q, K, \) and \( V \) respectively, and \( W_{msa} \) is the projection matrix for combining the outputs of the \( h \) self-attention heads. Then, the final learned features can be obtained by applying FFN to \( F_{msa} \) by

\[
F_{ffn} = (\max(0, F_{msa}W_1 + b_1) \cdot W_2 + b_2) \oplus F_{msa},
\]

where \( W_1 \) and \( W_2 \) are learnable projection matrices, and \( b_1 \) and \( b_2 \) are the offsets in FFN. The main benefit of vision transformers comes from the construction of \( A \) which builds patch-wise interactions for all to capture long-range, global dependency.

#### B. Challenges of Transformer Deployment for Medical Image Data

The success of transformers in the natural image domain relies on sufficient training time and data. However, when applying transformers to medical image segmentation, severe performance degradation is encountered due to:

- **Expensive training costs in learning self-attention matrices and position embedding:** As presented in Fig. 1, each self-attention matrix \( A \) can be regarded as \( N \) convolution kernels with the same size \( h \times w \) to capture features for each query patch based on global information, where \( h \times w \) is the number of patches. Here, \( N = h \times w = H/s \times W/s \) is determined according to the size of each patch \((s,s)\) (e.g. \( s = 2^0, 2^1, 2^2, 2^3, 2^4 \) in practice). Consequently, the calculation of \( A \) (i.e., \( N \times N \)) is highly expensive (e.g. \( N = 256, 1024 \) in practice), requiring a large amount of data for training. In addition, position embedding, which is introduced for compensating the spatial location information destroyed by tokenizing 2D features into 1D sequences, is a \( N \times d \) matrix, the scale and irregularity of which further exacerbates the data dependence problem. Unfortunately, different from natural images, collecting large-scale well-annotated medical images is both difficult and expensive.

- **Large model size:** The number of parameters in transformers is determined by the number of tokens and the size of each token, both of which are determined by the image resolution. Therefore, the model size grows exponentially with the image resolution.

- **Increased computational cost:** The computational cost of transformers is determined by the number of tokens, the number of self-attention heads, and the depth of the network. Therefore, the computational cost increases exponentially with the image resolution.

- **Data dependence problem:** Transformers rely heavily on sufficient training data and compute the data dependence of different tokens, which is highly nonlinear and difficult to approximate.
medical images is infeasible due to data privacy and high annotation cost, making it challenging to guarantee good convergence of both self-attention and position embedding. Based on the examples of $A$, which are illustrated in the second last column of Fig. 1, the dependency built for the background query patches (e.g., the rows framed in purple) is quite similar to that of the foreground patches (e.g., the rows framed in red) within each of the four self-attention matrices, indicating the poor performance of $A$, far from achieving the objective of building truly-important/-specific dependency for each query patch.

- **Severe computation and dependency redundancy in self-attention:** Transformers are known for their outstanding ability to build long-range dependency. However, as presented in Fig. 1, each patch would be taken as a query and then build dependency with all patches. Though building long-range dependency is beneficial, involving all patches could be highly redundant and would bring non-negligible noise due to perception redundancy. For example, as shown in the last column of Fig. 1, building long-range dependency for the green/purple patches (i.e., easy-to-identify background patches) is worthless, resulting in high computational redundancy. Meanwhile, for the patches in yellow/red, interactions with the patches highlighted in red via self-attention are beneficial, while the information from other patches can be redundant and even noisy.

Based on the above analysis, with the proper assistance of well-converged self-attention matrices and position embedding, pruning redundant dependency in transformers can be beneficial even for performance improvement in addition to computation cost reduction.

### III. Method

APFormer is constructed by first building a pruning-friendly CNN-Transformer hybrid architecture and then performing adaptive pruning. Details are provided in the following.

#### A. Overview

APFormer is designed as a simple U-shaped structure with pruned transformer blocks as the bridge between encoder and decoder as depicted in Fig. 2. Given a 2D input image $X \in \mathbb{R}^{H \times W}$ or $\mathbb{R}^{3 \times H \times W}$ with the resolution of $H \times W$, the CNNs encoder extracts shallow and fine features to capture short-range dependency and builds inner inductive biases. Then, the pruned transformer modules model long-term interactions with lower computation and perception redundancy. Finally, the decoder performs pixel-wise segmentation.

