

# Lightweight Structure-Aware Attention for Visual Understanding

Heeseung Kwon<sup>1</sup>, Francisco M. Castro<sup>2</sup>, Manuel J. Marin-Jimenez<sup>3</sup>, Nicolas Guil<sup>2</sup>,  
Karteek Alahari<sup>1\*</sup>

<sup>1</sup>Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK.

<sup>2</sup>Department of Computer Architecture, University of Málaga.

<sup>3</sup>Department of Computing and Numerical Analysis, University of Córdoba.

\*Corresponding author(s). E-mail(s): [karteek.alahari@inria.fr](mailto:karteek.alahari@inria.fr);

Contributing authors: [heeseung.kwon@inria.fr](mailto:heeseung.kwon@inria.fr); [fcastro@uma.es](mailto:fcastro@uma.es); [mjmarin@uco.es](mailto:mjmarin@uco.es);  
[nguill@uma.es](mailto:nguill@uma.es);

## Abstract

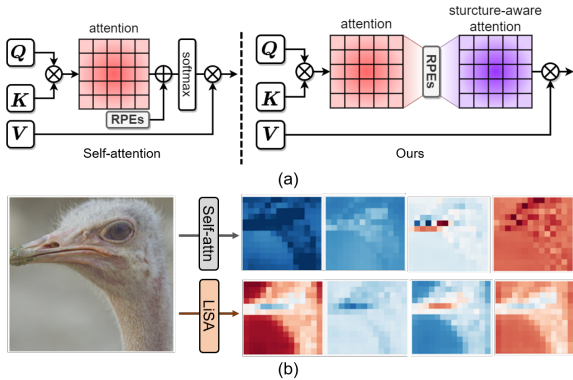
Attention operator has been widely used as a basic brick in visual understanding since it provides some flexibility through its adjustable kernels. However, this operator suffers from inherent limitations: (1) the attention kernel is not discriminative enough, resulting in high redundancy, and (2) the complexity in computation and memory is quadratic in the sequence length. In this paper, we propose a novel attention operator, called Lightweight Structure-aware Attention (LiSA), which has a better representation power with log-linear complexity. Our operator transforms the attention kernels to be more discriminative by learning structural patterns. These structural patterns are encoded by exploiting a set of relative position embeddings (RPEs) as multiplicative weights, thereby improving the representation power of the attention kernels. Additionally, the RPEs are approximated to obtain log-linear complexity. Our experiments and analyses demonstrate that the proposed operator outperforms self-attention and other existing operators, achieving state-of-the-art results on ImageNet-1K and other downstream tasks such as video action recognition on Kinetics-400, object detection & instance segmentation on COCO, and semantic segmentation on ADE-20K.

**Keywords:** Visual Understanding, Vision Transformer, Self-Attention, Image Recognition

## 1 Introduction

Since the emergence of the vision transformer (ViT) [1], transformers have become the dominant neural architecture for visual understanding, outperforming convolutional neural networks (CNNs). Self-attention, a core operator of ViT, has relative merits compared to convolution because of the adjustable attention kernel and its ability to capture long-range dependencies. However,

self-attention has inherent limitations for visual recognition. First, the attention kernel has difficulty learning discriminative features due to the lack of desirable inductive biases, resulting in high redundancy of the ViT layers [2, 3]. Thus, it usually requires a large amount of data [1] and aggressive augmentations [4] to obtain good performance. Second, the complexity of self-attention is quadratic in



**Fig. 1: Self-attention vs. LiSA.** (a) Process of self-attention & LiSA: LiSA updates the attention to the structure-aware attention via RPEs. (b) Feature visualization of self-attention & LiSA: compared to self-attention, LiSA learns better features by capturing geometric structural patterns.

the length of its input sequence, making the operator impractical for high-resolution images and difficult to adopt for hierarchical models.

Several approaches have proposed new types of operators to address the limitations of self-attention. Some of them have attempted to learn better discriminative features with self-attention by including relative position embeddings (RPEs) [5–7] or capturing geometric structures (*e.g.*, image gradients, video motion) [8, 9]. However, these operators still have high computational complexity, which makes it challenging to capture long-range dependencies [6–9]. Some other methods have proposed efficient attention operators to handle the complexity of self-attention [10–15]. Although these operators have a linear complexity with a factorized attention kernel, they often underperform, compared to the original attention [12, 15]. Recent methods have proposed convolutional attention operators by integrating attention with convolution [2, 16–20]. While these operators are effective for capturing geometric structures with convolutions, their learnability is still limited since convolution kernels are local and static.

In this paper, we propose an effective yet efficient operator, *lightweight structure-aware attention (LiSA)*. To address the limitations of self-attention, we focus on improving the attention kernel to be more discriminative. **By leveraging the fact that the feature correlation contains rich structural information [9, 21–24], we devise a**

**new attention operator that learns structural patterns within the query-key correlation via RPEs.**

As illustrated in Fig. 1a, while existing methods adopt RPEs for additive interaction, we exploit RPEs as multiplicative weights. The RPEs extract structural patterns from the attention kernel and recombine the kernel in a structure-aware manner. Fig. 1b illustrates a few sample feature maps from the early layers of self-attention and LiSA. LiSA effectively captures geometric structures in the image, while self-attention features are relatively weak and uninformative due to the lack of desirable inductive biases. Finally, since the complexity of RPEs is quadratic in the sequence length, we compute them efficiently with fast Fourier transforms (FFTs), achieving log-linear complexity.

Our main contributions are as follows: (1) we overcome the limitations of self-attention by proposing a new attention operator called LiSA, which learns structural patterns with log-linear complexity, and, (2) LiSANets, the models based on our LiSA operator, outperform their counterparts, achieving state-of-the-art results on visual understanding benchmarks such as ImageNet-1K [25], Kinetics-400 [26], COCO [27], and ADE-20K [28].

## 2 Related Work

**ViTs for visual understanding.** After the success of ViT [1], transformer architectures have been widely adopted in a variety of visual understanding tasks [3, 29–33]. Several approaches have proposed improvements to the original ViT [1], *e.g.*, using a teacher-student scheme [4], a better tokenization scheme [3], or using small splits of the tokens to obtain richer local information [34]. Recently, several approaches employ the hierarchical structure by adopting efficient local attention techniques [6, 35, 36]. However, their representation powers are still low due to the limitations of self-attention. In this paper, we propose a new hierarchical ViT model family, LiSANets, achieving state-of-the-art performance with less computation due to the high expressivity and efficiency of LiSA.

**Highly-expressive operators.** Recently proposed operators increase the representation power by developing self-attention [5–9, 37] or convolution [38–41]. Attention-based operators have achieved this by adding relative position embeddings [5, 6, 37] or capturing relational structures [7–

9]. Convolution-based operators have dynamically adapted convolution kernels based on the input features [38–41]. However, these highly-expressive operators require high computational complexity and are typically limited to local interactions [7–9, 39]. One example is the relational self-attention (RSA) [9], which is related to our work. RSA is one of the most expressive operators that captures structural patterns with relational components, but it is also limited to local interactions due to its high computational complexity. In contrast, our proposed LiSA shows the highest level of expressivity by capturing long-range structural patterns with log-linear complexity.

**Lightweight operators.** Some of the existing lightweight attention operators factorize the softmax attention kernel [10, 11, 42, 43]. While they have linear complexity, they usually perform worse than the original attention in terms of accuracy [12, 15]. Other approaches have attempted to linearize RPE-added attention operators [13–15], but they still underperform on visual recognition. Recently, a few approaches adopt FFTs for efficiently covering global receptive fields [15, 44, 45]. The Global Filter (GF) layer [44] is one such operator, which implements an efficient global circular convolution with FFTs. **However, the representation power of the GF layer is constrained since its static kernels hinder adaptation to general visual concepts. Our LiSA also adopts FFTs to improve efficiency, but it focuses on learning structural patterns with its dynamic attention kernels, leading to better performance.**

**Convolutional attention operators.** Recent approaches try to combine transformers and CNNs to leverage the best of each world. Some of them incorporate convolutions into the attention operators [2, 16, 17, 19, 20, 46] to increase the expressivity of self-attention. They often apply depthwise convolutions before computing attention to capture structural information. Inception Mixer [19] splits the input channels and processes convolution and attention in parallel to increase the representation power. While these operators are more expressive than the original attention, they highly depend on static convolution kernels for learning discriminative features. In contrast to these, our LiSA captures discriminative features by dynamic structure-aware attention kernels, which are beneficial for learning richer visual concepts.

### 3 Background

**Self-attention.** The self-attention operator [47] is a core component in transformer architectures that generates query-key attention for updating the value. Let  $N$  denote the sequence length (the number of tokens) and  $C$  the number of input channels. Given an input feature  $\mathbf{X} \in \mathbb{R}^{N \times C}$ , query, key, value,  $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N \times C}$ , are first produced by independent linear projections, and each element of the output  $\mathbf{Y} \in \mathbb{R}^{N \times C}$  of self-attention is expressed as

$$Y_{i,k} = \sum_j^N \sigma(A_{i,j}) V_{j,k}, \quad A_{i,j} = \frac{1}{\sqrt{C}} \sum_k^C Q_{i,k} K_{j,k}. \quad (1)$$

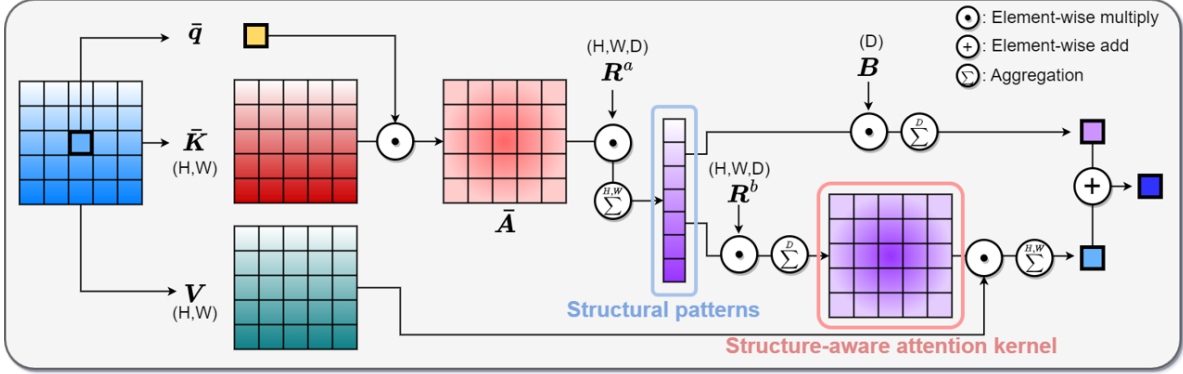
Note that  $\sigma$  is the softmax function along the  $j$ -axis. The two main characteristics of self-attention are that: (1) the operator represents a global interaction where the size of the attention kernel for each query is equal to  $N$ , and (2) the attention kernel dynamically changes according to the input feature. However, it is unable to encode the relative order of tokens due to the lack of convolutional inductive biases [37], resulting in performance degradation on visual recognition.

