Adaptive Complex Wavelet Informed Transformer Operator

Xiaotong Li, Licheng Jiao, *Fellow, IEEE*, Fang Liu, *Senior Member, IEEE*, Shuyuan Yang, *Senior Member, IEEE*, Hao Zhu, *Member, IEEE*, Xu Liu, *Senior Member, IEEE*, Lingling Li, *Senior Member, IEEE*, and Wenping Ma, *Senior Member, IEEE*

***Abstract*—Visual transformers have achieved great success in representation learning. This is mainly due to efficient token dependency modeling via self-attention. However, the computational burden increases sharply as the input pixels increase. Although recent Fourier-based global frequency-domain mixing methods attempt to improve the efficiency of transformers for high-resolution image inputs, the Fourier operator has limited ability to capture the local geometric structure. Complex wavelets can perform local attention in both the spatial domain and the frequency domain. Therefore, we propose the complex wavelet informed transformer operator that uses the real and imaginary wavelets of the dual-tree complex wavelet transform to simulate the interaction in the attention kernel. In order to further reduce the computational burden of operators, we introduce an adaptive local block shared attention mechanism in the channel domain for our wavelet informed operators. Further, we construct the deep multi-head operator network consisting of a hybrid stack of complex wavelet informed transformer operators and self-attention layers. This enables the Transformer to more sparsely capture multi-scale and multi-directional structured features in the process of learning dependencies. Extensive experimental results show that our adaptive complex wavelet informed transformer operator under the Transformer architecture achieves highly competitive accuracy performance on multiple image classification benchmark datasets. And the proposed operators can be flexibly and effectively migrated to vision tasks in dynamic video scenarios.**

***Index Terms*—Complex wavelet, multiscale frequency representation, neural operator, vision transformer.**

I. Introduction

Visual transformers have shown promise recently in understand and generation tasks due to their strong ability to capture context-dependent representations [1], [2], [3]. While long-range and multi-head dependencies are crucial for understanding relationship between objects in a scene [4], [5], [6], [7], [8], [9], [10], it suffers from quadratic complexity in the number of input pixels for high-dimensional images due to the use of dot product operations [11]. And sequential attention alone also lacks the capture of geometric structured features.

Recently, researchers have introduced an alternative method of self-attention for efficient mixing [12], [13]. They define token mixing as a neural operator that uses the geometry of Fourier transforms to learn mappings between continuous functions in an infinite-dimensional space [14], [15], [16], [17], [18]. The tokens are treated as constant elements in the function space, and the mixture of model labels as successive global convolutions, thus capturing the global relationship of the geometric space.

Similar to neural networks, the purpose of neural operators is to learn the nonlinear mapping of functions. The difference is that neural operators act in an infinite-dimensional function space [19]. Therefore, instantiated neural operators [20], [21], [22], [23], [24] are implemented through different degrees of discretization, and the parameterization therein is independent of the degree of discretization. And they have a substantial computational advantage over classical partial differential equation solvers. However, unlike the discrete finite-dimensional function space of the neural networks, the neural operators extend it to a continuous infinite-dimensional one, further enhancing the model's generalization ability.

The Fourier neural operator (FNO) [25] learns network parameters in Fourier space, and its core is to spectrally decompose the input using the fast Fourier transform (FFT) [26] and compute the convolution-integrating kernel in Fourier space [27]. However, a significant disadvantage of FNO is that the basis functions in the fast Fourier transform lack localized attention in spatial-frequencies. So it cannot exhibit multi-resolution (or multi-scale) resolution properties [28]. Usually, it is difficult to thoroughly analyze and interpret the visual scene only by extracting features at a certain resolution, but the multi-resolution decomposition of space can continuously approximate the visual scene from coarse to fine. Although Swin Transformer [29], Pyramid Vision Transformer [30], and Efficient Attention Pyramid Transformer [31] etc. have designed structures that provide multi-scale features for Transformer, they does not focus on the

Received 14 January 2024; revised 18 July 2024; accepted 10 October 2024. Date of publication 27 January 2025; date of current version 9 June 2025. This work was supported in part by the Key Scientific Technological Innovation Research Project of the Ministry of Education, in part by the Joint Funds of the National Natural Science Foundation of China under Grant U22B2054, in part by the National Natural Science Foundation of China under Grant 62076192 and Grant 62276199, in part by Project 111, in part by the Program for Cheung Kong Scholars and Innovative Research Team in University IRT 15R53, and in part by the Science and Technology Innovation Project through the Chinese Ministry of Education. The associate editor coordinating the review of this article and approving it for publication was Prof. Roger Zimmermann. *(Corresponding author: Licheng Jiao.)*

The authors are with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Joint International Research Laboratory of Intelligent Perception and Computation, School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: lixiaotong@stu.xidian.edu.cn; lchjiao@mail.xidian.edu.cn).

Codes are available at https://github.com/Dawn5786/ACWI-Former.

Digital Object Identifier 10.1109/TMM.2025.3535392

transformation of the attention kernel operator itself. In this case, we consider wavelet transforms with spatial and frequency localization properties.

Since wavelets have the property of focusing on spatial information locally, they can better process signals with discontinuities and spikes. Therefore, compared with the fast Fourier discrete transform, it can better learn the excellent features such as local high frequencies in visual scenes [32]. For example, [33] uses single-tree discrete wavelet transform to transform convolution kernal to the wavelet neural operator that learns in spectral space. Explained further, wavelets provide us with eigenfrequency and eigenspace information [34], [35]. In visual scenes, data often needs to be observed from multi-resolution to be better analyzed and understood. The element-wise multiplication in Transformer is usually not sensitive to changes in resolution space [36]. The wavelet transform can observe the data in the spatial and frequency domains simultaneously and has the advantage of multi-resolution, allowing the data to be observed step by step from coarse to fine.

Therefore, we creatively construct the adaptive complex wavelet informed transformer operator based on theoretical principles of operator learning. Based on the dual-tree complex wavelet structure, we rationally use the real and imaginary wavelets to simulate the key-value-query interaction in the self-attention kernel. And they can be flexibly embedded in Transformer to achieve frequency feature of high-dimensional image data, while reducing the redundancy of the network.

The main contributions of this paper are as follows.

1) We propose a new complex wavelet informed transformer operator learning paradigm with interaction and matching capabilities similar as that in the self-attention kernel. At the same time, it can mix structured representations of multi-resolution spaces in the spatial domain through scaling.
2) We introduce an adaptive local block shared attention mechanism for our complex wavelet informed transformer operator. In the frequency domain, it adaptively shares weight parameters between different frequencies, further reducing the extra computational burden brought by operators.
3) We construct a deep operator Transformer network fuses stacked complex wavelet informed transformer operators. This allows the Transformer to achieve more sparse and efficient transfer learning through the structural prior of the wavelet operator. Moreover, the phase interaction in complex wavelets operator improves the matching ability of the network in dynamic video scenes.