#### B. Construction of SSA and GRPE

The key to lossless pruning is to effectively identify and remove redundancy while preserving all useful information. Consequently, before pruning, we propose self-regularized self-attention (SSA) and Gaussian-prior relative positional embedding (GRPE) to achieve better convergence and to better recognize redundancy for lower training cost.
on the inherent properties of well-converged self-attention matrices, according to [49] and [51], we name this training strategy as self-regularized self-attention (SSA). Among the two inherent constraints, one is on the diagonal symmetry of $A$, defined as

$$S(A, A^T) = \sum_{i=0}^{N} \sum_{j=0}^{N} (A_{i,j} \times A_{i,j}^T),$$

(3)

where $A^T$ is the transpose of $A$, $\times$ is numerical multiplication, and $\| \|_2$ is the operation of calculating quadratic normal form. Then, the symmetry loss $L_{sym}$ is defined as

$$L_{sym} = Max(1 - S(A, A^T) - \alpha_{sym}, 0),$$

(4)

where $\alpha_{sym}$ is a smoothing factor, as pursuing $A$ and $A^T$ being identical can be counter-productive. The other constraint is on the entropy of $A$, where its entropy is obtained by

$$E(A_i) = -\sum_{j=0}^{N} (A_{i,j} \times \log_2(A_{i,j})).$$

(5)

To pursue low entropy, a entropy loss $L_{en}$ is defined as

$$L_{en} = Max(\min(E(A_i)) - \alpha_{en}, 0),$$

(6)

where $\alpha_{en}$ is a smoothing factor to release the penalty. Without $\alpha_{en}$, the attention matrix would just focus on very few patches. Finally, each self-regularized self-attention head in APFormer is trained by

$$L_{ssa} = \beta_1 L_{sym} + \beta_2 L_{en},$$

(7)

where $\beta_1$ and $\beta_2$ are balancing hyper-parameters set as 0.8 and 0.2 respectively in our experiments. Through $L_{ssa}$, APFormer is encouraged to converge to better dependency establishment.

2) Gaussian-Prior Relative Position Embedding: Inspired by the keypoint heatmaps in pose estimation [28], the prior importance of each key patch to the query patch can be expressed as a Gaussian distribution. As shown in Fig. 5, given a $32 \times 32$ feature map, position $(10, 10)$ as the query patch and all positions as the key patches, the closer the key patch is to the query patch, the more important it is to the query patch. Taking each position of the feature map as a query patch orderly, the complete Gaussian heatmap with the same resolution of the self-attention matrix is as shown in Fig. 5, formulated as

$$G_{i,j} = e^{-\frac{(j-10)^2+(i-10)^2}{2\sigma^2}},$$

(8)
where \((h, w)\) is the resolution of the feature map, \(\theta\) is a learnable parameter to control the range of important regions, \(\odot\) is the modulus operation, and \(\%\) is the remainder operation. Then, the generated \(G \in \mathbb{R}^{N \times N}\) can be regarded as the Gaussian prior knowledge of relative position information. Given that it only reflects the distance relationship of relative positions while ignoring the direction information, an objective learnable relative position embedding \(R \in \mathbb{R}^{4N}\) is introduced. The final GRPE is written as

\[
P_{i,j} = G_{i,j} + R_{ij},
\]

where \(R_{ij} = R(2w(i \odot w - j \odot w + h) + (i\%w - j\%w + w))\).

Compared to the number of position embedding parameters in naive vision transformer (i.e., \(N \times d\), generally \(d = 64/128/256/1024\)), GRPE (i.e., \(N \times 4 + h\), where \(h\) is the number of self-attention heads and generally equals to \(6/12\)) is much lighter. With the re-designed position embedding, the output of self-attention in Eq. 1 is rewritten as

\[
F_{sa} = \text{softmax}(\frac{F_p E_q F_p E_k}{\sqrt{d_m}} + P) F_p E_v.
\]

C. Adaptive Pruning

To pursue lightweight and efficient transformers for medical image segmentation, we propose a two-stage transformer pruning method, including query-wise pruning and dependency-wise pruning. In query-wise pruning, the easy-to-identify background patches are pruned before projecting to queries as building long-range dependency for them is worthless. Specifically, we introduce a learnable gate for each candidate query patch calculated by

\[
G_b(q_{ij}) = \frac{e^{W_b F_{i,j}}}{e^{W_b F_{i,j}} + e^{W_f F_{i,j}}},
\]

where \(W_b\) and \(W_f\) are learnable projection matrices trained by a cross-entropy loss \(L_g\) between \(G = G_b \odot G_f\) and the ground-true mask. Then, only the query patches with low gate scores \(F_p^g \in \mathbb{R}^{(1-\alpha)N \times d}\) (i.e., excluding the easy-to-identify background patches) would join long-term interactions with the key patches, leading to smaller self-attention matrices \(A_p \in \mathbb{R}^{(1-\alpha)N \times N}\) calculated by

\[
A_p = \text{softmax}(\frac{F_p^g E_q F_p E_k}{\sqrt{d_m}} + P),
\]

where \(\alpha\) is the query-wise pruning rate which is adaptive depending on inputs.

After capturing long-range dependency for the kept query patches, we conduct dependence-wise pruning based on adaptive thresholds \(T\). For each query \(q_i\), its pruning threshold \(T_i\) is generated based on the dependency distribution and its inner features, formulated as

\[
T_i = \min(A_{p1}) + \frac{(\max(A_{p1}) - \min(A_{p1}))}{(1 + e^{-W_i F_p E_q})(1 + e^{-g})},
\]

where \(W_i\) is the projection matrix and \(g\) is a learnable control parameter to avoid exorbitant \(T_i\) at the beginning of training.