**Relative position embedding (RPE).** One of the popular schemes to handle the lack of convolutional inductive biases is adopting an RPE for the self-attention operator [5, 6, 37]. A common RPE has the shape of a Toeplitz matrix, and it consists of learnable weights which can be expressed as

$$\mathcal{T}(\mathbf{e}) = \begin{pmatrix} e_N & e_{N+1} & e_{N+2} & \cdots & e_{2N-1} \\ e_{N-1} & e_N & e_{N+1} & \cdots & e_{2N-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_1 & e_2 & e_3 & \cdots & e_N \end{pmatrix}, \quad (2)$$

where  $\mathbf{e} = \{e_1, e_2, \dots, e_{2N-1}\}$ . When RPE is added, the attention operator is formulated as

$$Y_{i,k} = \sum_j^N \sigma(A_{i,j} + R_{i,j}) V_{j,k}, \quad \mathbf{R} = \mathcal{T}(\mathbf{e}) \in \mathbb{R}^{N \times N}. \quad (3)$$



**Fig. 2: Computational graph of Structure-aware Attention (SA) for each query.** After obtaining the query-key dot-product correlation ( $\bar{A}$ ), structural patterns in  $\bar{A}$  are encoded by  $R^a$ , and utilized in two ways: 1) the patterns are used for generating a structure-aware attention kernel with  $R^b$ , and 2) directly projected as a structural feature with  $B$ . Note that  $N = H \times W, C = 1$  in this figure.

By introducing relative positional information into attention, the self-attention operator obtains the ability to learn convolutional inductive biases.

#### Limitations of self-attention with RPE.

Despite several approaches showing the effectiveness of RPE, the attention operator still has some limitations. First, the expressivity of the operator is insufficient; it is difficult to capture geometric structures (*e.g.*, image gradients, video motion) since the softmax attention kernel may not be effective for encoding gradient information due to its non-negativity [9, 48]. Second, although the attention  $A$  suppresses photometric variations and reveals geometric structures [21, 23], the kernel is aggregated with the value  $V$  without leveraging structural patterns within  $A$ . Third, the operator suffers from quadratic complexity ( $\mathcal{O}(N^2)$ ) since the non-linear softmax function and RPE are hard to linearize. Although a few approaches [10–12, 15] have attempted to approximate the softmax function with kernelized methods to make the operator more efficient, they do not improve over the original transformer in accuracy due to its training instability [15] or approximation errors [12].

## 4 Structure-aware Attention

### 4.1 Basic Form of Structure-aware Attention (SA)

**Learning convolutional inductive biases.** To handle the limitations of self-attention, we devise

a new attention operator that leverages the advantages of convolution. Unlike the conventional usage of an RPE (Eq. 3), we employ it as multiplicative weights for learning relative token orders as follows:

$$Y_{i,k} = \sum_j^N \bar{A}_{i,j} R_{i,j} V_{j,k}, \quad \bar{A}_{i,j} = \sum_k^C \bar{Q}_{i,k} \bar{K}_{j,k}. \quad (4)$$

Note that  $\bar{Q}, \bar{K}$  are L2-normalized query and key, respectively. In Eq. 4, the RPE  $R$  not only learns relative token orders, but also actively adjusts the attention values. We remove the softmax function to allow the attention kernel to include negative values, which may be effective for encoding structural information. Instead, the query and key are L2-normalized to stabilize the training procedure [15]. Since the matrix multiplication between the Toeplitz matrix  $R$  and the value  $V$  is equivalent to a global convolution [49] that applies the convolution kernel  $e \in \mathbb{R}^{2N-1}$  for the value  $V$ , the operator can also be interpreted as a dynamic global convolution where the dynamic component of the convolution kernel is based on the attention  $\bar{A}$ . Thus, the proposed operator merges the characteristics of self-attention and convolution.

**Learning structural patterns.** Nevertheless, the above operator (Eq. 4) is still limited for capturing rich structural patterns within the attention  $\bar{A}$ . The RPE  $R$  is only element-wise multiplied with  $\bar{A}$ , but it cannot extract meaningful features within the query-key correlation. To handle this,

we directly extract structural patterns from  $\bar{A}$  and regenerate the attention kernel using multiple RPEs. The updated operator is formulated as:

$$Y_{i,k} = \sum_j \sum_d \sum_n^N (\bar{A}_{i,n} R_{i,n,d}^a) (R_{i,j,d}^b V_{j,k} + B_{k,d}) \quad (5)$$

where  $\mathbf{R}^a = \{\mathcal{T}(\mathbf{e}_1^a), \mathcal{T}(\mathbf{e}_2^a), \dots, \mathcal{T}(\mathbf{e}_D^a)\}$ ,  $\mathbf{R}^b = \{\mathcal{T}(\mathbf{e}_1^b), \mathcal{T}(\mathbf{e}_2^b), \dots, \mathcal{T}(\mathbf{e}_D^b)\} \in \mathbb{R}^{N \times N \times D}$  are RPE tensors composed of sets of Toeplitz matrices and  $\mathbf{B} \in \mathbb{R}^{C \times D}$  is a learnable projection matrix, respectively. Note that  $D$  is the size of structural patterns. The computational graph of Eq. 5 is illustrated in Fig. 2. For each query, the learnable RPE tensor  $\mathbf{R}^a$  captures structural patterns by encoding the  $N$ -size attention kernel as a  $D$ -size vector. We utilize this vector in two ways: first, to generate a new structure-aware attention kernel along the  $j$ -axis using the RPE tensor  $\mathbf{R}^b$ ; second, to project it as a feature representation with the learnable matrix  $\mathbf{B}$ . In summary, the generated attention kernel updates  $\mathbf{V}$  in a structure-aware manner, and the projected feature represents the encoded structural patterns.

Note that our method differs from convolutional attention operators [2, 16, 17, 19, 20] in the way of learning discriminative features. Convolutional attention operators process convolution and attention separately in a sequential [2, 16, 17, 20] or a parallel [19] way. These operators rely on convolution to obtain structural information from the input feature; however, their ability is limited because the convolution kernels are static and have restricted receptive fields. They cannot leverage rich structural patterns within the query-key correlation. In contrast to these operators, our structure-aware attention obtains structural information from the query-key correlation and has global receptive fields.

## 4.2 Improving the Expressivity of SA

We can further improve its expressivity by exploiting semantic information of the input channels. Here we describe the advanced form of our structure-aware attention.

**Capturing channel-wise structural patterns.** To exploit the semantics of the input

operator	computation	memory
self-attention [47]	$\mathcal{O}(N^2C)$	$\mathcal{O}(N^2 + NC)$
structure-aware attention (Eq. 4)	$\mathcal{O}(N^2C)$	$\mathcal{O}(N^2 + NC)$
structure-aware attention (Eq. 5)	$\mathcal{O}(N^2CD)$	$\mathcal{O}(NCD)$
LiSA (FFT approximation)	$\mathcal{O}(NCD \log_2 N)$	$\mathcal{O}(NCD)$

**Table 1: Comparison of complexity of the operators.**  $N, C, D$  denote the sequence length, the number of channels, and the size of structural patterns, respectively. Our operator has log-linear and linear complexity in computation and memory, respectively.

channels, we employ a different type of query-key correlation, the Hadamard-product correlation. A few approaches [8, 9] have demonstrated that Hadamard-product correlation is more effective than the dot-product one due to the use of richer query-key semantics. Considering the Hadamard correlation is a 3-dimensional tensor  $\bar{A}_{i,n,c} = \bar{Q}_{i,c} \bar{K}_{n,c} \in \mathbb{R}^{N \times N \times C}$ , we expand the RPE tensor  $\mathbf{R}^a$  by  $C$  channels for encoding the Hadamard correlation. The modified operator is formulated as follows:

$$Y_{i,k} = \sum_c \sum_d \sum_{j,n}^N (\bar{A}_{i,n,c} \tilde{R}_{i,n,c,d}^a) (R_{i,j,d}^b V_{j,k} + B_{k,d}). \quad (6)$$

Note that  $\tilde{\mathbf{R}}^a \in \mathbb{R}^{N \times N \times C \times D}$  is the expanded RPE tensor and the number of learnable weights increases from  $\mathbf{E}^a = \{\mathbf{e}_1^a, \mathbf{e}_2^a, \dots, \mathbf{e}_D^a\} \in \mathbb{R}^{(2N-1) \times D}$  to  $\tilde{\mathbf{E}}^a = \{\mathbf{e}_1^a, \mathbf{e}_2^a, \dots, \mathbf{e}_{CD}^a\} \in \mathbb{R}^{(2N-1) \times C \times D}$ . In Eq. 6, for each query, the expanded tensor  $\tilde{\mathbf{R}}^a$  captures channel-wise structural patterns by encoding an  $N \times C$  Hadamard correlation matrix as a  $D$ -size vector. Since this process does not require additional computation, the operator can exploit rich semantics through the Hadamard correlation by only increasing the number of parameters.

## 4.3 Lightweight Structure-aware Attention (LiSA)

Although our operator (SA) is highly-expressive, it is not easy to apply in neural architectures due to its high computational complexity. Here, we describe its final form, LiSA, which significantly reduces the complexity by efficiently processing the heavy RPE tensors through FFTs.

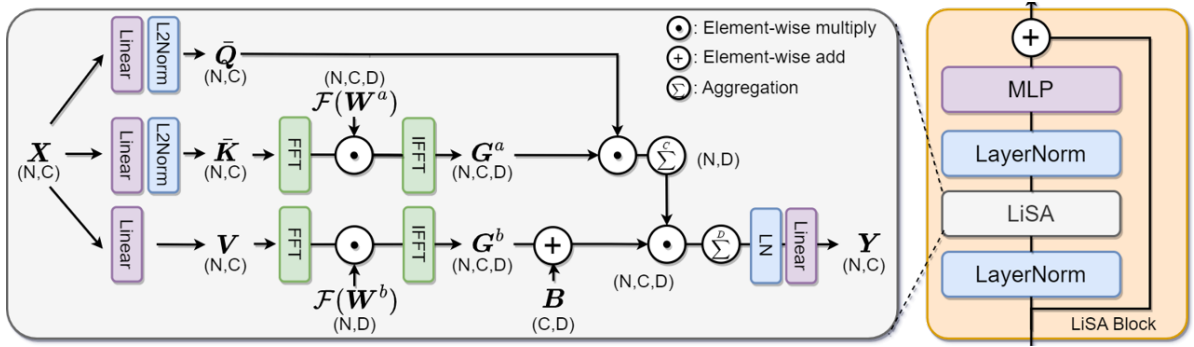


Fig. 3: Computational graph of LiSA and its block configuration. See text for details.

**Approximating RPEs with FFTs.** Unlike the original attention (Eq. 3), the multiplicative RPEs ( $\tilde{R}^a, R^b$ ) and the absence of softmax enable the permutation of the computation order, and thus, the rearranged equation is expressed as:

$$\begin{aligned} Y_{i,k} &= \sum_c \bar{Q}_{i,c} \sum_d \sum_{j,n} \bar{K}_{n,c} \tilde{R}_{i,n,c,d}^a (R_{i,j,d}^b V_{j,k} + B_{k,d}), \\ &= \sum_c \bar{Q}_{i,c} \sum_d G_{i,c,d}^a (G_{i,k,d}^b + B_{k,d}), \end{aligned} \quad (7)$$

where  $G^a = \bar{K} * \tilde{E}^a$ ,  $G^b = V * E^b \in \mathbb{R}^{N \times C \times D}$ .