The sections of this article are organized as follows: In Section II, we introduce recent related work in the field of visual transformers and neural operators. In Section III, we elaborate on the theory and method of our complex wavelet attention neural operator. Section IV includes experimental setup, ablation and comparative experiments, transfer learning, analysis and visualization. Section V is a conclusion and outlook.

II. Related Work

A. Vision Transformer

Recently, many vision Transformer models are proposed for various vision tasks [2], [37]. The self-attention layers are stacked as the encoder and decoder in [38], and are placed after the convolutional features to achieve target discrimination and bounding box regression. On the other hand, ViT [39] evenly divides the image into local patches, and uniformly reshapes them into annotated and stacked attention encoding, so as to avoid convolution operations and realize image classification under pure Transformer structure. The training features are processed by the encoder, while the decoder uses a cross-attention layer to fuse the training and test features to compute discriminative features [40], [41].

However, due to the pure Transformer's lack of attention to features of different scales, its recognition capabilities for objects of different sizes vary significantly. Therefore, researchers try to design multi-scale structures for it. For example, Swin Transformer [29] designs a hierarchical structure based on shift operations so that attention can be switched between features of different scales. Pyramid Vision Transformer [30] uses the pyramid structure to improve the Transformer's recognition ability in pixel-level fine-grained scenes. In order to reduce the high-frequency information lost in the down-sampling process of multi-scale attention operations, Wave-ViT [42] uses reversible wavelet transform in the frequency domain to achieve low-loss sampling.

Moreover, transformer structures are also effectively used for dynamic vision scene. Taking visual tracking as an example, TrDiMP [43] employs a model predictor with attention layers to generate model weights. It fuses the template features as training samples. Then, it uses the encoder of the Transformer as a coupler for a set of trusted template features. Afterwards, the decoder is used to achieve matching and decoupling of the search area features of the frame to be tested with template features, thereby predicting the credibility score of the tracking target. TrSiam also encodes the target state information, but integrates it through two different cross-notice modules in the decoder module, instead of using two encoding modules before the transformer.

B. Neural Operator Learning

The essence of the process classical neural network is to find a mapping between the source and target domains in finite-dimensional spaces. Unlike some commonly used methods in recent years, such as diversity-driven active learning to address uncertainty estimation and data bias [44], or using causal inference modeling to resolve semantic confusion bias [45], neural operator learning [28], [46], [47] is an extension of neural networks, which is to learn the mapping operations between infinite-dimensional function spaces.

Operator learning usually uses a combination of linear integral operators and nonlinear activation functions to approximate operators. Activation functions and combinatorial operators can be approximated to complex nonlinear operators. The learning

optimization of operators can be based on the solution process of partial differential equations. Experiments show that the neural operator has the ability to learn the solution of a series of physical constraint equations such as Burgers and Darcy equations. And it can map the learning coefficient to the solution. The learning process of neural operators is also orders of magnitude faster than that of traditional partial differential equation solvers. It can also be well applied to scenarios such as zero-shot and super-resolution. Recently, a variety of neural operators have been proposed and we introduce several representative of them.

Graph neural operators are a type of neural operator that is used for structured data such as graphs. It expands the graph neural network in continuous space and associates the mesh with a set of network parameters based on the Nystrom approximation formula. Due to their typical local connectivity and weight-sharing mechanisms, graph neural networks have been used to simulate a wide range of physical models, such as molecular and rigid body systems. Therefore, the graph neural operator generalizes the basic structures such as graph convolution and graph pooling by constructing a graph with nodes located on the spatial domain of the output function. Graph neural operators can directly learn the core of the network that is very close to the solution of the partial differential equation through gradient transfer. If querying locations, interpolation errors can be avoided by simply adding new nodes to the spatial graph and connecting them to existing nodes. In this way, the graph neural operator not only gains the ability of the graph neural network to process structured association data, but also avoids interpolation errors by using the ability of Nystrom to extend the integration operator.

Low-rank neural operator constructs the low-rank decomposition commonly used in kernel methods and Gaussian processes as the product kernel of a two-factor kernel network. The framework of the factorial low-rank neural operator network consists of a backbone network and a branch network.

The factor network of LNOs is defined on the physical domain, while non-local signals are accrued using integrals relative to Lebesgue metrics. Rather than integrating based on incremental metrics on a predefined set of nodes (which are often thought of as the grid where the data resides).

The Fourier neural operator replaces the kernel integral operator with the Fourier transform, which is commonly used to solve spectra for differential equations, and can also be seen as based on convolution operators defined in Fourier space. (Because the essence of multiplication in the Fourier field is differential.) FNO uses the resolution-invariant solution of the operator to realize the sharing of unadjusted weights in different resolution spaces. And compared with the traditional partial differential equation solving method, its reasoning time is greatly shortened.

As spectral methods of partial differential equations have been extended to neural networks, FNOs have been deeply stacked. The modified adaptive Fourier neural operator can also be adapted to the sequential form input of Transformers. It makes use of the global convolution of Fourier space to achieve efficient blending of input units.

The wavelet neural operator transforms the convolutional kernel of the convolutional neural network into a single-tree wavelet kernel. It takes advantage of the wavelet's strong ability to capture high-frequency features such as mutations or boundaries to improve the information loss in the convolution smoothing process. Such as the transparent operator network [48] utilizes a learnable Morlet wavelet operator and physically understandable modules to achieve superior diagnostic performance with fewer parameters and enhanced fault feature transparency. However, it is difficult for single-tree wavelets to construct aliasing-free operators with matching or contrasting capabilities by using reasonable phase information intervals. Therefore, it is slightly difficult to construct a similar self-attention kernel with key-value-query matching function in Transformer. So, we propose a complex wavelet informed transformer operator. We use real wavelets and imaginary wavelets with specific phase differences in the dual-tree complex wavelet transform to simulate the interactive operations of the attention kernel. Furthermore, we utilize the local block structure attention in the channel domain to reduce the complexity of our wavelet informed transformer operator.

III. Method

A. Neural Operator Learning

Operator Learning: Operator learning is a process of approximating complex nonlinear mappings in infinite-dimensional continuous function spaces with combinations of integral and nonlinear operators. For the sake of brevity, we take the functions of one variable $x \in D$ as an example. Specifically, we assume that the input and output functions $a(x)$ and $u(x)$ in two separable Banach function spaces $\mathcal{A}$ and $\mathcal{U}$ defined on bounded domains $D \subset \mathbb{R}^d$.

$$\mathcal{G}_{\theta^\star} : \mathcal{A} \to \mathcal{U}, \theta^\star \in \Theta$$
$$s.t. \min \|\mathcal{G}^\star - \mathcal{G}_{\theta^\star}\|^2_{L^2_\mu(\mathcal{A},\mathcal{U})} \tag{1}$$

The purpose of operator learning is to approximate the nonlinear mapping:$\mathcal{G}^\star : \mathcal{A} \to \mathcal{U}$ by constructing the operator $\mathcal{G}_\theta$ in the space $L^2_\mu(\mathcal{A},\mathcal{U})$ with parameters θ from a finite-dimensional space Θ so that $\mathcal{G}_{\theta^\star} \approx \mathcal{G}^\star$ where $\theta^\star \in \Theta$ and μ is a probability measure:

In order to describe a pattern that the operator learns from data-driven input-output pairs, we set N-point discretization of the source domain D as $X_N = \{x_1, \ldots, x_N\}$. And then, we denote one of the J input-output observation pairs with index j as $\{a_j|X_N, u_j|X_N\}_{j=1}^J$.