According to the adaptively generated \(T_i\), a binary decision mask \(M \in \{0, 1\}^{(1-\alpha)N \times N}\) is used to determine whether to prune the dependency or not. Specifically, all elements in \(M\) are first initialized to 1 and then updated to 0 when \(A_{p1,i,j} < T_i\) is satisfied. To make the sum of all dependency scores of each query 1, the attention matrix \(A_p\) is updated according to

\[
A_{p1,i,j} = \frac{M_{i,j} e^{A_{p1,i,j}}}{10^{-6} + \sum_k M_{i,k} e^{A_{p1,k}}}.
\]

Through the above two-stage pruning, compared to the naive self-attention module whose computational complexity is

\[
\Omega_{sa} = 3Ndd_m + 2N^2d_m + Nd_m^2,
\]

the computational complexity of the pruned self-attention module in APFormer becomes

\[
\Omega_{psa} = (3 - \alpha)Ndd_m + \alpha(2 - \lambda)N^2d_m + \alphaNd_m^2,
\]

where \(\alpha \in (0, 1)\) and \(\lambda \in (0, 1)\) are the query-wise pruning rate and the dependency-wise pruning rate respectively. It should be noted that both \(\alpha\) and \(\lambda\) are adaptively determined not pre-defined.

IV. Evaluation

A. Datasets

1) ISIC 2018: This dataset is for skin lesion segmentation [41], [42], consisting of 2596 images with pixel-level annotations. Due to the lack of official or commonly-used data split strategies, three-fold cross-validation is employed for evaluation.

2) Synapse: This dataset involves 30 cases of abdominal CT scans and each CT slice is annotated with 13 organs. Following the setting of [9] and the following transformer-related works [22], [35], [36], [37], [38], [39] on the Synapse dataset, eight out of 13 organs are used for evaluation, and 18 cases are selected for training while the rest are built as the test set.

B. Implementation Details

All learning frameworks are implemented based on PyTorch and trained for 400 rounds by an Adam optimizer with an initial learning rate of \(10^{-4}\) and a batch size of 4. The parameter \(g\) is initialized as \(-2\) for stable pruning. For data augmentation, methods including contrast adjustment, gamma augmentation, random rotation, and scaling are adopted.

C. Evaluation on ISIC 2018

1) Learning Frameworks for Comparison: Four CNN-based methods including Att-UNet [8], CENet [30], CANet [30], and CPFNet [32], and three transformer-based approaches including SETR-PUP [5], TransUnet [9], and FAT-Net [29], have been included for comparison. All experimental results of the comparison methods are re-implemented according to the released source codes. The results of statistical testing between comparison methods and APFormer are reported as p-values.
**TABLE I**

| Model Type | Method         | Dice (%) | IoU (%) | ACC (%) | SE (%) | SP (%) | P (M) | P (G) | FPS | p-value |
|------------|----------------|----------|---------|---------|--------|--------|-------|-------|-----|---------|
| CNNs       | Att-UNet [8]   | 87.75±0.38 | 80.33±0.74 | 93.13±0.37 | 89.98±0.23 | 97.07±0.34 | 35    | 67    | 109 | <0.001  |
|            | CENet [30]     | 88.99±0.50 | 82.05±0.62 | 95.48±0.21 | 90.66±0.63 | 96.64±0.65 | 29    | 7     | 22  | <0.001  |
|            | CAnet [31]     | 89.03±0.30 | 82.01±0.40 | 95.79±0.15 | 90.24±0.69 | 97.12±0.40 | 2.8   | 6     | 47  | <0.001  |
|            | CPNeFNet [32]  | 89.13±0.31 | 82.19±0.38 | 95.56±0.27 | 89.98±0.75 | 97.04±0.15 | 43    | 8     | 66  | <0.001  |
| Hybrid/Trans | SETR-PUP [5]   | 88.10±0.37 | 80.74±0.74 | 95.16±0.15 | 90.21±1.57 | 96.69±0.52 | 77    | 79    | 51  | <0.001  |
|            | TransUnet [9]  | 89.18±0.25 | 82.40±0.34 | 95.77±0.32 | 90.09±0.56 | 97.21±0.61 | 105   | 32    | 39  | <0.001  |
|            | FAT-Net [29]   | 88.27±0.45 | 81.11±0.47 | 95.23±0.21 | 89.36±1.30 | 97.38±0.76 | 34    | 31    | 41  | <0.001  |
|            | APFormer       | 90.07±0.37 | 83.47±0.40 | 96.06±0.10 | 90.99±0.87 | 97.05±0.12 | 2.6   | 4.1   | 47  | -       |

1) **Learning Frameworks for Comparison**: CNN-based approaches including R50 U-Net [9], ResUNet [7], Att-UNet [8], SETR-PUP [5], TransUnet [9], FAT-Net [29] and APFormer respectively, and the ground truth.