Note that  $E^b = \{e_1^b, e_2^b, \dots, e_D^b\} \in \mathbb{R}^{(2N-1) \times D}$  are learnable weights of  $R^b$ , and  $*$  denotes convolution.  $G^b$  is a global convolution that applies the global kernels  $E^b \in \mathbb{R}^{(2N-1) \times D}$  to the value  $V$  by sharing the kernels across  $C$  channels, and  $G^a$  is a depth-wise global convolution that applies the global kernels  $\tilde{E}^a \in \mathbb{R}^{(2N-1) \times C \times D}$  to the key  $\bar{K}$ . In Eq. 8, improving the efficiency of SA has shifted to efficiently processing global convolutions.

To reduce the complexity of global convolutions, we approximate them as *global circular convolutions* [44], indicating that the RPEs are replaced by circular position embeddings consisting of circulant matrices. These circular convolutions can be efficiently computed by FFTs via the convolution theorem of Fourier transform: *multiplication in the frequency domain is equal to circular convolution in the time domain*. Thus, we can rewrite the equation as follows:

$$G^a \approx \bar{K} \circledast W^a = \mathcal{F}^{-1}(\mathcal{F}(\bar{K}) \odot \mathcal{F}(W^a)), \quad (8)$$

$$G^b \approx V \circledast W^b = \mathcal{F}^{-1}(\mathcal{F}(V) \odot \mathcal{F}(W^b)). \quad (9)$$

Note that  $W^a = \{w_1^a, \dots, w_{CD}^a\} \in \mathbb{R}^{N \times C \times D}$ ,  $W^b = \{w_1^b, \dots, w_D^b\} \in \mathbb{R}^{N \times D}$  are learnable weights of the circular convolutions and  $\circledast, \odot, \mathcal{F}, \mathcal{F}^{-1}$  denote circular convolution, element-wise multiplication, FFT, and IFFT, respectively. The circular convolutions have half the size of parameters since the kernel size reduces from  $2N-1$  to  $N$ . As shown in Tab. 1, we reduce the complexity in computation and memory on a log-linear scale by approximating heavy global interactions with FFTs. The computational graph of LiSA is illustrated in Fig. 3. Moreover, we further reduce the complexity by half using RFFT and inverse RFFT. Implementing Eq. 9 using standard GPU libraries has an IO bottleneck in throughput since it reads and writes the intermediate results repeatedly. We adopt the kernel fusion strategy [50] to overcome this by fusing the entire computation of Eq. 9 into a single kernel and computing it in SRAM.

## 5 Experiments

We first describe the implementation details and then present extensive results. This includes a set of comprehensive analyses and a state-of-the-art comparison on ImageNet-1K [25] and ImageNet-21k [51]. Finally, we also verify the effectiveness of LiSA on video action recognition with Kinetics-400 [26], object detection with COCO [27], and semantic segmentation with ADE-20K [28].

### 5.1 Implementation details

**LiSA block.** Our proposed block follows the traditional transformers sequence of layers [1, 4, 31]: layer normalization (LN), attention operator, LN

Model	#Blocks	#Channels (#heads)
LiSANet-I	12	192 (12)
LiSANet-S	[2, 4, 12, 4]	[64 (4), 128 (8), 320 (20), 384 (24)]
LiSANet-B	[4, 8, 18, 3]	[96 (6), 192 (12), 384 (24), 576 (36)]
HyLiSANet-S	[3, 6, 12, 4]	[64, 128, 320 (20), 384 (24)]
HyLiSANet-B	[3, 12, 18, 3]	[96, 192, 384 (24), 576 (36)]

**Table 2:** Detailed configurations of different variants of LiSANet. For hierarchical models, we provide the number of channels and blocks in 4 stages.

and MLP. Instead of using a traditional attention operator, we use LiSA. The overall block configuration is shown in Fig. 3.

**LiSANet.** To demonstrate the effectiveness of LiSA, we define three different variant architectures as shown in Tab. 2. The first variant is an *isotropic model* (LiSANet-I), which has no down-sampling layers and fixes the number of tokens ( $14 \times 14$ ) at all depths. The second is *hierarchical ViTs* (LiSANet-S, LiSANet-B) consisting of LiSA blocks. All of the hierarchical ViTs are composed of 4 stages with a different number of blocks, and the number of tokens is downsampled in each stage. And, the third is *hybrid ViTs* (HyLiSANet-S, HyLiSANet-B) following the strategy of recent hybrid models [2, 19, 37]. We adopt depthwise convolutions for the early two stages, which allows for constructing deeper layers for high-resolution stages. The structural pattern size  $D$  is set to 16 for LiSANet-I and 8 for the other models. Like many previous backbones [16, 18, 52–55], we adopt convolutional position embedding [56] and convolutional MLP [18] for our models. All the details of our variants are summarized in Sec.A.1. of the supplementary material.

## 5.2 LiSA Analysis

**Setup.** We use the isotropic model (LiSANet-I) for analyses since the operators with quadratic complexity are hard to be adopted for hierarchical models due to their extreme memory consumption. LiSANet-I is trained for 150 epochs on ImageNet1K, and we follow the rest of the training recipes suggested in [4, 6] for a fair comparison. Unless specified otherwise, we use  $224 \times 224$  resolution for input.

**Comparison with other operators.** In Tab. 3a, we compare our LiSA operator with

several others, including self-attention [5, 47], convolution [44, 57], and the other expressive operators [7, 9, 16, 19]. For a fair comparison, we only replace our operator with those others in the LiSA blocks, and all the receptive fields are set as global, except for the  $7 \times 7$  depthwise convolution [57]. LiSA substantially outperforms self-attention with and without RPE [5] (first two rows) in accuracy, showing the impact of learning structural patterns. The accuracy of LiSA is even 1.5% higher than the self-attention with larger channels (3<sup>rd</sup> row), indicating that the gain does not come from the increased parameters. **LiSA also performs better than the GF layer [44] (5, 6<sup>th</sup> rows), which is equal to a global circular convolution.** This implies that the adjustable attention kernel of LiSA is a better fit for learning various visual concepts rather than the static convolution kernel.

For the convolutional attention operator (8, 9<sup>th</sup> rows), we adopt multi-scale attention [16], one of the advanced operators. The operator applies a  $3 \times 3$  depthwise convolutions for the input features before computing attention, and the stride is set to  $1 \times 1$ . Inception mixer [19] splits input channels and processes convolution and attention in parallel for learning both high and low-frequency features, and the stride is set to  $1 \times 1$ . To demonstrate fair comparisons between these advanced attention operators with ours, we match the number of parameters and FLOPs in a similar scale by increasing the number of channels (9, 10<sup>th</sup> rows). Convolutional attention and Inception mixer perform better than self-attention in accuracy, but their accuracies are lower than that of LiSA, even with larger FLOPs and number of parameters. We conjecture that LiSA learns better discriminative features by capturing structural patterns inside the attention. RSA [9] shows high accuracy by learning structural patterns via its relational components, but it requires a significantly large computation budget due to the larger correlation. In contrast, LiSA shows the best trade-off between accuracy and FLOPs, achieving the best accuracy among the operators with lower FLOPs.

**Effectiveness of LiSA components.** In Tab. 3b, we provide a detailed analyses of the components of LiSA. We first compare how convolutional inductive biases are learned in self-attention by our (Eq. 4) parameters. For the same FLOPs and the number of parameters, our approach

index	operator	FLOPs	#params	top-1	top-5
1	Self-attn [47]	1.25 G	5.72 M	71.0	90.0
2	Self-attn w/ RPE [5]	1.25 G	5.72 M	72.2	90.9
3	Self-attn w/ RPE ( $C \uparrow$ ) [5]	1.40 G	6.44 M	73.4	91.6
4	Depthwise conv ( $7 \times 7$ ) [57]	0.84 G	4.49 M	69.0	89.2
5	GF layer [44]	0.82 G	4.90 M	69.5	89.4
6	GF layer ( $C \uparrow$ ) [44]	1.27 G	7.37 M	72.4	91.0
7	Lambda convolution [7]	2.41 G	5.41 M	72.6	91.0
8	Convolutional attn [16]	1.26 G	5.78 M	73.3	91.7
9	Convolutional attn ( $C \uparrow$ ) [16]	1.41 G	6.49 M	74.1	92.1
10	Inception mixer ( $C \uparrow$ ) [19]	1.28 G	6.36 M	74.4	92.1
11	RSA [9]	5.34 G	8.23 M	74.5	92.2
12	LiSA (ours)	1.21 G	6.36 M	<b>74.9</b>	<b>92.4</b>

(a) Comparison with other operators.

index	$D$	FLOPs	#params	top-1	top-5
1	1	1.11 G	5.76 M	71.3	90.5
2	4	1.13 G	5.88 M	73.6	91.6
3	8	1.17 G	6.04 M	74.4	92.2
4	16	1.21 G	6.36 M	<b>74.9</b>	<b>92.4</b>

(c) Effect of the structural patterns ( $D$ ).

index	operator	FLOPs	#params	top-1	top-5
1	Self-attn	1.25 G	5.72 M	71.0	90.0
2	Self-attn w/ RPE	1.25 G	5.72 M	72.2	90.9
3	Self-attn w/ $h$ RPEs	1.25 G	5.81 M	72.7	91.3
4	SA (Eq. 4)	1.25 G	5.72 M	72.4	91.2
5	SA (Eq. 4 w/ $h$ RPEs)	1.25 G	5.81 M	73.6	91.7
6	SA (Eq. 5)	3.92 G	5.99 M	74.2	92.1
7	+ Hadamard corr	3.92 G	8.09 M	74.9	92.3
8	LiSA (ours)	1.21 G	6.36 M	<b>74.9</b>	<b>92.4</b>

(b) Effectiveness of LiSA components.

index	kernel size	FLOPs	#params	top-1	top-5
1	Local - $3 \times 3$	1.22 G	5.75 M	71.9	90.7
2	Local - $5 \times 5$	1.45 G	5.80 M	73.8	91.7
3	Local - $7 \times 7$	1.80 G	5.88 M	73.8	91.8
4	Global (ours)	1.21 G	6.36 M	<b>74.9</b>	<b>92.4</b>

(d) Effect of global interactions.

**Table 3: LiSA analysis on ImageNet-1K.** Top-1, top-5 accuracy (%), FLOPs (G) and the number of parameters (M) are shown.

block	$(H, W) = (7, 7)$			$(H, W) = (14, 14)$			$(H, W) = (28, 28)$			$(H, W) = (56, 56)$			$(H, W) = (84, 84)$		
	FLOPs (M) $\downarrow$	mem (M) $\downarrow$	latency (ms) $\downarrow$	FLOPs (M) $\downarrow$	mem (M) $\downarrow$	latency (ms) $\downarrow$	FLOPs (M) $\downarrow$	mem (M) $\downarrow$	latency (ms) $\downarrow$	FLOPs (G) $\downarrow$	mem (G) $\downarrow$	latency (ms) $\downarrow$	FLOPs (G) $\downarrow$	mem (G) $\downarrow$	latency (ms) $\downarrow$
Self-attn [47]	22.7	<b>18.1</b>	<b>0.9</b>	102.0	21.7	1.6	584.0	282.4	13.0	5.2	14.7	174.7	22.2	OOM	OOM
Convolutional attn [16]	23.1	<b>18.1</b>	1.7	104.0	23.9	2.7	413.0	135.4	8.5	1.6	2.6	30.5	3.7	3.4	68.0
LiSA (ours)	<b>21.8</b>	18.3	1.2	<b>87.3</b>	<b>20.3</b>	<b>1.3</b>	<b>349.0</b>	<b>87.4</b>	<b>4.6</b>	<b>1.4</b>	<b>0.8</b>	<b>19.8</b>	<b>3.1</b>	<b>1.7</b>	<b>45.0</b>

**Table 4: Comparisons among LiSA & attention operators in FLOPs, memory, and latency.** The memory & latency is measured by an RTX A5000 (batch=32, channels=192). OOM is an abbreviation of out-of-memory.

that uses an RPE as multiplicative weights (“SA (Eq. 4)”, 4<sup>th</sup> row in the table) is better than standard self-attention with RPE (2<sup>nd</sup> row) in accuracy. The accuracy gap between self-attention and ours becomes more clear when we use independent learnable weights (RPE) for each attention head (3<sup>rd</sup> vs. 5<sup>th</sup> row), thus showing that our RPEs defined in Eq. 4 are more beneficial than those of standard self-attention. This indicates that our attention containing negative values is potentially more effective for learning spatial features such as gradient information compared to softmax attention. Next, we demonstrate our structure-aware attention variants in the third part of the table. Comparing Eq. 4 with Eq. 5 (4<sup>th</sup> vs. 6<sup>th</sup> row), we validate the effectiveness of learning structural patterns, which improves by 1.8% in top-1 accuracy. With the Hadamard correlation, our operator improves

the top-1 accuracy by 1.0% without any additional FLOPs. Lastly, LiSA with FFT approximation dramatically reduces the computational cost (3 $\times$ ) without compromising accuracy.