Neural Operator: The neural operators extend the neural network from finite-dimensional discrete function space to infinite-dimensional continuous function space, thereby enhancing the generalization ability of the structure. For neural operators with multiple hidden layers, the operator $\mathcal{G}_\theta$ can be constructed in the following iterative form as:

$$\mathcal{G}_\theta := \mathcal{Q} \circ \sigma_L(\mathcal{K}_{L-1}) \circ \cdots \circ \sigma_1(\mathcal{K}_0) \circ \mathcal{P} \tag{2}$$

The operator $\mathcal{P} : \{a : \mathbb{R}^d \to \mathbb{R}^{d_a}\} \mapsto \{v_0 : \mathbb{R}^d \to \mathbb{R}^{d_{v0}}\}$ lift the input $a(x)$ to the initial high dimensional hidden $v_0(x)$. Then, the iterative operator $\mathcal{K}_{l-1} : \{v_{l-1} : \mathbb{R}^{d_{l-1}} \to \mathbb{R}^{d_{v_{l-1}}}\} \mapsto \{v_l : \mathbb{R}^{d_l} \to \mathbb{R}^{d_{v_l}}\}$ maps the hidden representation to the next

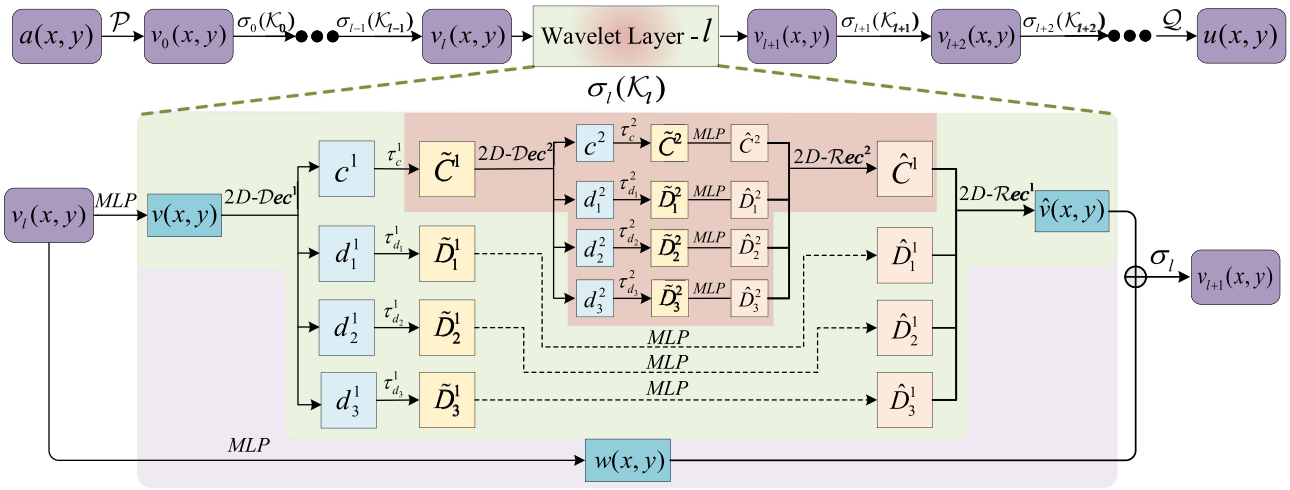

Fig. 1. The schematic diagram of 2-level wavelet informed transformer operator layer. The transformation from $a(x, y)$ to $u(x, y)$ represents the iterative process of the multi-hidden layer. The detailed display on the figure shows the process of obtaining the output $v_{l+1}(x, y)$ from the input $v_l(x, y)$ through operator $\sigma_l(\mathcal{K}_l)$, as follows: Step 1: Perform two independent MLP operations on input $v_l(x, y)$ to obtain $v(x, y)$ and $w(x, y)$, respectively. Step 2: Perform wavelet decomposition operator (2D-$\mathcal{D}ec^1$, 2D-$\mathcal{D}ec^2$) and wavelet attention operator ($\mathcal{T}_c^1, \mathcal{T}_c^2$) on $v(x, y)$ level-by-level to obtain the informed decomposition coefficients $\{\tilde{C}^1, \tilde{D}_1^1, \tilde{D}_2^1, \tilde{D}_3^1\}$, $\{\tilde{C}^2, \tilde{D}_1^2, \tilde{D}_2^2, \tilde{D}_3^2\}$. Step 3: Perform MLP on $\{\tilde{D}_1^1, \tilde{D}_2^1, \tilde{D}_3^1\}$, $\{\tilde{C}^2, \tilde{D}_1^2, \tilde{D}_2^2, \tilde{D}_3^2\}$ and reconstruct them into the original space to obtain $\hat{v}(x, y)$ by wavelet reconstruction operator (2D-$\mathcal{R}ec^1$, 2D-$\mathcal{R}ec^2$). Step 4: Add $w(x, y)$ to $\hat{v}(x, y)$ and pass through the activation map σ_l to get the output $v_{l+1}(x, y)$, i.e., the input of the $(l+1)$-th layer.

layer, where $l \in \{0, \ldots, L-1\}$. $\sigma_l : \mathbb{R}^{d_{v_l}} \to \mathbb{R}^{d_{v_l}}$ is the activation map for each layer. Finally, another operator $\mathcal{Q} : \{v_L : \mathbb{R}^{d_L} \to \mathbb{R}^{d_{v_L}}\} \to \{u : \mathbb{R}^{d_L} \to \mathbb{R}^{d_u}\}$ maps $v_L(x)$ back to the solution space $u(x)$.