2) **Quantitative Results**: According to the quantitative comparison results summarized in Table I, the state-of-the-art transformer-based approaches achieve competitive performance to the well-designed CNN-based approaches in skin lesion segmentation. Among these state-of-the-art approaches, CPFNet and TransUnet achieve the best performance with the average Dice score of 89.13% and 89.18% respectively. Compared to CPFNet, APFormer consistently achieves better performance across all evaluation metrics with lower computational complexity. Compared to TransUnet, despite a slightly lower SP, APFormer achieves greater performance for all other metrics including Dice, IoU, ACC, and SE by an average increase of 0.89%, 1.07%, 0.29%, and 0.9% respectively. All p-values of the methods selected for comparison are smaller than 0.001, indicating that the performance improvement achieved by APFormer is statistically extremely significant. Quantitative efficiency analysis of various approaches measured in FPS (i.e., frames per second) is summarized in Table I. In general, Att-UNet achieves the highest FPS outperforming others. Among the transformer-based methods, APFormer achieves a comparable FPS with noticeable performance improvements, validating its value in deployment.

3) **Qualitative Results**: Qualitative skin lesion segmentation results of various approaches including U-Net [6], ResUNet [7], Att-UNet [8], SETR-PUP [5], TransUnet [9], FAT-Net [29], and APFormer are presented in Fig. 6. As shown in the first two rows, both CNN-based and hybrid/transformer-based approaches suffer from local contextual similarity due to either limited receptive fields or superabundant dependencies. Comparatively, APFormer achieves the best segmentation performance based on the truly-valuable long-range dependency after pruning.

**D. Evaluation on Synapse**

1) **Learning Frameworks for Comparison**: CNN-based approaches including R50 U-Net [9], R50 Att-UNet [9], U-Net [6] and Att-UNet [8], and transformer-based methods including TransUnet [9], Swin-UNet [35],
TABLE II

| Type        | Method                   | Avg. | Aorta | Gall | L-Kid | R-Kid | Liver | Pancreas | Spleen | Stomach | P  | FPS  | p-value |
|-------------|--------------------------|------|-------|------|-------|-------|-------|----------|--------|---------|----|------|---------|
| CNNs        | R30 U-Net [9]            | 74.68| 87.74 | 63.66| 80.60 | 78.19 | 93.74 | 56.90    | 85.87  | 74.16   | 144| 221 | 59      |
|             | R30 Att-U-Net [6]        | 75.57| 55.92 | 63.91| 79.20 | 72.71 | 93.56 | 49.37    | 87.39  | 74.95   | 348| 700 | 44      |
|             | U-Net [6]                | 76.85| 89.07 | 69.72| 77.77 | 68.60 | 93.43 | 53.98    | 86.67  | 75.58   | 35 | 67   | 109     |
|             | Att-U-Net [8]            | 77.77| 89.55 | 68.88| 77.98 | 71.11 | 93.57 | 58.04    | 87.30  | 75.75   | 35 | 67   | 109     |
| Hybrid/Trans| TransUnet [9]            | 77.48| 87.23 | 63.13| 81.87 | 77.02 | 94.08 | 55.86    | 85.08  | 75.62   | 105| 32   | 39      |
|             | Swin-UNet [35]           | 79.13| 85.47 | 66.53| 83.28 | 79.61 | 94.29 | 56.58    | 90.66  | 76.60   | 41 | 11   | 49      |
|             | TransClaw U-Net [39]     | 78.09| 85.87 | 61.38| 84.83 | 79.36 | 94.28 | 57.65    | 87.74  | 73.55   | -  | -    | -       |
|             | LeVit-Unet-384 [36]      | 78.53| 87.33 | 62.23| 84.61 | 80.25 | 93.11 | 59.07    | 88.86  | 72.76   | 22 | 22   | 79      |
|             | MT-UNet [37]             | 78.59| 87.92 | 64.99| 81.47 | 77.29 | 93.06 | 59.46    | 87.75  | 76.81   | 79 | 40   | 9       |
|             | MISSFormer [22]          | 81.96| 86.99 | 68.65| 85.21 | 82.00 | 94.41 | 65.67    | 91.92  | 80.81   | 42 | 7.2  | 3.002   |
|             | CA-GANFormer [38]        | 82.55| 89.05 | 67.48| 86.05 | 82.17 | 95.61 | 67.49    | 91.00  | 81.55   | -  | -    | -       |
|             | APFormer                 | 83.53| 90.84 | 64.36| 90.54 | 85.99 | 94.93 | 72.16    | 91.88  | 77.55   | 2.6| 3.9  | 52      |

Fig. 7. Qualitative comparison results on the Synapse dataset. From left to right: the raw images, the segmentation results produced by U-Net [6], ResUNet [7], Att-U-Net [8], SETR-PUP [5], TransUnet [9], FAT-Net [29] and the proposed APFormer respectively, and the ground truth.