**Effect of the structural patterns  $D$ .** Our  $D$  parameter represents the size of the encoded vector that learns structural patterns from the query-key correlation. Thus, larger values are expected to result in better performances. Tab. 3c shows the impact of the structural patterns by varying the values of  $D$ . As shown in Tab. 3c, the size of  $D$  significantly impacts the accuracy, which implies the importance of structural patterns. The accuracy is indeed improved with larger values of  $D$ , 16 being the best one. Compared to the results of Tab. 3a, the case of  $D = 8$  already surpasses all the other operators in accuracy except for RSA. Note that

type	model	FLOPs	#params	top-1
CNN	ResNet-50 [58]	4.1 G	26 M	76.1
	RegNetY-4.0GF [59]	4.0 G	21 M	80.0
	GFNet-H-S [44]	4.6 G	32 M	81.5
	ConvNext-T [60]	4.5 G	29 M	82.1
	InternImage-T [54]	5.0 G	30 M	83.5
ViT	PVT-S [61]	3.8 G	25 M	79.8
	Deit-S [4]	4.6 G	22 M	79.9
	Swin-Ti [6]	4.5 G	29 M	81.2
	T2T-ViT-14 [3]	4.8 G	22 M	81.5
	CSwin-T [36]	4.3 G	23 M	82.7
	LiSANet-S (ours)	3.9 G	25 M	<b>83.4</b>
Hybrid	CvT-13 [17]	4.5 G	20 M	81.6
	CoAtNet-0 [37]	4.2 G	25 M	81.6
	MViTv2-T [16]	4.7 G	24 M	82.3
	iFormer-S [19]	4.8 G	20 M	83.4
	Slide-PVTv2-B2 [55]	4.2 G	23 M	82.7
	UniFormer-S [53]	3.6 G	22 M	82.9
	SMT-S [52]	4.7 G	21 M	<b>83.7</b>
	HyLiSANet-S (ours)	4.5 G	24 M	<b>83.7</b>
CNN	ResNet-101 [58]	7.9 G	45 M	77.4
	RegNetY-8.0GF [59]	8.0 G	39 M	81.7
	GFNet-H-B [44]	8.6 G	54 M	82.9
	ConvNext-B [60]	15.4 G	89 M	83.8
	InternImage-S [54]	8.0 G	50 M	84.2
ViT	PVT-L [61]	9.8 G	61 M	81.7
	Deit-B [4]	17.5 G	86 M	81.8
	T2T-ViT-24 [3]	13.8 G	64 M	82.3
	Swin-B [6]	15.4 G	88 M	83.5
	CSwin-B [36]	15.0 G	78 M	84.2
	LiSANet-B (ours)	10.4 G	51 M	<b>84.6</b>
Hybrid				
	UniFormer-B [53]	8.3 G	50 M	83.9
	CoAtNet-2 [37]	15.7 G	75 M	84.1
	VAN-B4 [62]	12.2 G	60 M	84.2
	Slide-Swin-B [55]	15.5 G	89 M	84.2
	MViTv2-B [16]	10.2 G	52 M	84.4
	MaxViT-S [20]	11.7 G	69 M	84.5
	iFormer-B [19]	9.4 G	48 M	84.6
	SMT-L [52]	17.7 G	81 M	84.6
	Slide-CSwin-B [55]	15.0 G	78 M	84.7
	iFormer-L [19]	14.0 G	87 M	84.8
	HyLiSANet-B (ours)	11.7 G	50 M	<b>85.0</b>

**Table 5: Comparison to the state-of-the-art models on ImageNet-1K.** FLOPs (G), the number of parameters (M), top-1 accuracy (%) on the ImageNet validation set are shown. All the models use  $224 \times 224$  resolution images.

the cases over  $D = 16$  are not reported since the accuracy becomes saturated.

**Effect of global interactions.** Tab. 3d studies the influence of local and global LiSA kernels. Focusing on the local LiSA kernels (first three rows in the table), the larger the kernel, the larger the number of parameters, FLOPs, and accuracies. This shows that a big kernel with more trainable parameters produces better results but at an increased computational cost. However, our global version (last row) allows a larger number of parameters with smaller FLOPs due to FFT, resulting in the best performance among all the variants.

**Efficiency of the LiSA block.** In Tab. 4, we demonstrate the efficiency of LiSA in terms of FLOPs, memory consumption, and latency. We

compare our LiSA block with a standard attention block and measure the performance of a single block ( $C = 192$ ) by varying the number of tokens. We also compare the LiSA block with a convolutional attention block [16], which downsamples the query and key to  $14 \times 14$  before computing attention by depthwise convolutions with multiple strides. This attention block is more efficient than the standard one on high-resolutions (*e.g.*,  $28 \times 28$ ,  $56 \times 56$ ), but it becomes less efficient on low-resolutions due to added depthwise convolutions. In comparison, the efficiency values of LiSA increase gracefully due to the log-linear complexity, and thus LiSA achieves the best computational performance except for  $7 \times 7$ , where LiSA is slightly slower due to a higher number of sequential operations. Additional details are presented in the supplementary material.

### 5.3 Image classification

**ImageNet-1K.** We train our hierarchical models (LiSANet-S, LiSANet-B, HyLiSANet-S, HyLiSANet-B) for 300 epochs, and follow the rest of the training recipes suggested in [4, 6] for a fair comparison. In Tab. 5, we compare our hierarchical models with state-of-the-art approaches on ImageNet-1K, including CNNs [44, 58, 59], ViTs [3, 4, 6, 36, 61], and hybrid models containing both convolution and self-attention [16, 17, 19, 20, 37, 52–55]. In the top half of the table, we present the results of small models with comparable FLOPs and number of parameters. Our pure ViT model, LiSANet-S, clearly outperforms all the other ViTs in terms of accuracy and FLOPs, demonstrating the effectiveness of LiSA. Compared with other hybrid models, our HyLiSANet-S achieves better or similar results than other approaches with comparable FLOPs and number of parameters.

In the case of larger models, grouped in the bottom half of the table, again our pure ViT model, LiSANet-B, obtains the best accuracy with the lowest FLOPs and number of parameters, showing the benefits of our LiSA operator. Focusing on the hybrid models, our proposed HyLiSANet achieves the best accuracy (85.0%) with much lower FLOPs and number of parameters than other approaches. Thus, our proposed models achieve state-of-the-art results, especially for large models, improving traditional ViTs and more complex models that rely on additional techniques such as

model	img size	FLOPs	#params	top-1
ConvNext-T [60]	224×224	4.5 G	29 M	82.9
ViT-B/16 [1]	384×384	55.4 G	88 M	84.0
ConvNext-S [60]	224×224	8.7 G	50 M	84.6
Swin-B [6]	224×224	15.4 G	88 M	85.2
ConvNext-B [60]	224×224	15.4 G	89 M	85.8
CSwin-B [36]	224×224	15.0 G	78 M	85.9
HyLiSANet-S	224×224	4.5 G	24 M	84.3
HyLiSANet-B	224×224	11.7 G	50 M	<b>86.2</b>

**Table 6: Comparison with other models pre-trained on ImageNet-21K and fine-tuned on ImageNet-1K.** Top-1 accuracy (%), FLOPs (G), and the number of parameters (M) are shown.

window attention (Swin [6]), depthwise convolutions plus self-attention (MVITv2 [16], SMT [52]) or processing convolution and attention in parallel (iFormer [19]). This indicates an essential difference between convolution and LiSA. Convolution-based models converge fast and are data-efficient by their strong inductive biases such as 2D locality, but their performance is restricted when the models are scaled up since convolution may hinder adaptation to general visual concepts due to its static kernels. Whereas, our model achieves great performance even with a larger scale since LiSA can adaptively aggregate spatial information by its structure-aware attention kernels.

**ImageNet-21K.** We have demonstrated the results with HyLiSNets fine-tuned from ImageNet-21k pre-training in Table 6 to check the behavior of our models in the large-scale data regime. We train 90 epochs for ImageNet-21K pre-training, and fine-tune 30 epochs on ImageNet-1K. We follow the setup of ConvNext [60] models for a fair comparison, and all the other details are summarized in Sec.A.2 of the supplementary material. When our models are pre-trained on ImageNet-21K, the accuracies are improved substantially. HyLiSANet-S & -B obtain 0.9% & 1.4% accuracy gains compared to the previous ImageNet-1K results, respectively. Considering iFormer [19] could not be trained on ImageNet-21K due to the manually defined channel ratio [19], these results demonstrate our models can learn general visual concepts when they can access a larger amount of training data. Both HyLiSANet-S and HyLiSANet-B outperform the other models with lower computation, showing the superior scalability of HyLiSNets.

**Throughput.** In Tab. 7, we compare our models with other low-latency ViTs [6, 60, 63] in terms of throughput to show the efficiency of FFTs in

model	img size	FLOPs	#params	mem	imgs/s	top-1
Swin-B [6]	224×224	15.4 G	88 M	1.8 G	306.7	83.5
Swin-B [6]	384×384	47.1 G	88 M	2.9 G	96.5	84.5
ConvNext-B [60]	224×224	15.4 G	89 M	1.8 G	312.5	83.8
ConvNext-L [60]	224×224	34.4 G	198 M	3.1 G	171.2	84.3
FastViT-SA36 [63]	384×384	12.6 G	30 M	4.2 G	194.6	84.5
FastViT-MA36 [63]	384×384	17.7 G	43 M	5.0 G	168.4	84.9
LiSANet-B	224×224	10.4 G	51 M	2.0 G	347.2	84.4
HyLiSANet-B	224×224	11.7 G	50 M	1.8 G	299.4	85.0

**Table 7: Throughput comparison among modern ViT models on ImageNet-1K.** For the images per second (imgs/s) metric, higher is better.

type	model	pretrain	frame× crop×clips	FLOPs	top-1
CNN	X3D-XL [64]	-	16×3×10	1452 G	79.1
	SlowFast+NL [65]	-	16×3×10	7020 G	79.8
ViT	X-ViT [66]	IN-21K	16×3×1	850 G	80.2
	Mformer-L [67]	IN-21K	16×3×10	35553 G	80.2
	ViViT-L [31]	IN-21K	16×3×4	17352 G	80.6
	Swin-B [68]	IN-1K	32×3×4	3384 G	80.6
	TimeSformer-L [69]	IN-21K	16×3×1	7140 G	80.7
Hybrid	MViT-B [70]	-	16×1×5	353 G	78.4
	UniFormer-S [53]	IN-1K	16×1×4	167 G	80.8
	MViTv2-S [16]	-	16×1×5	320 G	81.0
	UniFormer-B [53]	IN-1K	16×1×4	389 G	82.0
	HyLiSANet-S	IN-1K	16×1×4	165 G	81.1
	HyLiSANet-B	IN-1K	16×1×4	428 G	<b>82.5</b>

**Table 8: Comparison to the other models on Kinetics-400.** Pre-trained weights, FLOPs (G), and top-1 accuracy (%) on Kinetics-400 validation set are shown. All the models use 16 input frames except for Swin [68].