Attention neural operator: To facilitate the study of the iterations of neural operators for a single-head self-attention hidden layer, we record the the iterative operator $\mathcal{K}_{l,(R_q,R_k,R_v)}$ step-update from v_l to v_{l+1} as:

$$v_{l+1}(x) = \sigma_{l+1}(W_l v_l(x) + \mathcal{K}_{l,(R_q,R_k,R_v)} v_l(x)) \tag{3}$$

where W_l denote local linear operators, and $\{R_q, R_k, R_v\} \in \mathbb{R}^{d_{v_l} \times d_{v_{l+1}}}$ are learnable free parameters. For attention neural operator in continuous space, there is:

$$(\mathcal{K}_{l,(R_q,R_k,R_v)}(v_l))(x) = \int_{D_l \in \mathbb{R}^{d_{v_l}}} \mathcal{S} \cdot R_v v_l(y) d(v_l(y))$$
$$s.t. \quad \mathcal{S} = \frac{exp(\langle R_k v_l(x), R_q v_l(y) \rangle / \sqrt{d_{v_l}})}{\int_{D_l \in \mathbb{R}^{d_{v_l}}} exp(\langle R_k v_l(s), R_q v_l(y) \rangle / \sqrt{d_{v_l}}) ds} \tag{4}$$

where $\langle \cdot, \cdot \rangle$ is the inner-product. We define the operator kernel $\kappa_{l,(R_q,R_k)}(x, y) = softmax(exp(\langle R_k v_l(x), R_q v_l(y) \rangle / \sqrt{d_{v_l}})$, and let $\mathcal{L}_{l,R_v}(\cdot) = R_v v_l(\cdot)$. So the above formula can be organized as follows:

$$(\mathcal{K}_{l,(R_q,R_k,R_v)}(v_l))(x) = \int_{D_l} \kappa_l(x, y) \mathcal{L}_l(y) dy \tag{5}$$

In the Transformer networks, the above integral form of the attention operator in the continuous space is discretized into an element-wise multiplication operation.

B. Wavelet Informed Transformer Operator.

In visual scenes, data often faces the flow of multi-resolution space, and the above-mentioned attention neural operator in Transformer has quadratic complexity, which makes it usually inflexible to changes in resolution space. In contrast, the wavelet transform can comprehensively observe the data through the spatial and frequency domains. It has the advantage of multi-resolution so that visual data can be observed step by step from coarse to fine. And as a type of shift-equivariant convolution operation, it has lower computational complexity than integrals. Therefore, we define the wavelet neural operator as Fig. 1 to replace the attention neural operator in the Transformer structure.

Wavelet Operator: We define the wavelet operator $\mathcal{K}_\Psi$ with the kernel operator $\mathcal{W}_\Psi$ in the parametric space for a mother wavelet $\Psi(x) \in L^2(\mathbb{R})$, and $\mathcal{W}_\Psi^{-1}$ as its inverse. $\mathcal{T}_\Psi$ is a wavelet attention operator in the frequency domain.

$$(\mathcal{K}_\Psi(v_l))(x) = \mathcal{W}_\Psi^{-1}\{\mathcal{T}_\Psi(\mathcal{W}_\Psi(v_l))\}(x) \tag{6}$$

Specifically, the operators $\mathcal{W}_\Psi$ and $\mathcal{W}_\Psi^{-1}$ are:

$$(\mathcal{W}_\Psi(v_l))(\alpha, \beta) = \int_D v_l(x) \Psi^*_{\alpha,\beta}(x) dx \tag{7}$$

$$(\mathcal{W}_\Psi^{-1}(\omega_l))(x) = \frac{1}{c_\Psi} \int_{D^*} \frac{d\alpha}{\alpha^2} \int_D \omega_l(\alpha, \beta) \Psi_{\alpha,\beta}(x) d\beta \tag{8}$$

where ω_l represents the corresponding wavelet coefficient function. The scale wavelets are generated with scale factor α and shift factor β as $\Psi_{\alpha,\beta}(x) = 1/|\alpha|^{1/2} \Psi((x-\beta)/\alpha)$. c_Ψ is the admissible constant. Ψ^* refers to the conjugate of Ψ.

The operator $\mathcal{T}_\Psi$ is defined as (9).

$$(\mathcal{T}_\Psi(\omega_l))(\alpha) = W_\Psi(\tau_\Psi \cdot \omega_l(\alpha)) + b_\Psi \tag{9}$$

where $\tau_\Psi \in \mathbb{R}^{d_{w_l} \times d_{w_l}}$ is a learnable free parameter, W_Ψ is a local linear operator, and b_Ψ is a bias function.

For simplicity and practicality, discrete wavelet transform (DWT) uses scale basis $\{\phi_{k,j}(x); j = 0, 1, \ldots, 2^k - 1\}$ and wavelet basis $\{\psi_{k,j}(x); j = 0, 1, \ldots, 2^k - 1\}$ to construct discrete wavelet decomposition and reconstruction in multi-resolution spaces $V_k = span(\phi_{k,j}(x) : j \in \mathbb{N})$ and $W_k = span(\psi_{k,j}(x) : j \in \mathbb{N})$.

$$\phi_{k,j}(x) = 2^{k/2}\phi(2^{k/2}x - j) \tag{10}$$

$$\psi_{k,j}(x) = 2^{k/2}\psi(2^{k/2}x - j) \tag{11}$$

The two sequences of coefficients known as discrete low-pass filter h_n^ϕ and high-pass filter h_n^ψ, that satisfy the following equations for fast discrete wavelet transformations:

$$h_n^\phi = \langle \phi_{0,0}, \phi_{1,n} \rangle \quad s.t. \phi(x) = \sqrt{2}\sum_n h_n^\phi \phi(2x - n) \tag{12}$$

$$h_n^\psi = \langle \psi_{0,0}, \phi_{1,n} \rangle \quad s.t. \psi(x) = \sqrt{2}\sum_n h_n^\psi \phi(2x - n) \tag{13}$$

There is a nesting relationship of V_k that $\{\cdots \subset V_k \subset V_{k-1} \subset \cdots \subset V_1 \subset V_0\}$, where $V_{m+1} = V_0 \oplus W_0 \oplus W_1 \oplus \cdots \oplus W_m$. W_k is the orthogonal complement of V_k inside the subspace V_{k-1}.

Wavelet Informed Transformer Operator: According to the above nesting and orthogonal relationship, there is a representation based on ϕ and ψ of the multi-scale subspaces for $v_l(x)$ as (14). The discrete wavelet decomposition operator is denoted as $\mathcal{D}ec$. c and d are corresponding low and high-frequency decomposition coefficients. $\hat{\mathcal{W}}$ is the discrete representation of wavelet kernel operator $\mathcal{W}$.

$$(\mathcal{D}ec_{(\phi,\psi)}v_l)(k,j) = \begin{cases} c_{k,j} := (\hat{\mathcal{W}}_\phi v_l)(k,j) = \langle v_l, \phi_{k,j} \rangle; \\ d_{k,j} := (\hat{\mathcal{W}}_\psi v_l)(k,j) = \langle v_l, \psi_{k,j} \rangle; \end{cases} \tag{14}$$

The discrete wavelet reconstruction operator is denotes as $\mathcal{R}ec$.

$$\begin{aligned}(\mathcal{R}ec_{(\phi,\psi)}(\omega_\phi, \omega_\psi))(x) &= (\hat{\mathcal{W}}_\phi^{-1}\omega_\phi + \hat{\mathcal{W}}_\psi^{-1}\omega_\psi))(x) \\ &= \sum_j \omega_\phi(j)\phi_j + \sum_k \sum_j \omega_\psi(k,j)\psi_{k,j}\end{aligned} \tag{15}$$