TransClaw U-Net [39], LeVit-Unet-384 [36], MT-UNet [37], MISSFormer [22], and CA-GANFormer [38] are used for comparison. All quantitative results are from the published papers. Methods with open-source code are re-implemented for calculating parameters, GFLOPs, FPS and p-value.

2) Quantitative Results: Quantitative comparison results of different methods on the Synapse dataset are summarized in Table II. Compared to the CNN-based methods which struggle to capture long-range dependency, the transformer-based methods generally achieve better segmentation results at the cost of a slower inference speed. Among the approaches selected for comparison, CA-GANFormer achieves the best performance with an average of 82.55% in Dice and MISSFormer has the lowest computational complexity. Compared to the state-of-the-art transformer-based methods, APFormer achieves the best overall performance with the lowest computational complexity, the fewest parameters, and the second-fastest inference speed. In terms of segmentation performance for individual organs, despite the deficiency in gallbladder segmentation, APFormer achieves the best performance for the segmentation of aorta, left kidney, right kidney, and pancreas (i.e., four of eight organs), the second best performance for segmenting liver and spleen, and the third best segmentation performance for the stomach. Furthermore, most p-values between the comparison methods and APFormer are smaller than 0.001, indicating that APFormer’s performance improvement is statistically extremely significant. It should be noted that, though the p-value compared to MISSFormer...
Fig. 8: Visual results of self-attention matrix changes by applying transformer pruning in SETR [5], TransUnet [9], and APFormer on the Synapse and ISIC 2018 datasets.

is slightly greater than 0.001, \( p = 0.002 \) still satisfies the definition of “statistically significant” (i.e., \( p < 0.05 \) or \( p < 0.005 \) in practice). As the number of patients/scans for the Synapse dataset is relatively limited for evaluation, compared to the ISIC 2018 dataset, it is more likely to make prediction results somewhat similar and result in a higher \( p \)-value given the same performance improvement. More importantly, APFormer achieves better segmentation performance on five out of eight organs with much better model efficiency.

To evaluate the deployment capability, we measure the efficiency of each approach in FPS (i.e., frames per second) as summarized in Table II. In general, APFormer achieves the second highest FPS, lower than LeViT-Unet-384, outperforming all other transformer-based methods. Compared to CNN-based methods, U-Net and Att-UNet achieve the highest FPS and surpass other approaches by a large margin, demonstrating their lightweight design. It should be noted that the main motivation of APFormer is to reduce redundancy in transformers to make them more trainable with relatively insufficient data, instead of pursuing real-time inference. In addition, we prove, through experimental comparison, pruning to reduce training complexity is beneficial to improve the performance ceiling of transformers. It explains why we emphasize more on performance improvement for the segmentation problem.

3) Qualitative Results: Qualitative segmentation results of different methods on the Synapse dataset are shown in Fig. 7. As analyzed in [9] and [10], the CNN-based methods tend to make segmentation errors in local similarity regions due to their inner inductive biases. Theoretically, the hybrid-/transformer-based methods can overcome this issue by modeling long-range dependency. However, perception redundancy as analyzed in Sec. II would negatively affect the segmentation performance. It explains why transformer pruning is beneficial.

V. DISCUSSION

In this section, a series of ablation studies and discussions are provided for a comprehensive evaluation of APFormer.

A. Generalization of Transformer Pruning

1) Generalization to Other Hybrid-/Transformer-Based Segmentation Approaches: Based on the availability of source code, SETR [5] and TransUnet [9] were re-implemented for adaptive pruning. For SETR, only dependency-wise pruning is applied as it directly feeds raw images to transformers, making it impossible to generate learnable gate scores for query-wise pruning. For TransUnet, we apply both query-wise and dependency-wise pruning.

Quantitative comparison results of different methods before and after pruning are summarized in Table III. With dependency-wise pruning, the segmentation performance of SETR can be improved by 1.2% and 0.89% respectively in Dice on the Synapse and the ISIC 2018 datasets. By introducing both query-wise and dependency-wise pruning, on the
Synapse and the ISIC 2018 datasets, the computational complexity of TransUnet can be reduced by 25.07% and 10.88%, and its segmentation performance is improved by an average increase of 1.33% and 0.64% respectively in Dice. In addition, $p < 0.001$ is satisfied in most cases, validating the value and architecture independence of adaptive pruning.