LiSA. We measure the model throughputs with their respective official source codes by using an RTX6000 with batch size of 32. With 224 × 224 input, LiSANet-B outperforms both Swin-B [6] and ConvNext models (ConvNext-B, -L) [60] with higher accuracy. In addition, HyLiSANet-B outperforms much larger models (Swin-B with 384 × 384 input, ConvNext-L) in all metrics. FastViT [63] aims to improve the runtime of conventional ViTs by removing skip connections and applying the reparameterization technique. Regarding FastViT [63], our models are better in throughput with higher accuracies (e.g., 6<sup>th</sup> vs 8<sup>th</sup> rows), and HyLiSANet-B outperforms FastViT-MA36 in both metrics. Note that FastViT did not report the results on the 224 × 224 resolution.

model	FLOPs	#params	Mask R-CNN 1× schedule					
			AP <sup>b</sup>	AP <sup>b</sup> <sub>50</sub>	AP <sup>b</sup> <sub>75</sub>	AP <sup>m</sup>	AP <sup>m</sup> <sub>50</sub>	AP <sup>m</sup> <sub>75</sub>
Res-50 [58]	260 G	44 M	38.0	58.6	41.4	34.4	55.1	36.7
PVT-S [61]	245 G	44 M	42.9	65.8	47.1	40.0	62.7	42.9
Twins-S [71]	238 G	44 M	24.0	50.0	41.4	34.4	55.1	36.7
Swin-T [6]	264 G	48 M	42.2	64.6	46.2	39.1	61.6	42.0
ViL-S [72]	218 G	45 M	44.9	67.1	49.3	41.0	64.2	44.1
Focal-T [73]	291 G	49 M	44.8	67.7	49.2	41.0	64.7	44.2
iFormer-S [19]	263 G	40 M	46.2	68.5	50.6	41.9	65.3	45.0
CSwin-T [36]	279 G	42 M	46.7	68.6	51.3	42.2	65.6	45.4
LiSANet-S	258 G	43 M	47.2	68.3	52.0	42.2	65.7	45.2
HyLiSANet-S	265 G	42 M	<b>47.7</b>	<b>69.0</b>	<b>52.5</b>	<b>42.6</b>	<b>65.9</b>	<b>45.9</b>

**Table 9: Comparison with other models on COCO validation set.** FLOPs (G), the number of parameters (M), box mAP (AP<sup>b</sup>) and mask mAP (AP<sup>m</sup>) are shown. Note that FLOPs are measured at resolution 800 × 1280.

## 5.4 Downstream tasks on other domains

**Video action recognition.** We conduct experiments on Kinetics-400 by adjusting HyLiSANet-S & -B for video representation learning. Depth-wise convolutions in the models are transformed from 2D to 3D, and LiSA takes ( $T \times H \times W$ ) tokens for spatio-temporal modeling. We temporally downsample at the first patch embedding layer, and keep the temporal dimension for the rest of the models. For training, we fine-tune the ImageNet-1K trained weights by inflating 2D convolution kernels and LiSA weights. We follow the training recipe of Uniformer [53] and all the experimental details are in Sec.A.2 of the supplementary material. Table 8 demonstrates the state-of-the-art results on Kinetics-400 including 3D CNNs [64, 65], video transformers [31, 66–69], and hybrid models [16, 53, 70]. For a fair comparison, we compare the models taking the same number of input frames (16 frames). Our small model, HyLiSANet-S, outperforms CNN-based models [64, 65] and ViT models [31, 66–69] in accuracy while consuming much fewer FLOPs. Our base model, HyLiSANet-B, outperforms all the other models including hybrid ViTs [16, 53, 70] in accuracy with comparable FLOPs. The results demonstrate the transferability of our models, and further verify that LiSA is highly beneficial for spatio-temporal modeling.

**Object detection & instance segmentation.** To show the generalization ability of LiSA, we conduct object detection experiments on the

model	FLOPs	#params	mIOU (%)
ResNet-50 [58]	183 G	29 M	36.7
PVT-S [18]	161 G	28 M	39.8
Twins-S [71]	144 G	28 M	43.2
Swin-T [6]	182 G	32 M	41.5
UniFormer-S <sub>h32</sub> [53]	199 G	25 M	46.2
UniFormer-S [53]	147 G	25 M	46.6
CSwin-T [36]	202 G	26 M	48.2
LiSANet-S	176 G	27 M	49.2
HyLiSANet-S	184 G	26 M	<b>49.3</b>

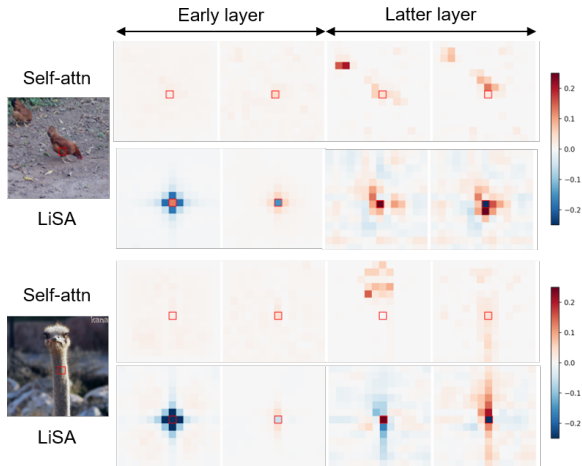
**Table 10: Comparison with other models on ADE-20K.** mIOU (%), FLOPs (G) and the number of parameters (M) are shown. Note that FLOPs are measured at resolution 512×2048.

COCO dataset. We adopt standard Mask R-CNN [74] detection frameworks, which employ ImageNet-1K pre-trained weights for fine-tuning. We use a 1× schedule (12 epochs) and follow the same recipe as in [6]. In Tab. 9, we show the results of object detection and instance segmentation tasks. Our HyLiSANet-S shows the best performances among CNN [58], ViT [6, 61, 72, 73], and Hybrid ViT [19] backbones in AP<sup>b</sup> and AP<sup>m</sup>, while maintaining its efficiency. Since these are high-resolution computer vision tasks (*e.g.*, 800 × 1280), the results demonstrate that LiSA is a proper fit for processing a large number of tokens compared to other attention methods [6, 61, 73].

**Semantic segmentation.** We also evaluate our model on ADE-20K dataset [28]. We adopt the popular Semantic FPN [75] as a basic framework and the model is trained for 80k iterations. The stochastic depth rate is set as 0.15, and we follow the same setting of PVT [18] for a fair comparison. Table 10 summarizes the results on ADE-20K. HyLiSANet-S achieves the best mIOU among different models while requiring fewer FLOPs and the number of parameters, indicating that LiSA is beneficial for processing high-resolutions.

## 5.5 Visualization

In Fig. 4, we visualize both self-attention and LiSA kernels of different layers and heads from isotropic models. As expected, LiSA kernels contain much more diverse patterns compared to self-attention kernels. Self-attention kernels in the early layers often fail to capture relevant context, and those in the latter layers are effective, but they usually capture redundant information. Unlike self-attention, LiSA kernels in the early layers focus on encoding local features. Some of these look similar to



**Fig. 4: Attention kernels of self-attention & LiSA.** Attention kernels from different layers and heads are visualized. For each sample, the top row is self-attention, and the bottom is LiSA. Note that the red box in the center of each subfigure is the query pixel.

Sobel or Laplacian filters, which are beneficial for learning local structural information. Considering that modern hybrid models [2, 19, 33, 37], which replace self-attention with convolution in early layers obtain an extra accuracy gain, the behavior of LiSA kernels in early layers seems reasonable. Meanwhile, LiSA kernels in the latter layers concentrate on the context relevant to the target object like self-attention does. LiSA, however, generates more diverse shapes of kernels than self-attention, which implies that they aggregate the relevant context and further consider structural patterns inside the context at the aggregation. Therefore, this visualization demonstrates that our structure-aware attention kernel can be more expressive and flexible compared to the self-attention kernel.

## 6 Conclusion

In this paper, we have presented LiSA, a novel expressive, yet efficient, attention operator that learns rich structural patterns with log-linear complexity. Our comprehensive analyses have shown that the ViTs based on LiSA, LiSANets, outperform their counterparts in accuracy and computational complexity. LiSANet & HyLiSANet have achieved competitive performance on various kinds

of visual understanding tasks such as image classification, video classification, object detection, and semantic segmentation. While LiSA is effective yet efficient for understanding visual concepts, it still leaves room for improvement in several aspects. First, LiSA encodes structural patterns as a fixed size of the vector, which is set as a parameter  $D$ , but we can further investigate a method to dynamically change the size of structural patterns depending on the visual context. Second, although LiSA is much more efficient than the other attention methods due to the log-linear complexity, it is still heavy to construct multiple layers for processing high resolutions. It would be interesting to examine the effect of LiSA with the local or sparse attention techniques. We believe that LiSA can be a guideline for designing a better basic brick for visual understanding, and we further expect our structure-aware attention could be applied for cross-attention mechanisms, which are widely used for handling multiple modalities such as image-text or video-text.

## Acknowledgments

This work has been supported in part by the ANR grant AVENUE (ANR-18-CE23-0011), the Junta de Andalucía of Spain (P18-FR-3130 and P20\_00430, including European Union funds) and the Ministry of Education of Spain (PID2019-105396RB-I00). We also thank the EuroHPC JU for the GPU computing hours.