So the discrete wavelet informed transformer operator is denoted as follows:

$$\begin{aligned}(\mathcal{K}(\phi,\psi)v_l)(x) &= \mathcal{R}ec_{(\phi,\psi)}\{\mathcal{T}(\mathcal{D}ec_{(\phi,\psi)}v_l)\}(x) \\ &= \mathcal{R}ec_{(\phi,\psi)}\{\mathcal{T}_\phi(\hat{\mathcal{W}}_\phi v_l)(j) + \mathcal{T}_\psi(\hat{\mathcal{W}}_\psi v_l)(k,j))(x)\end{aligned} \tag{16}$$

Complex Wavelet Informed Transformer Operator: In order to further avoid the displacement sensitivity of the discrete wavelet transform, we further introduce the complex wavelet neural operator. In order to use the phase information of complex wavelets to construct a informed operator with matching advantages, we further introduce the wavelet transformer operator in complex case. Specifically, in the wavelet transform, when the corresponding wavelet basis approximately satisfies the Hilbert transform relation, a dual-tree complex wavelet transform can be formed as follows.

$$\psi^c(x) = \psi_h(x) + i\psi_g(x) \tag{17}$$

where ψ_h and ψ_g are orthogonal wavelets and form a Hilbert transform pair respectively, i.e. $\psi_h = \mathcal{H}(\psi_g)$. Then the corresponding complex wavelet transform can be denoted as:

$$\langle v(x), \psi^c(x) \rangle = \langle v(x), \psi_h(x) \rangle + i\langle v(x), \psi_g(x) \rangle \tag{18}$$

And the decomposition complex wavelet operator $\mathcal{D}ec^c$ can be expressed as:

$$(\mathcal{D}ec^c_{(\phi^c,\psi^c)}v) = \begin{cases} c^c = (\hat{\mathcal{W}}_{\phi^c}v) = \langle v, \phi^c \rangle \\ d_k^c = (\hat{\mathcal{W}}_{\psi_k^c}v) = \langle v, \psi_k^c \rangle \\ d_{k+3}^c = (\hat{\mathcal{W}}_{\psi_{k+3}^c}v) = \langle v, \psi_{k+3}^c \rangle \end{cases} \tag{19}$$

Based on the saliency mechanism of the image, we assume attention operator g^c with a specific form for kernel, which does not explicitly depend on the spatial variables x but on the input pair $(\hat{\mathcal{W}}_{*^c}v_t(x_t), \hat{\mathcal{W}}_{*^c}v_{t'}(x_{t'}))$. Thus, we let:

$$\begin{aligned}&g^c_{tt'}(\hat{\mathcal{W}}_{*^c}v_t(x_t), \hat{\mathcal{W}}_{*^c}v_{t'}(x_{t'})) = g^r_{tt'}(x) + ig^i_{tt'}(x) = \\ &(|A(\hat{W}_{*^c}v_t)||A(\hat{W}_{*^c}v_{t'})|)^{-1}[(\hat{\mathcal{W}}_{*^r}v_t \cdot \hat{\mathcal{W}}_{*^r}v_{t'} - \hat{\mathcal{W}}_{*^i}v_t \cdot \hat{\mathcal{W}}_{*^i}v_{t'}) \\ &+ i(\hat{\mathcal{W}}_{*^r}v_t \cdot \hat{\mathcal{W}}_{*^i}v_{t'} + \hat{\mathcal{W}}_{*^i}v_t \cdot \hat{\mathcal{W}}_{*^r}v_{t'})]\end{aligned} \tag{20}$$

Then, $\mathcal{T}$ in (16) is redefined as $\mathcal{T}^c$, in particular:

$$\begin{aligned}&\mathcal{T}^c_*(\hat{\mathcal{W}}_{*^c}v_t, g^c_{tt'}(\hat{\mathcal{W}}_{*^c}v_t, \hat{\mathcal{W}}_{*^c}v_{t'})) \\ &= \mathcal{T}^r_* \cdot (\hat{\mathcal{W}}_{*^r}v_t \cdot g^r_{tt'} + \hat{\mathcal{W}}_{*^i}v_t \cdot g^i_{tt'}) \\ &\quad + i\mathcal{T}^i_* \cdot (\hat{\mathcal{W}}_{*^i}v_t \cdot g^r_{tt'} - \hat{\mathcal{W}}_{*^r}v_t \cdot g^i_{tt'})\end{aligned} \tag{21}$$

Thus, the dual-tree complex wavelet informed transformer operator can be credited as:

$$\begin{aligned}&(\mathcal{K}^c(\phi^c,\psi^c))(v_t(x), v_{t'}(x)) = \mathcal{R}ec_{(\phi,\psi)}\{\mathcal{T}^c_*(\mathcal{D}ec^c_{(\phi,\psi)}v)\}(x) \\ &= \mathcal{R}ec^c_{(\phi,\psi)}\left\{\sum \mathcal{T}^c_*(\hat{\mathcal{W}}_{*^c}v_t, g^c_{tt'}(\hat{\mathcal{W}}_{*^c}v_t, \hat{\mathcal{W}}_{*^c}v_{t'}))\right\}\end{aligned} \tag{22}$$

Then, the gradient message of the complex wavelet informed transformer operator is passed through the network and indicates the update in the following route:

$$\Delta\phi(x,s) = \phi_t(x - s/2) - \phi_{t'}(x + s/2) \tag{23}$$

$$ds = \Delta\phi(x,s)/(\partial\frac{\Delta\phi(x,s)}{\partial s}) \tag{24}$$

where $s = (\varphi_t(x - s^t/2) - \varphi_t(x + s^t/2))/\omega$.

The complex expansion of wavelet informed operator constructed above can also reduce the translational sensitivity and improves the direction selectivity compared with the single real one. These advantages enable the operator to extract the spectral features more efficiently while utilizing phase information for interactive matching.

C. Adaptive Local Block Attention Mechanism.

Neural networks usually have high-dimensional hidden layer channel sizes. In a complex wavelet informed transformer operator in space $\mathbb{R}^{W \times H \times D_l}$, since the wavelet attention operator in it has a $O(HWD_l^2)$ complexity along the channel direction,