To analyze the effect of adaptive pruning, we visualize the changes in self-attention matrices before and after pruning. As shown in Fig. 8, in the raw attention matrices, though for training on small-scale datasets. Comparatively, GRPE separately are provided in Fig. 10. The main contribution of each component in APFormer, namely SSA, GRPE, and AP, are separately introduced to the same backbone network and trained on the ISIC 2018 dataset for evaluation. As shown in Table V, the p-value of each component’s sole contribution to performance improvement is $< 0.01$, indicating the effectiveness of each component. To illustrate the effectiveness of SSA, we visualized the attention matrices of the same baseline model with and without SSA. As depicted in Fig. 9, by introducing SSA, the self-attention matrices can escape from equal dependency distributions and compactly focus on the relevant regions. Furthermore, compared with directly applying AP, introducing SSA and GRPE can not only improve the segmentation performance but also increase the pruning effect from 0.46% to 0.76% in improvement to Dice, being consistent with the analysis in Sec. II.

Training curves of the baseline network with SSA and GRPE separately are provided in Fig. 10. The main contribution of SSA is to speed up convergence, which is quite useful for training on small-scale datasets. Comparatively, GRPE provides more useful position information, leading to better convergence performance and segmentation results.

2) Ablation Studies of $\alpha_{\text{sym}}$ and $\alpha_{\text{en}}$ in APFormer: Quantitative results on the Synapse dataset are summarized in Table III. The three components in APFormer, namely SSA, GRPE, and AP, are separately introduced to the same backbone network and trained on the ISIC 2018 dataset for evaluation. As shown in Table V, the p-value of each component’s sole contribution to performance improvement is $< 0.01$, indicating the effectiveness of each component. To illustrate the effectiveness of SSA, we visualized the attention matrices of the same baseline model with and without SSA. As depicted in Fig. 9, by introducing SSA, the self-attention matrices can escape from equal dependency distributions and compactly focus on the relevant regions. Furthermore, compared with directly applying AP, introducing SSA and GRPE can not only improve the segmentation performance but also increase the pruning effect from 0.46% to 0.76% in improvement to Dice, being consistent with the analysis in Sec. II.

Training curves of the baseline network with SSA and GRPE separately are provided in Fig. 10. The main contribution of SSA is to speed up convergence, which is quite useful for training on small-scale datasets. Comparatively, GRPE provides more useful position information, leading to better convergence performance and segmentation results.

2) Ablation Studies of $\alpha_{\text{sym}}$ and $\alpha_{\text{en}}$ in APFormer: Quantitative results on the Synapse dataset are summarized in Table III. The three components in APFormer, namely SSA, GRPE, and AP, are separately introduced to the same backbone network and trained on the ISIC 2018 dataset for evaluation. As shown in Table V, the p-value of each component’s sole contribution to performance improvement is $< 0.01$, indicating the effectiveness of each component. To illustrate the effectiveness of SSA, we visualized the attention matrices of the same baseline model with and without SSA. As depicted in Fig. 9, by introducing SSA, the self-attention matrices can escape from equal dependency distributions and compactly focus on the relevant regions. Furthermore, compared with directly applying AP, introducing SSA and GRPE can not only improve the segmentation performance but also increase the pruning effect from 0.46% to 0.76% in improvement to Dice, being consistent with the analysis in Sec. II.

Quantitative results of different approaches before and after pruning are summarized in Table IV. With AP, the classification accuracy of ConViT and BoTNet can be improved by 2.07% and 3.11% respectively with little computational complexity reduction. Comparatively, the computational complexity of ViT is reduced by 49.92% and its classification accuracy is improved by 1.04%. One interesting observation is that, with AP, CNN-Transformer hybrid methods achieve greater performance improvement while pure transformer-based methods achieve greater complexity reduction, validating the effectiveness of AP in removing irrelevant query patches and redundant dependency.

B. Ablation Study

1) Effectiveness of Each Component in APFormer: The three components in APFormer, namely SSA, GRPE, and AP, are separately introduced to the same backbone network and trained on the ISIC 2018 dataset for evaluation. As shown in Table V, the p-value of each component’s sole contribution to performance improvement is $< 0.01$, indicating the effectiveness of each component. To illustrate the effectiveness of SSA, we visualized the attention matrices of the same baseline model with and without SSA. As depicted in Fig. 9, by introducing SSA, the self-attention matrices can escape from equal dependency distributions and compactly focus on the relevant regions. Furthermore, compared with directly applying AP, introducing SSA and GRPE can not only improve the segmentation performance but also increase the pruning effect from 0.46% to 0.76% in improvement to Dice, being consistent with the analysis in Sec. II.

Training curves of the baseline network with SSA and GRPE separately are provided in Fig. 10. The main contribution of SSA is to speed up convergence, which is quite useful for training on small-scale datasets. Comparatively, GRPE provides more useful position information, leading to better convergence performance and segmentation results.
TABLE V
ABBLATION STUDY ON DIFFERENT COMPONENT COMBINATIONS OF APFormer ON THE ISIC 2018 DATASET. APFormer- REPRESENTS APFormer WITHOUT ADAPTIVE PRUNING. “BASELINE” IS THE BASELINE MODEL TO CALCULATE THE P-VALUES