## References

- [1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *Proc. International Conference on Learning Representations (ICLR)* (2020)
- [2] Li, K., Wang, Y., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unified transformer for efficient spatiotemporal representation learning. *Proc. International Conference on Learning Representations (ICLR)* (2022)
- [3] Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.-H., Tay, F.E., Feng, J., Yan,

- S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 558–567 (2021)
- [4] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: *Proc. International Conference on Machine Learning (ICML)*, pp. 10347–10357 (2021). PMLR
- [5] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019)
- [6] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 10012–10022 (2021)
- [7] Bello, I.: Lambdanetworks: Modeling long-range interactions without attention. In: *Proc. International Conference on Learning Representations (ICLR)* (2020)
- [8] Zhao, H., Jia, J., Koltun, V.: Exploring self-attention for image recognition. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10076–10085 (2020)
- [9] Kim, M., Kwon, H., Wang, C., Kwak, S., Cho, M.: Relational self-attention: What’s missing in attention for video understanding. *Proc. Neural Information Processing Systems (NeurIPS)* **34**, 8046–8059 (2021)
- [10] Wang, S., Li, B., Khabsa, M., Fang, H., Ma, H.: Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768* (2020)
- [11] Choromanski, K., Likhoshesterov, V., Dohan, D., Song, X., Kane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., *et al.*: Rethinking attention with performers. In: *Proc. International Conference on Learning Representations (ICLR)* (2021)
- [12] Qin, Z., Sun, W., Deng, H., Li, D., Wei, Y., Lv, B., Yan, J., Kong, L., Zhong, Y.: cosformer: Rethinking softmax in attention. *arXiv preprint arXiv:2202.08791* (2022)
- [13] Liutkus, A., Cifka, O., Wu, S.-L., Simsekli, U., Yang, Y.-H., Richard, G.: Relative positional encoding for transformers with linear complexity. In: *Proc. International Conference on Machine Learning (ICML)*, pp. 7067–7079 (2021). PMLR
- [14] Chen, P.: Permuteformer: Efficient relative position encoding for long sequences. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10606–10618 (2021)
- [15] Luo, S., Li, S., Cai, T., He, D., Peng, D., Zheng, S., Ke, G., Wang, L., Liu, T.-Y.: Stable, fast and accurate: Kernelized attention with relative positional encoding. *Proc. Neural Information Processing Systems (NeurIPS)* **34**, 22795–22807 (2021)
- [16] Li, Y., Wu, C.-Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C.: Mvitv2: Improved multiscale vision transformers for classification and detection. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4804–4814 (2022)
- [17] Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 22–31 (2021)
- [18] Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media* **8**(3), 415–424 (2022)
- [19] Si, C., Yu, W., Zhou, P., Zhou, Y., Wang, X., Yan, S.: Inception transformer. *Proc. Neural Information Processing Systems (NeurIPS)* **35**, 23495–23509 (2022)

- [20] Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: Maxvit: Multi-axis vision transformer. In: *Proc. European Conference on Computer Vision (ECCV)*, pp. 459–479 (2022). Springer
- [21] Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8 (2007). IEEE
- [22] Wang, H., Tran, D., Torresani, L., Feiszli, M.: Video modeling with correlation networks. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 352–361 (2020)
- [23] Kwon, H., Kim, M., Kwak, S., Cho, M.: Learning self-similarity in space and time as generalized motion for video action recognition. In: *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 13065–13075 (2021)
- [24] Sun, D., Yang, X., Liu, M.-Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8934–8943 (2018)
- [25] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255 (2009). IEEE
- [26] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017)
- [27] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: *Proc. European Conference on Computer Vision (ECCV)*, pp. 740–755 (2014). Springer
- [28] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torrallba, A.: Scene parsing through ade20k dataset. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 633–641 (2017)
- [29] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: *Proc. European Conference on Computer Vision (ECCV)*, pp. 213–229 (2020). Springer
- [30] Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 7262–7272 (2021)
- [31] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 6836–6846 (2021)
- [32] Vaswani, A., Ramachandran, P., Srinivas, A., Parmar, N., Hechtman, B., Shlens, J.: Scaling local self-attention for parameter efficient visual backbones. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12894–12904 (2021)
- [33] Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., Girshick, R.: Early convolutions help transformers see better. *Proc. Neural Information Processing Systems (NeurIPS)* **34**, 30392–30400 (2021)
- [34] Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y.: Transformer in transformer. *Proc. Neural Information Processing Systems (NeurIPS)* **34**, 15908–15919 (2021)
- [35] Zhang, Z., Zhang, H., Zhao, L., Chen, T., Arik, S.Ö., Pfister, T.: Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding. In: *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, vol. 36, pp. 3417–3425 (2022)
- [36] Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B.: Cswin transformer: A general vision transformer backbone

- with cross-shaped windows. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12124–12134 (2022)
- [37] Dai, Z., Liu, H., Le, Q.V., Tan, M.: Coatnet: Marrying convolution and attention for all data sizes. *Proc. Neural Information Processing Systems (NeurIPS)* **34**, 3965–3977 (2021)
- [38] Jia, X., De Brabandere, B., Tuytelaars, T., Gool, L.V.: Dynamic filter networks. *Proc. Neural Information Processing Systems (NeurIPS)* **29** (2016)
- [39] Li, D., Hu, J., Wang, C., Li, X., She, Q., Zhu, L., Zhang, T., Chen, Q.: Involution: Inverting the inherence of convolution for visual recognition. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12321–12330 (2021)
- [40] Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z.: Dynamic convolution: Attention over convolution kernels. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11030–11039 (2020)
- [41] Ma, N., Zhang, X., Huang, J., Sun, J.: Weightnet: Revisiting the design space of weight networks. In: *Proc. European Conference on Computer Vision (ECCV)*, pp. 776–792 (2020). Springer
- [42] Shen, Z., Zhang, M., Zhao, H., Yi, S., Li, H.: Efficient attention: Attention with linear complexities. In: *Proc. Winter Conference on Applications of Computer Vision (WACV)*, pp. 3531–3539 (2021)
- [43] Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are rnns: Fast autoregressive transformers with linear attention. In: *Proc. International Conference on Machine Learning (ICML)*, pp. 5156–5165 (2020). PMLR
- [44] Rao, Y., Zhao, W., Zhu, Z., Lu, J., Zhou, J.: Global filter networks for image classification. *Proc. Neural Information Processing Systems (NeurIPS)* **34**, 980–993 (2021)
- [45] Lee-Thorp, J., Ainslie, J., Eckstein, I., Ontanon, S.: Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824* (2021)
- [46] d’Ascoli, S., Touvron, H., Leavitt, M.L., Morcos, A.S., Biroli, G., Sagun, L.: Convit: Improving vision transformers with soft convolutional inductive biases. In: *Proc. International Conference on Machine Learning (ICML)*, pp. 2286–2296 (2021). PMLR
- [47] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Proc. Neural Information Processing Systems (NeurIPS)* **30** (2017)
- [48] Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-alone self-attention in vision models. *Proc. Neural Information Processing Systems (NeurIPS)* **32** (2019)
- [49] Strang, G.: A proposal for toeplitz matrix calculations. *Applied Mathematics* **74**(2), 171–176 (1986)
- [50] Fu, D.Y., Dao, T., Saab, K.K., Thomas, A.W., Rudra, A., Re, C.: Hungry hungry hippos: Towards language modeling with state space models. In: *Proc. International Conference on Learning Representations (ICLR)* (2022)
- [51] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* **115**(3), 211–252 (2015) <https://doi.org/10.1007/s11263-015-0816-y>
- [52] Lin, W., Wu, Z., Chen, J., Huang, J., Jin, L.: Scale-aware modulation meet transformer. In: *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 6015–6026 (2023)
- [53] Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern*

- [54] Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., *et al.*: Internimage: Exploring large-scale vision foundation models with deformable convolutions. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14408–14419 (2023)
- [55] Pan, X., Ye, T., Xia, Z., Song, S., Huang, G.: Slide-transformer: Hierarchical vision transformer with local self-attention. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2082–2091 (2023)
- [56] Chu, X., Tian, Z., Zhang, B., Wang, X., Shen, C.: Conditional positional encodings for vision transformers. In: *Proc. International Conference on Learning Representations (ICLR)* (2023)
- [57] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017)
- [58] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (2016)
- [59] Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollár, P.: Designing network design spaces. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10428–10436 (2020)
- [60] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986 (2022)
- [61] Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 568–578 (2021)
- [62] Guo, M.-H., Lu, C.-Z., Liu, Z.-N., Cheng, M.-M., Hu, S.-M.: Visual attention network. *arXiv preprint arXiv:2202.09741* (2022)
- [63] Vasu, P.K.A., Gabriel, J., Zhu, J., Tuzel, O., Ranjan, A.: Fastvit: A fast hybrid vision transformer using structural reparameterization. In: *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 5785–5795 (2023)
- [64] Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020)
- [65] Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: *Proc. IEEE International Conference on Computer Vision (ICCV)* (2019)
- [66] Bulat, A., Perez Rua, J.M., Sudhakaran, S., Martinez, B., Tzimiropoulos, G.: Space-time mixing attention for video transformer. *Proc. Neural Information Processing Systems (NeurIPS)* **34**, 19594–19607 (2021)
- [67] Patrick, M., Campbell, D., Asano, Y., Misra, I., Metze, F., Feichtenhofer, C., Vedaldi, A., Henriques, J.F.: Keeping your eye on the ball: Trajectory attention in video transformers. *Proc. Neural Information Processing Systems (NeurIPS)* **34**, 12493–12506 (2021)
- [68] Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3202–3211 (2022)
- [69] Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: *Proc. International Conference on Machine Learning (ICML)*, vol. 2, p. 4 (2021)
- [70] Fan, H., Xiong, B., Mangalam, K., Li, Y.,

- Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 6824–6835 (2021)
- [71] Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H., Shen, C.: Twins: Revisiting the design of spatial attention in vision transformers. *Proc. Neural Information Processing Systems (NeurIPS)* **34**, 9355–9366 (2021)
- [72] Zhang, P., Dai, X., Yang, J., Xiao, B., Yuan, L., Zhang, L., Gao, J.: Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In: *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 2998–3008 (2021)
- [73] Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., Gao, J.: Focal attention for long-range interactions in vision transformers. *Proc. Neural Information Processing Systems (NeurIPS)* **34**, 30008–30022 (2021)
- [74] He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969 (2017)
- [75] Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6399–6408 (2019)
- [76] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: *Proc. International Conference on Learning Representations (ICLR)* (2018)
- [77] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: *Proc. International Conference on Learning Representations (ICLR)* (2018)
- [78] Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 6023–6032 (2019)
- [79] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826 (2016)
- [80] Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: *Proc. European Conference on Computer Vision (ECCV)*, pp. 646–661 (2016). Springer
- [81] Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.: Randaugment: Practical automated data augmentation with a reduced search space. *Proc. Neural Information Processing Systems (NeurIPS)* **33**, 18613–18624 (2020)
- [82] Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, pp. 13001–13008 (2020)
- [83] Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803 (2018)
- [84] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: *Proc. European Conference on Computer Vision (ECCV)*, pp. 20–36 (2016). Springer
- [85] Fu, D.Y., Kumbong, H., Nguyen, E., Ré, C.: Flashfftconv: Efficient convolutions for long sequences with tensor cores. arXiv preprint arXiv:2311.05908 (2023)

# Supplementary Material of "Lightweight Structure-Aware Attention for Visual Understanding"

## A Implementation details

### A.1 Architecture Details

**LiSA.** In addition to the details included in the main paper, we provide pseudo-code of structure-aware attention (Eq. 4 & Eq. 5 in the main paper) and LiSA in Fig. 6 and Fig. 7, respectively. The notation of multi-head is omitted for clarity and simplicity. As shown in the pseudo-code, we effectively reduce the computational complexity with the FFT approximation.