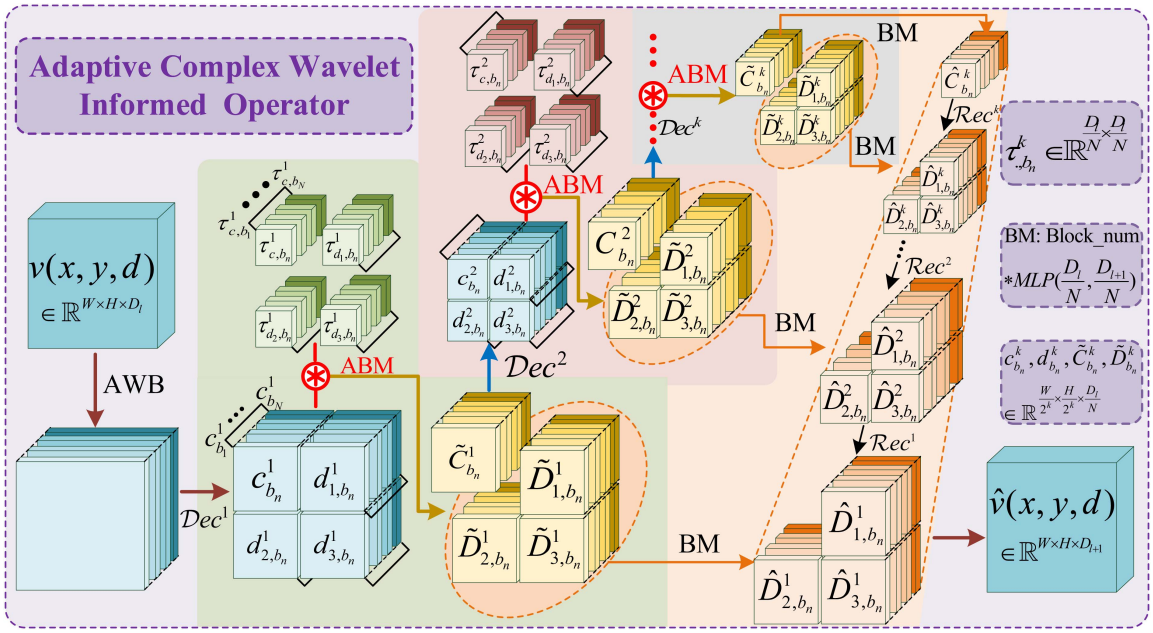

Fig. 2. The structure of adaptive complex wavelet informed operator with adaptive local block attention mechanism. The input of the operator is $v(x,y,d) \in \mathbb{R}^{W\times H\times D_l}$, AWB represents the adaptive weight block operation that divides channel dimension D_l into N blocks; ABM represents the adaptive block multiplication operator; BM represents the blocked MLP operation. The output of the operator is $\hat{v}(x,y,d) \in \mathbb{R}^{W\times H\times D_{l+1}}$.

it may limit the deep stacking of operators in the network. To improve the operator's efficiency, we construct a streamlined adaptive local block attention mechanism by adaptively partitioning the attention operator in the wavelet frequency domain. The operator structure is shown in Fig. 2.

Assume that in a neural operator network, $v(x,y,d) \in \mathbb{R}^{W\times H\times D_l}$ is an intermediate variable obtained by l-th layer input $v_l(x,y,d)$ through multi-layer perceptron (MLP). First, we equally divide $v(x,y,d)$ along the channel dimension into N blocks of dimension D_l/N (i.e., adaptive weight block, AWB). Then each block is decomposed by a k-level two-dimensional discrete wavelet decomposition operator, and k-resolution blocked low-frequency components $c^k_{b_n}$ and high-frequency component $d^k_{b_n}$ in space $\mathbb{R}^{W/2^k\times H/2^k\times D_l/N}$ are obtained. After that, we propose an adaptive block multiplication (ABM) operator to replace the operator τ_Ψ in (9) as:

$$\begin{aligned} ABM(\omega_\phi,\omega_\psi)[k] &= ABM(c^k, d^k_j) \\ &= \sum_n^N \tau^k_{c,b_n} c^k_{b_n} + \sum_n^N \sum_{j=1}^3 \tau^k_{d_j,b_n} d^k_{j,b_n} \end{aligned} \tag{25}$$

where $\tau^k_{b_n}$ in space $\mathbb{R}^{D_l/N\times D_l/N}$ is the learnable free parameter matrix of the ABM operator. Through this adaptive block weight operation, the calculation amount is reduced to $O(NHW(D_l/N)^2)$. After that, we map the frequency components of each resolution from $\mathbb{R}^{D_l}$ space to $\mathbb{R}^{D_{l+1}}$ space by N blocked MLP (BM). Finally, all levels of wavelet domain representation are reconstructed back to the source space as $\hat{v}(x,y,d)$ by the discrete wavelet reconstruction operator.

In summary, we design an adaptive local block attention structure to sharing weight along the channel axis in the wavelet domain. By this, we reduce the network's computational cost and parameter memory usage. Specifically, it reduces the storage occupancy of the parameter matrix $\tau^k_{b_n}$ from $D\times D$ to $(D/N)\times(D/N)$, the storage occupancy of the component matrix $c^k_{b_n}$, $x^k_{b_n}$, etc. from $(W/2^k)\times(H/2^k)\times D$ to $(W/2^k)\times(H/2^k)\times(D/N)$, and the computational cost from WHD^2 to $NWH(D/N)^2$.

ACWI-Former: This work also constructs an ACWI-Former network for visual scenes based on the Transformer framework. Taking the image classification scene as an example, we show the specific network structure in Fig. 3. We construct the network's backbone by sandwiching stacked ACWI layers and several multi-head self-attention layers. The head of the network is a linear classifier to obtain classification results.

We mainly study model variants at two magnitudes: ACWI-Tiny and ACWI-Small. We design these two models with similar computational complexity to ViT-tiny and ViT-small by simply adjusting the number of combined layers and heads, referring to [39]. Detail settings are shown in Table I. ACWI-Tiny/4 represents a variant model with 4×4 input patch size. We keep the ratio of hidden size to the number of layers at 64. The input resolution is uniformly set to 224×224. We set block=1 as a general CWI-Former network and block=12 as an example of ACWI-Former. The following experimental section shows a more fine-grained comparison of block number values.

IV. Experimental Results and Discussion

We conduct extensive experiments on datasets ImageNet-1k [51] to verify the feasibility and effectiveness of our ACWI-Former. And simultaneously, we compare and analyze the evaluation results on vision Transformers [39] and Convnets [52] with

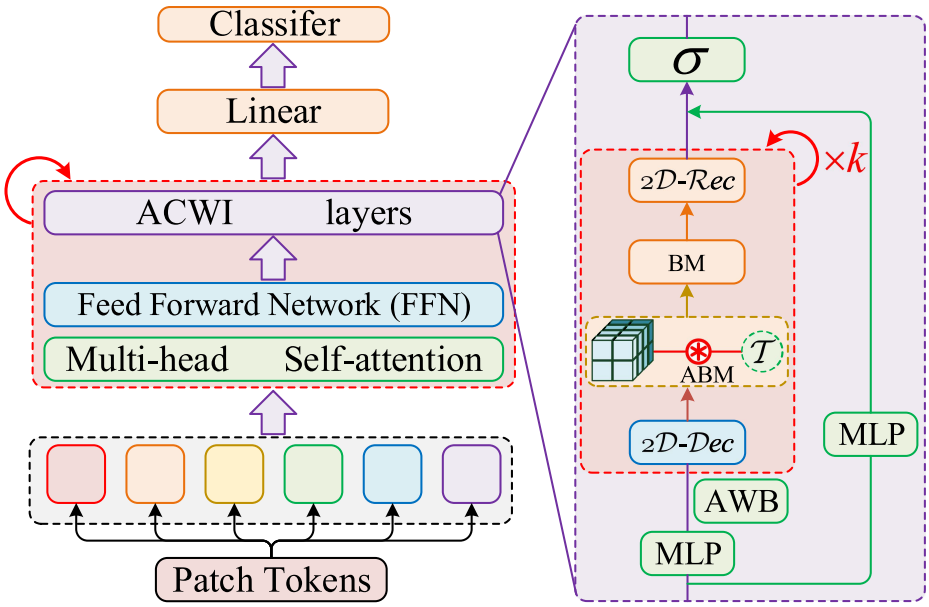

Fig. 3. ACWI-former network for image classification. The input is patch tokens that are divided and encoded from image samples. The backbone is constructed by sandwiching stacked MHSA (multi-head self-attention) layers and adaptive complex wavelet informed operators. The head is a combination of a linear and softmax layer for classification.