| Method          | Dice (%) | IoU (%) | ACC (%) | SE (%) | SP (%) | p-value |
|-----------------|----------|---------|---------|--------|--------|---------|
| baseline        | 88.37±0.45 | 81.08±0.76 | 95.61±0.31 | 89.73±2.38 | 96.94±0.81 |         |
| + SSA           | 88.78±0.31 | 81.70±0.39 | 95.80±0.24 | 89.41±0.76 | 97.35±0.08 | 0.001   |
| + GRPE          | 88.97±0.40 | 81.93±0.50 | 95.69±0.18 | 90.70±0.07 | 96.88±0.36 | <0.001  |
| + AP            | 88.73±0.50 | 81.62±0.73 | 95.60±0.35 | 90.91±1.85 | 96.73±1.09 | 0.006   |
| APFormer-       | 89.31±0.51 | 82.37±0.68 | 95.77±0.20 | 89.57±0.60 | 97.37±0.11 | <0.001  |
| APFormer        | 90.07±0.37 | 83.47±0.40 | 96.01±0.10 | 90.99±0.87 | 97.05±0.12 | <0.001  |

C. V.s. State-of-the-Art Transformer Pruning Methods
Three vision transformer pruning methods including VTP [24], AdaViT [27], and DViT [26] are combined with APFormer- (i.e., the backbone network of APFormer without pruning) for comparison. Quantitative comparison results of different transformer pruning methods on the Synapse dataset are summarized in Table VIII. Among those pruning methods, AdaViT fails to improve the efficiency of APFormer- with noticeable performance degradation. Though DViT achieves the most significant complexity reduction in FLOPs, it is at the cost of introducing the most parameters. VTP reduces both the computational cost and the parameters but suffers from severe performance degradation. In summary, all these transformer pruning methods which were proposed for the classification problem encounter severe performance degradation for the segmentation problem, which is consistent with the analysis in Sec. I. Comparatively, APFormer is superior for both performance improvement and computational complexity reduction.

D. Comparison Between 2D and 3D Architectures
It should be noted that APFormer is designed in a 2D manner without exploring spatial information across slices, resulting in under-performance compared to the state-of-the-art 3D algorithms on 3D datasets. To make a fair comparison between 2D and 3D models, we evaluate and compare their respective performance in terms of segmentation accuracy and model efficiency as illustrated in Figs. 11 and 12.

Tables VI and VII. With smaller αsym and αen, the network would enforce self-attention matrices A to be completely symmetric with the lowest entropy, resulting in poor feature diversity. Comparatively, using larger αsym and αen would be more likely to build better representations but struggle to produce well-converged self-attention matrices under insufficient training data. In our experiments, setting αsym = 0.3 and αen = 0.4 achieves the best performance on the Synapse dataset. In other segmentation tasks, the two parameters are suggested to be tuned in ascending order.
Fig. 11. Organ-level performance comparison on the Synapse dataset. The dashed and solid lines represent the 3D and 2D models respectively. Furthermore, circle, triangle, square, and rhombus markers indicate the CNN-based methods, the transformer-based methods requiring training for at least 1000 epochs, the transformer-based methods requiring pre-training, and the normal transformer-based methods, respectively.

Fig. 12. Efficiency comparison in terms of parameters and computational complexity on the Synapse dataset. Circles filled with grid, horizontal lines, slashes, and solid colors indicate the 3D transformer-based, the 3D CNN-based, the 2D CNN-based, and the 2D transformer-based methods respectively. Parameters represents the number of model parameters and GFLOPs is the abbreviation of floating-point operations per second measured in billions (G). The larger the circle, the better the corresponding model. The closer the circle is to the bottom-left corner, the more efficient the model is.

In terms of model efficiency, 3D models [23], [33], [34], [40] generally incur higher computational complexity, have more parameters, and require more training rounds than most 2D models. As for segmentation performance, all models except R50 U-Net follow a similar Dice trend across organs. Specifically, pancreas, gallbladder, right kidney, and stomach are the four most challenging organs to segment. One interesting observation is that most performance improvement achieved by the 3D models comes from these four organs.

To better understand why the 3D models are superior, especially for segmenting these four organs, we visualize the inputs and the corresponding annotations of several sequential examples in Fig. 13. The shape and intensity of both stomach and pancreas can vary greatly across individuals. In addition, suffering from low contrast, accurate boundary detection of stomach and pancreas is challenging. As for gallbladder and right kidney, the interference from neighboring regions with similar features makes the segmentation task challenging. In clinical applications, experts would rely on neighboring slices to achieve accurate pixel-wise annotation, which is a 3D matter. This explains why 3D models are more likely to achieve better segmentation performance with the assistance of spatial information which is only available in 3D input.

E. Limitation and Future Work

Despite the impressive success of APFormer, its SSA and AP components still encounter limitations. In terms of SSA, three constraints are adopted to promote the convergence of transformers and avoid attention collapse, laying a solid foundation for the adaptive pruning operation. However, these constraints may also limit the diversity of self-attention matrices. In terms of AP, the redundancy in the heads and dimensions of transformers is ignored in this study, which may hinder the upper bound of lightweight. In our future work, we will explore additional, as well as alternative, effective self-regularized constraints for SSA and a more comprehensive pruning program including all transformer components.