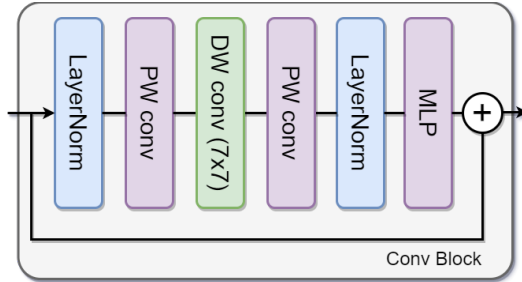
**LiSAnet.** A detailed overview of our proposed architectures is shown in Tab. 11. LiSAnet-I is composed of a single stage following the traditional ViT guidelines [4, 31]. The number of tokens is constant in this model:  $14 \times 14$ . Focusing on the blocks, we have an initial patch embedding layer with a stride of 16 pixels and then, 12 LiSA blocks with an embedding size of 192. On the other hand, our hierarchical ViTs follow the guidelines proposed in [6, 36, 44]. Both models (LiSAnet-S and LiSAnet-B) are composed of the same four stages with different token sizes and numbers of tokens. For hybrid models (HyLiSAnet-S and HyLiSAnet-B), we follow the strategies proposed in [2, 37]. We adopt convolutional blocks for the early two stages and employ LiSA blocks for the other stages. Each convolutional block consists of two pointwise convolutions and one depthwise convolution as illustrated in Fig. 5. We adopt the overlapping patch embedding strategy [18] for hierarchical models. MLP ratios are set to 4 for the early two stages and 3 for the last two stages. We adopt convolutional position embedding [56] and convolutional MLP [18] for our hierarchical models.

### A.2 Experimental Setup

**Image classification (ImageNet-1K).** Our models are trained with AdamW [76] with a weight decay of 0.05 and a learning rate of  $\frac{0.0005}{512} \cdot batch\_size$  with a cosine decay scheduler and 20 warm-up epochs. Our isotropic model (LiSAnet-I) is trained for 150 epochs and hierarchical models are trained for 300 epochs. Following the training recipe proposed in [6], we apply several regularization techniques such as Mixup [77], Cutmix [78], label smoothing [79] and stochastic depth [80]. The stochastic depth strategy is applied only for the hierarchical models with a probability of 0.1 and 0.4 for the LiSAnet-S and LiSAnet-B models, respectively. In addition, we also apply several data augmentation techniques like Rand-Augment [81], random erasing [82], and repeated augmentation. Note that all these hyperparameter values and data-augmentation techniques are selected following the training recipes of the previous works [4, 6].

**Image classification (ImageNet-21K).** For ImageNet-21K pre-training, models are trained with AdamW [76] with a weight decay of 0.05 and a base learning rate of  $4e-3$  & batch size 4096 with a cosine decay scheduler and 5 warm-up epochs. The total number of epochs are set to 90, and the stochastic depth rate is set to 0.1 & 0.2 for HyLiSAnet-S & -B models, respectively. For the other details, we follow the training setup proposed in [60]. For ImageNet-1K fine-tuning, models are trained with AdamW [76] with a weight decay of  $1e-8$  and a base learning rate of  $5e-5$  & batch size 512 with a cosine decay scheduler. The total number of epochs are set to 30, and the stochastic depth rate is set to 0.1 & 0.2 for HyLiSAnet-S & -B models, respectively. For the other details, we follow the training setup proposed in [60].

**Video action recognition (Kinetics-400).** ImageNet-1K pre-trained models (HyLiSAnet-S, HyLiSAnet-B) are adapted to video models and utilized as backbones, and the models are trained with AdamW [76] using a weight decay of 0.05 and  $\frac{0.0001}{32} \cdot batch\_size$  with a cosine decay scheduler and 10 warm-up epochs. The total number of epochs are set to 110, and the stochastic depth rate is set to 0.15 & 0.3 for HyLiSAnet-S & -B models, respectively. For the other regularization or data-augmentation details, we follow the training setup proposed in [53]. Number of input frames is set to 16, and we adopt the dense sampling strategy [83] for training and multi-clip ( $16 \times 1 \times 4$ ) inference for testing. All scores are averaged for the final result.



**Fig. 5: Convolutional block in hybrid models.** ‘PW conv’ denotes a pointwise convolution and ‘DW conv’ denotes a depthwise convolution.

	Output Size	LiSANet-I	LiSANet-S	LiSANet-B	HyLiSANet-S	HyLiSANet-B
Stage1	$\frac{H}{4} \times \frac{W}{4}$	-	Overlap Patch Embed $\downarrow$ 4 LiSA Block (64) $\times$ 2	Overlap Patch Embed $\downarrow$ 4 LiSA Block (96) $\times$ 4	Overlap Patch Embed $\downarrow$ 4 Conv Block (64) $\times$ 3	Overlap Patch Embed $\downarrow$ 4 Conv Block (96) $\times$ 3
Stage2	$\frac{H}{8} \times \frac{W}{8}$	-	Overlap Patch Embed $\downarrow$ 2 LiSA Block (128) $\times$ 4	Overlap Patch Embed $\downarrow$ 2 LiSA Block (192) $\times$ 8	Overlap Patch Embed $\downarrow$ 2 Conv Block (128) $\times$ 6	Overlap Patch Embed $\downarrow$ 2 Conv Block (192) $\times$ 12
Stage3	$\frac{H}{16} \times \frac{W}{16}$	Patch Embed $\downarrow$ 16 LiSA Block (192) $\times$ 12	Overlap Patch Embed $\downarrow$ 2 LiSA Block (320) $\times$ 12	Overlap Patch Embed $\downarrow$ 2 LiSA Block (384) $\times$ 18	Overlap Patch Embed $\downarrow$ 2 LiSA Block (320) $\times$ 12	Overlap Patch Embed $\downarrow$ 2 LiSA Block (384) $\times$ 18
Stage4	$\frac{H}{32} \times \frac{W}{32}$	-	Overlap Patch Embed $\downarrow$ 2 LiSA Block (384) $\times$ 4	Overlap Patch Embed $\downarrow$ 2 LiSA Block (576) $\times$ 3	Overlap Patch Embed $\downarrow$ 2 LiSA Block (384) $\times$ 4	Overlap Patch Embed $\downarrow$ 2 LiSA Block (576) $\times$ 3
Classifier		Global Average Pooling, Linear				

**Table 11: Details of LiSANet variants.** Patch Embed $\downarrow n$  denotes a patch embedding layer that downsamples features with a stride  $n$ .

**Video action recognition (Something-V2).** Video models are trained with AdamW [76] using a weight decay of 0.05 and  $\frac{0.0002}{32} \cdot batch\_size$  with a cosine decay scheduler and 5 warm-up epochs. The total number of epochs is set to 60, and we apply several regularization or data-augmentation techniques following [2, 16, 31]. We adopt the uniform sampling strategy [84] for training and a single crop inference for testing.

**Object detection (COCO).** We adopt standard Mask R-CNN [74] detection frameworks, and ImageNet-1K pre-trained models (LiSANet-S, HyLiSANet-S) are utilized as backbones. We use a 1 $\times$  schedule (12 epochs) with total batch size 16, and follow the same recipe as in [6]. For training, the shorter side of the image is resized to 800 pixels while keeping the longer side no more than 1333 pixels. AdamW [76] with a weight decay of 0.05 is adopted as an optimizer, and the initial learning rate is set to 0.0001. The stochastic depth rate is set to 0.1 and we follow the other details proposed in [6].

**Semantic segmentation (ADE-20K).** We adopt Semantic FPN [75] as a basic framework. ImageNet-1K pre-trained models (LiSANet-S, HyLiSANet-S) are utilized as backbones. The framework is trained for 80k iterations with a cosine decay scheduler. The stochastic depth rate is set to 0.15 and we follow the other details proposed in [18] for a fair comparison.

## B Additional Analyses

**Latency.** We demonstrate both inference and training latency in Table 12. As shown in the table, the training latency follows the tendency of the inference ones. LiSA becomes much faster than self-attention as the number of tokens increases due to its log-linear complexity. We observed that training and inference latency values vary according to the FLOP size except for 7  $\times$  7, where LiSA is slightly slower due to a

---

```

# B: batches, N: tokens, C: channels, D: latent_channels
def structure_aware_attn_Eq4(input, e):
# input shape: [B,N,C], e shape: [2N-1]
qkv = linear_proj(input, channels=3C) # shape: [B,N,3C]
query,key,value = split(qkv, [C,C,C], dim=-1)
# query,key,value shape: [B,N,C]
query = L2norm(query) # shape: [B,N,C]
key = L2norm(key) # shape: [B,N,C]

R = Toeplitz(e) # shape: [N,N]
attn = einsum(query,key, 'BNC,BMC->BNM') # shape: [B,N,N]
attn_R = attn * R # shape: [B,N,N]

out = einsum(attn_R,value,'BNM,BMC->BNC') # shape: [B,N,C]
out = linear_proj2(out, channels=C) # shape: [B,N,C]
return out

def structure_aware_attn_Eq5(input, Ea, Eb, B):
# input shape: [B,N,C], Ea,Eb shape: [2N-1,D], Bb shape: [C,D]
qkv = linear_proj(input, channels=3C) # shape: [B,N,3C]
query,key,value = split(qkv, [C,C,C], dim=-1)
# query,key,value shape: [B,N,C]
query = L2norm(query) # shape: [B,N,C]
key = L2norm(key) # shape: [B,N,C]

Ra = Toeplitz(Ea) # shape: [N,N,D]
Rb = Toeplitz(Eb) # shape: [N,N,D]
K_Ra = einsum(key,Ra,'BMC,NMD->BNCD') # shape: [B,N,C,D]
Rb_V = einsum(Rb,value,'NMD,BMV->BNVD') # shape: [B,N,C,D]
Rb_V_B = Rb_V + B # shape: [B,N,C,D]

out = einsum(query,K_Ra,Rb_v_B,'BNC,BNCD,BNVD->BNV') # shape: [B,N,C]
out = linear_proj2(out, channels=C) # shape: [B,N,C]
return out

```

---

**Fig. 6: Pseudo-code for structure-aware attention.** We describe the way of learning convolutional inductive biases in structure-aware attention (Eq. 4) and its basic form (Eq. 5) presented in Sec. 4.1 of the main paper.

higher number of sequential operations. We employ FlashConv [50, 85] for FFT acceleration and adapt it to our implementation. The kernel fusion in FlashConv addresses the I/O bottleneck by fusing the entire calculation into a single kernel and computing it in GPU SRAM.

block	$(H, W) = (7, 7)$			$(H, W) = (14, 14)$			$(H, W) = (28, 28)$			$(H, W) = (56, 56)$			$(H, W) = (84, 84)$		
	FLOPs (M)↓	infer (ms)↓	train (ms)↓	FLOPs (M)↓	infer (ms)↓	train (ms)↓	FLOPs (M)↓	infer (ms)↓	train (ms)↓	FLOPs (G)↓	infer (ms)↓	train (ms)↓	FLOPs (G)↓	infer (ms)↓	train (ms)↓
Self-attn [47]	22.7	<b>0.9</b>	<b>2.6</b>	102.0	1.6	4.9	584.0	13.0	39.6	5.2	174.7	531.8	22.2	OOM	OOM
Convolutional attn [16]	23.1	1.7	4.6	104.0	2.7	11.0	413.0	8.5	32.1	1.6	30.5	120.8	3.7	68.0	266.5
LiSA (ours)	<b>21.8</b>	1.2	3.2	<b>87.3</b>	<b>1.3</b>	<b>4.3</b>	<b>349.0</b>	<b>4.6</b>	<b>13.6</b>	<b>1.4</b>	<b>19.8</b>	<b>59.2</b>	<b>3.1</b>	<b>45.0</b>	<b>139.5</b>

**Table 12: Comparisons among LiSA & attention operators in FLOPs, inference latency, and training latency.** The latency is measured by an RTX A6000 (batch=32, channels=192). OOM is an abbreviation of out-of-memory.