TABLE I
DETAILED STRUCTURE OF MULTIPLE VARIANTS OF ACWI-FORMER

Model	Modules	Hidden size	Blocks	Batch size	Heads	Resolution	Params
CWI-Tiny/16	[6,1]×2	192	1	256	3	224	6.47M
ACWI-Tiny/4	[6,1]×2	192	12	256	3	224	6.87M
ACWI-Tiny/16	[6,1]×2	192	12	256	3	224	5.89M
CWI-Small/16	[3,1]×4	384	1	128	6	224	24.16M
ACWI-Small/4	[3,1]×4	384	12	128	6	224	23.35M
ACWI-Small/16	[3,1]×4	384	12	128	6	224	22.30M

different magnitudes, AWNO-Tiny and ACWI-Small. Moreover, we analyze the superiority of our operator over other attention-alternating mixing modalities on a consistent network architecture. Finally, we present the transfer learning effectiveness of the proposed methods based on public benchmarks, including Flower-102 [53], Stanford-Car [54], CIFAR-10 [55], and CIFAR-100 [56]. In addition, we also verify the scalability of ACWI transformer operatpr in the dynamic video tracking task. On the tracking benchmarks OTB [57] and UAV123 [58], our ACWI-Former layer achieves benefits at both Siamese [59] and DiMP [60] tracking frameworks.

A. ACWI-Former Setups

We first build our adaptive complex wavelet informed Former in visual classification scene based on large-scale public dataset ImageNet-1k [61]. It consists of 1.2 million training, 50,000 validation and 50,000 testing images belonging to 1,000 categories with a resolution of 224× 224pixels. In this paper, we mainly verify the performance of the algorithm through downstream tasks such as image classification and video tracking.

In image classification tasks, we mainly use GFLOPs, Top-1 accuracy (Top-1 Acc), and Top-5 accuracy (Top-5 Acc) for the large Imagenet-1 K dataset; and use Accuracy (Acc) for the relatively small Flowers-102, Stanford Cars, CIFAR-10, and CIFAR-100 datasets as evaluation metrics. GFLOPs evaluate the computational complexity of the model and help understand the computational resources required for its runtime; Top-1 Acc reflect the most accurate predictions; and Top-5 Acc provides a more relaxed accuracy assessment. While Acc refers to the ratio of correctly classified samples to the total number of samples. It can directly reflect the overall correctness of the model across all samples, which facilitates a quick assessment of the model's overall performance.

In video tracking tasks, we use the Area Under the Success Rate Curve (AUC) as the evaluation metric. AUC refers to the area enclosed by the success rate curve and the coordinate axis. A high AUC value indicates that the tracker maintains a high success rate across various overlap rate thresholds, demonstrating strong robustness and accuracy.

The CWI-Tiny and the CWI-Small are two control models designed based on complex wavelet informed transformer operators with no weight-adaptive diagonal module at different

TABLE II
ABLATION STUDY AND COMPARISON WITH OTHER CLASSIFICATION ARCHITECTURES ON IMAGENET-1K WITH THE EVALUATION INDICATORS GFLOPS, TOP-1 ACC, AND TOP-5 ACC

Model	Params	Patch size	GFLOPs	Top-1 Acc	Top-5 Acc
DeiT-Tiny [49]	5M	16	1.2	72.20%	91.10%
GFNet-Tiny [13]	7M	4	1.3	74.60%	92.20%
AFNO-Tiny [50]	4M	4	2.1	74.75%	92.52%
CWI-Tiny/16	6M	16	15.4	74.89%	92.63%
ACWI-Tiny/4	6M	4	1.8	75.28%	92.90%
ACWI-Tiny/16	5M	16	1.4	75.46%	93.87%
GFNet-XS	16M	4	2.9	78.60%	94.20%
DeiT-Small [49]	22M	16	4.6	79.80%	95.00%
GFNet-Small [13]	25M	4	4.5	80.00%	94.90%
AFNO-Small [50]	16M	12	6.8	80.89%	95.39%
ViT-Base/16 [39]	86M	16	–	77.90%	83.60%
GFNet-Base [13]	43M	4	7.9	80.70%	95.10%
CWI-Small/16	24M	16	42.2	80.84%	95.35%
ACWI-Small/4	23M	4	4.6	81.26%	95.47%
ACWI-Small/16	22M	16	3.7	81.43%	95.63%

parameter scales. The size of their input patch is set to 16 referring to the model settings of ViT. Details are also shown in Table I. The input patch sizes for the ACWI-*/4 and ACWI-*/16 are 4×4 and 16×16, respectively. We keep the ratio of hidden size to the number of layers at 64in our models. The number of network heads for tiny- and small- models is set to 3 and 6, respectively. So the corresponding hidden layer dimensions are 192 and 384.

During training, we randomly initialize the network with the AdamW optimizer and a cosine-decayed learning rate. The initial learning rate is 0.0005, the weight decay is 0.05, and the warm-up 10 epochs. We does not use additional data augmentation operations for fairness. All the models are trained with 300 total epochs.

Our entire experiment was conducted on two Nvidia GeForce RTX 3090 GPUs, each with 24 GB of memory. The proposed framework was implemented on the PyTorch platform under the Ubuntu 18.04 system, using a 16-core Intel i9-11900K@3.50 GHz processor.

B. Ablation and Comparison Experiments.

We compare our proposed adaptive complex wavelet informed Former with different style deep neural networks for image classification. Specifically, it includes the transformer neural network with corresponding parameter magnitudes like DeiT-Tiny/Small [49], ViT [39], GFNet-Tiny/Small [13] and the neural operator network AFNO-Tiny/Small [50]. In addition, ViT-base and GFNet-Base with larger parameter magnitudes are also compared.

First, we conduct comparative experiments with various Transformer-type architectures that have recently received much attention. We design multiple layer-matched complex wavelet informed operator models for the widely used visual Transformer model ViT for fairness comparison. The results in Table II indicate that no matter for the lightweight DeiT-tiny or the small DeiT-Small model, the complex wavelet informed Former network we proposed can improve the classification accuracy with less parameter growth cost. At the same time, in order to verify the effectiveness of the large model, we set the ACWI-Small/16 model to compare with the ViT-Base/16 model that trained with large-scale JFT [62] datasets. It can be seen that our model obtains higher Top-1 accuracy of 81.43% than others.