VI. Conclusion

In this paper, we propose a lightweight yet effective hybrid CNN-Transformer framework APFormer for medical image segmentation from the perspective of transformer pruning. Specifically, self-regularized attention is proposed to boost the convergence of transformers which is helpful to alleviate the data dependence problem. Then, Gaussian-prior relative position embedding is designed to effectively complement the position information to transformers. After that, we perform
query-wise and dependency-wise pruning sequentially for not only redundancy reduction but also performance improvement. Experimental results on two public datasets demonstrate the superior performance of APFormer compared to the state-of-the-art methods with significantly lower computational complexity. Furthermore, we demonstrate, through extensive evaluation, that adaptive pruning is architecture-independent and applicable to other hybrid-/transformer-based frameworks for performance improvement. We believe the idea of transformer pruning would inspire future work on developing ultra-light and even better-performing transformers for medical image segmentation.

REFERENCES

[1] A. Vaswani et al., “Attention is all you need,” in Proc. NeurIPS, 2017, pp. 5998–6008.
[2] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” 2020, arXiv:2010.11929.
[3] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” 2021, arXiv:2103.14030.
[4] N. Carion et al., “End-to-end object detection with transformers,” in Proc. ECCV, 2021, pp. 213–229.
[5] S. Zheng et al., “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,” in Proc. CVPR, Jun. 2021, pp. 6881–6890.
[6] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. MICCIAI, 2015, pp. 234–241.
[7] X. Xiao, S. Lian, Z. Luo, and S. Li, “Weighted res-UNet for high-quality retina vessel segmentation,” in Proc. 9th Int. Conf. Inf. Technol. Med. Educ. (ITME), Oct. 2018, pp. 327–331.
[8] J. Schlemper et al., “Attention gated networks: Learning to leverage salient regions in medical images,” Med. Image Anal., vol. 53, pp. 197–207, Apr. 2019.
[9] J. Chen et al., “TransUNet: Transformers make strong encoders for medical image segmentation,” 2021, arXiv:2102.04306.
[10] Y. Zhang, H. Liu, and Q. Hu, “TransFuse: Fusing transformers and CNNs for medical image segmentation,” 2021, arXiv:2102.08005.
[11] D. Karimi, S. Vasylychko, and A. Gholipour, “Convolution-free medical image segmentation using transformers,” 2021, arXiv:2102.13645.
[12] W. Wang, C. Chen, M. Ding, J. Li, H. Yu, and S. Zha, “TransBTS: Multimodal brain tumor segmentation using transformers,” 2021, arXiv:2103.04430.
[13] B. Chen, Y. Liu, Z. Zhang, G. Lu, and A. Wai Kin Kong, “TransAttUNet: Multi-level attention-guided U-Net with transformer for medical image segmentation,” 2021, arXiv:2107.03527.
[14] Y. Xie, J. Zhang, C. Shen, and Y. Xia, “CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation,” 2021, arXiv:2103.03024.
[15] Y. Gao, M. Zhou, and D. N. Metaxas, “UTNet: A hybrid transformer architecture for medical image segmentation,” in Proc. MICCIAI, 2021, pp. 61–71.
[16] J. Maria Jose Valanarasu, P. Oza, I. Hachhalioglu, and V. M. Patel, “Medical transformer: Gated axial-attention for medical image segmentation,” 2021, arXiv:2102.10662.
[17] J. Wang, L. Wei, L. Wang, Q. Zhou, L. Zhu, and J. Qin, “Boundary-aware transformers for skin lesion segmentation,” in Proc. MICCIAI, 2021, pp. 206–216.
[18] Y. Gao, M. Zhou, D. Liu, Z. Yan, S. Zhang, and D. N. Metaxas, “A data-scalable transformer for medical image segmentation: Architecture, model efficiency, and benchmark,” 2022, arXiv:2203.00131.
[19] Y. Li, W. Cai, Y. Gao, C. Li, and X. Hu, “More than encoder: Introducing transformer decoder to upsampling,” 2021, arXiv:2106.10387.
[20] A. Lin, B. Chen, J. Xu, Z. Zhang, and G. Lu, “DS-TransUNet: Dual swin transformer U-Net for medical image segmentation,” 2021, arXiv:2106.07616.
[21] Y. Ji et al., “Multi-compound transformer for accurate biomedical image segmentation,” in Proc. MICCIAI, 2021, pp. 326–336.
[22] X. Huang, Z. Deng, D. Li, and X. Yuan, “MISSFormer: An efficient medical image segmentation transformer,” 2021, arXiv:2109.07162.
[23] Y. Wu et al., “D-former: A U-shaped dilated transformer for 3D medical image segmentation,” 2022, arXiv:2201.00462.