**Fine-tuning on higher resolutions.** We have verified that fine-tuning LiSAnet on higher resolutions can boost image recognition accuracy. Tab. 13 summarizes the results of LiSAnet-I on ImageNet. LiSAnet obtains a 2.5% gain when we use  $384 \times 384$  resolution. While MLP-mixer models are hard to adapt to higher resolutions since they process a fixed number of tokens, LiSAnet can be easily interpolated to higher resolutions due to the property of Discrete Fourier transform, where each element of the time (*i.e.* spatial) domain is a sampling of a continuous spectrum in the frequency domain. Since the circular embeddings  $\mathbf{W}^a$ ,  $\mathbf{W}^b$  can be considered as samplings of continuous spectrums, changing the resolution is

---

```

# B: batches, N: tokens, C: channels, D: latent_channels
def LiSA(input, Wa, Wb, B):
# input shape: [B,N,C], Wa shape: [N,C,D], Wb shape: [N,D], B shape: [C,D]
qkv = linear_proj(input, channels=3C) # shape: [B,N,3C]
query,key,value = split(qkv, [C,C,C], dim=-1)
# query,key,value shape: [B,N,C]
query = L2norm(query) # shape: [B,N,C]
key = L2norm(key) # shape: [B,N,C]

K_fft = rfft(key, dim=1) # shape: [B,N//2+1,C]
Wa_fft = rfft(Wa, dim=0) # shape: [N//2+1,C,D]
K_Wa = einsum(K_fft, Wa_fft, 'BMK,MKD->BMKD') # shape: [B,N//2+1,C,D]
K_Wa = irfft(K_Wa, dim=1) # shape: [B,N,C,D]

V_fft = rfft(value, dim=1) # shape: [B,N//2+1,C]
Wb_fft = rfft(Wb, dim=0) # shape: [N//2+1,D]
V_Wb = einsum(V_fft, Wb_fft, 'BMV,MD->BMVD') # shape: [B,N//2+1,C,D]
V_Wb = irfft(V_Wb, dim=1) # shape: [B,N,C,D]
V_Wb_B = V_Wb + B # shape: [B,N,C,D]

out = einsum(query,K_Wa,V_Wb_B,'BNK,BNKD,BNVD->BNV') # shape: [B,N,C]
out = layer_norm(out)
out = linear_proj2(out, channels=C) # shape: [B,N,C]
return out

```

---

**Fig. 7: Pseudo-code for LiSA.** We describe the final form of LiSA described in Sec. 4.3 of the main paper.

index	model	Image size	FLOPs	#params	top-1
1	LiSAnet-I	224 × 224	1.21 G	6.36 M	74.9
2	LiSAnet-I	384 × 384	3.62 G	7.82 M	77.4

**Table 13: Fine-tuning to higher resolutions on ImageNet.** Image size, Top-1, accuracy (%), FLOPs (G) and the number of paramaters (M) are shown.

operator	FLOPs	#params	top-1	top-5
Self-attention [47]	7.36 G	5.82 M	18.0	40.9
Self-attention w/ RPE [5]	7.36 G	5.87 M	24.0	50.0
Depthwise conv (3 × 7 × 7) [57]	3.75 G	4.82 M	33.0	60.9
GF layer [44]	3.53 G	6.54 M	28.7	40.0
Lambda convolution [7]	26.87 G	6.34 M	34.5	63.1
RSA [9]	72.58 G	23.45 M	34.1	62.7
LiSA (ours)	5.16 G	8.37 M	<b>38.1</b>	<b>67.1</b>

**Table 14: Comparison with other basic operators on SS-V2.** Top-1, top-5 accuracy (%), FLOPs (G) and the number of parameters (M) are shown.

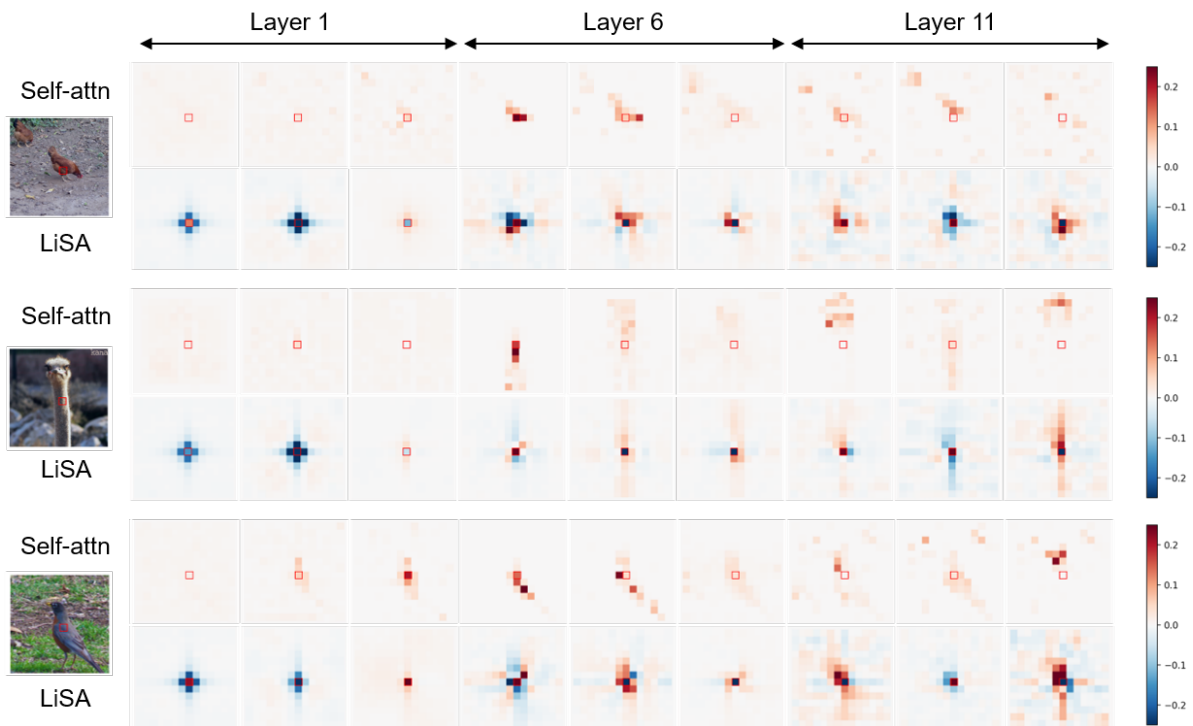
equal to changing the sampling interval of spectrums [44]. Thus, LiSA can be adapted to higher resolutions by simple interpolation.

**LiSA analysis on videos (Something-V2).** In Tab. 14, we compare different types of operators to check the feasibility of LiSA on videos, as done in Tab. 3a for images. The isotropic model (LiSAnet-I) is trained for 60 epochs from scratch, and we adopt the uniform sampling strategy [84] for training and a single crop inference for testing. We sample 8 frames per video, and the rest of the details are the same as in Sec. 5.2. Since structural patterns of videos, *i.e.*, motion patterns, are important cues for recognizing video actions, the operators that learn convolutional inductive biases [7, 8, 44, 57] or geometric structures [9] are more effective than their self-attention counterparts [5, 47]. LiSA shows even better performance on video than image, in terms of both accuracy and complexity. While FLOPs of the attention [8, 47] and highly-expressive [7, 9] operators significantly grow due to the increased number of tokens ( $T \times H \times W$ ), LiSA remains efficient due to its log-linear complexity.

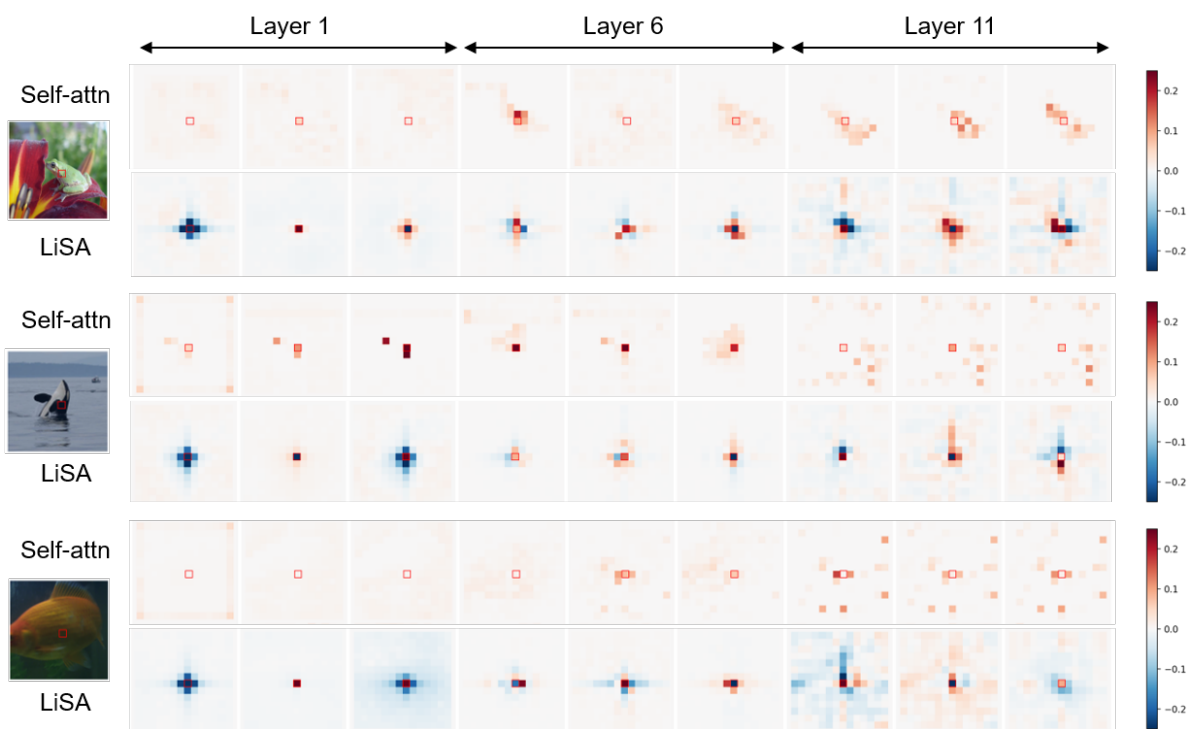
## C Visualization

In Fig. 8, we additionally visualize both self-attention and LiSA kernels of different layers and heads from isotropic models. In the early layers, while self-attention kernels often fail to capture relevant context, LiSA kernels focus on encoding local features. In the latter layers, both self-attention kernels and LiSA kernels are concentrating on the context relevant to the target object. Therefore, LiSA kernels are much more flexible than self-attention kernels due to their ability to capture structural patterns.

In Fig. 9, we visualize both self-attention and LiSA feature maps of different layers from isotropic models. We use the L2-norm of features for each layer. Compared to self-attention, we observed that LiSA captures the geometric layout and positions of the main objects more effectively as the feature map progresses through the layers, verifying the effectiveness of capturing structural patterns.



(a)



(b)

**Fig. 8: Attention kernels of self-attention & LiSA.** Attention kernels from different layers and heads are visualized. For each sample, the top row is self-attention and the bottom is LiSA. Note that the red box in the center of each subfigure is the query pixel.



**Fig. 9: L2-norms of self-attention & LiSA feature maps.** L2-norms of different intermediate feature maps are visualized. For each sample, the top row is self-attention and the bottom is LiSA.