Second, we compare the proposed method with Transformer structures based on Fourier operator learning. For example, compared to GFNet, which learns long and short spatial dependencies in the Fourier frequency domain, our model achieves both an improvement in accuracy and a reduction in parameter memory at both the tiny and small scales. For ACWI-small with less parameter amount as GFNet-base, the accuracy improvement is significant, reaching 81.43%. Additionally, we compare our method with AFNO networks based on a principled foundation of operator learning. In the case of the same number of parameters, our wavelet neural operator learning has achieved higher classification accuracy due to the advantages of multi-scale shrinkage and richer orientation attention.

Thirdly, the ablation comparison of the CWI-Tiny/16 and ACWI-Tiny/16 models shows that the adaptive local block attention mechanism improves the model's classification accuracy while reducing the computational complexity. Similarly, the above rules also hold for CWI-Small/16 and ACWI-Small/16. In addition, comparing ACWI models of different magnitudes based on the dataset ImageNet-1 k, it is found that the model's performance gradually improves with the increase of the model and parameter scale.

C. Transfer Learning

To validate the general applicability of our method, we evaluate the model's transfer learning performance on downstream benchmarks and extension performance on dynamic video tasks.

TABLE III
THE DETAIL COMPOSITION OF DATASETS

Dataset	Training Set	Evalution Set	Testing Set	Classes	Average images per class	Resolution
ImageNet-1K	1,281,167	50,000	100,000	1,000	1,281	224^2
Flower-102	1020	1020	6129	102	40∼250	500^2
Stanford-Car	8144	–	8041	196	40∼50	360×240
CIFAR-10	50,000	10,000	10,000	10	5,000	32^2
CIFAR-100	50,000	10,000	10,000	100	500	32^2

TABLE IV
COMPARISON OF EXPERIMENTAL RESULTS OF TRANSFER LEARNING TO MULTIPLE DATASETS WITH THE EVALUATION INDICATOR ACC

Method	Params	Flower-102 [53]	Stanford-Car [54]	CIFAR-10 [55]	CIFAR-100 [56]
ViT-Base/16 [39]	86M	89.5%	–	98.1%	87.1%
ResNet50 [65]	26M	96.2%	90.0%	–	–
GFNet-Tiny	7M	97.4%	91.5%	97.6%	86.9%
GFNet-XS [13]	16M	98.1%	92.8%	98.6%	89.1%
GFNet-Small	25M	98.3%	92.9%	98.7%	89.5%
ResMLP-12 [64]	15M	97.4%	84.7%	98.1%	87.0%
ResMLP-24 [64]	30M	97.9%	89.5%	98.7%	89.5%
ACWI-Tiny/16	5M	97.8%	92.2%	98.3%	87.3%
ACWI-Small/16	22M	98.2%	92.8%	98.7%	89.7%

Transfer learning to other classification datasets: To validate the transfer learning capability of the proposed architecture [63], we first evaluate the transfer performance on several commonly used downstream public datasets for visual classification. The datasets we choose are Flowers-102, Stanford Cars, CIFAR-10, and CIFAR-100. Details of the dataset, such as sample size, type, distribution characteristics, and resolution, are presented in Table III.

Our models have initialized with ImageNet-1 k pre-trained weights and fully fine-tuned on the respective datasets on the new target dataset. We evaluate the transfer learning performance of ACWI models on multiple scales. As shown in Table IV, the experimental results show that our ACWI-tiny and ACWI-Small models achieve better results than regardless of the classic convolutional supervised learning structure ResNet or the Transformer model ViT-Base/16 with a larger parameter scale. And ViT-Base/16 for better Top-1 and Top-5 accuracy. The transfer learning process only uses ImageNet-1k [51] as the pre-training set. The final accuracy is obtained by fine-tuning the target dataset. Our model achieves higher accuracy with fewer computational parameters than ResMLP [64] and GFNet [13]. Specifically, ACWI-Tiny shows a 0.4% to 0.7% accuracy improvement over GFNet-Tiny while reducing parameter count by 2 M. Similarly, ACWI-Small demonstrates comparable accuracy to GFNet-Small but with 3 M fewer parameters. In fact, all our models basically achieve good performance on downstream datasets. Therefore, our model can also achieve flexible transfer adaptation for target small datasets with small data volumes that do not have the conditions to train visual deformers from scratch.

Fine-tuning at higher resolution: The visual transform model is often blinded by lack of theoretical basis when adjusting the input resolution and the corresponding patch size. The adaptive wavelet neural operator learning we propose uses the stacking of multi-layer wavelet size functions to more flexibly and adaptively shrink and focus on features of different scales. We take the fine-tuning from $224\times$ 224 input resolution in ImageNet-1k to a higher 500×500 input resolution in in Flower-102 as an example. Our model also achieves better accuracy performance at higher resolution inputs with fine-tuning uniformly.

Transfer learning to dynamic video tracking: To verify the scalability of our adaptive complex wavelet informed transformer operator, we use ACWI-Former as the skeleton network to implement more dynamic video tasks. Here, the dynamic video object tracking task are used to demonstrate the application potential of our method. We take the object tracking task as an example to verify the application potential of our adaptive wavelet neural operator network in dynamic video scenes.

We utilize the tracking framework TrSiam and TrDiMP based on the Transformer structure in [43]. In these tracking frameworks, the multi-head self-attention modules are used to match the template, and then the frame features are to be tested. We replace these self-attention layers with our ACWI layers, recorded as ACWI-TrSiam and ACWI-TrDiMP. During training, we initialize the feature extraction module weights with the pre-trained baseline model and keep them fixed. Then fully fine-tune the ACWI module for feature matching in the classification and regression modules of the head.

Finally, we test the performance of the trained model on the widely used tracking benchmark dataset OTB [57] and UAV123 [58]. The specific results are reported in Table V. Clearly, the introduction of ACWI brings accuracy gains to the baseline trackers. Although the tracking efficiency is slightly reduced, it does not affect the real-time effect of tracking. We use

TABLE V
EXPERIMENTAL RESULTS OF ACWI ON THE VISUAL TRACKING BENCHMARKS WITH THE EVALUATION INDICATOR AUC AND FPS

Benchmark	TrSiam [43]	ACWI-TrSiam	TrDiMP [43]	ACWI-TrDiMP	CFSiam-DPSE [66]	ACWI-CFSiam	OSTrack [67]	ACWI-OSTrack
OTB	70.8%	71.0%	71.1%	71.3%	71.2%	71.2%	71.9%	72.1%
UAV123	67.4%	67.9%	67.5%	68.0%	68.2%	68.4%	68.3%	68.4%
fps	37.6	32.3	25.4	21.2	33.1	28.8	63.3	58.6

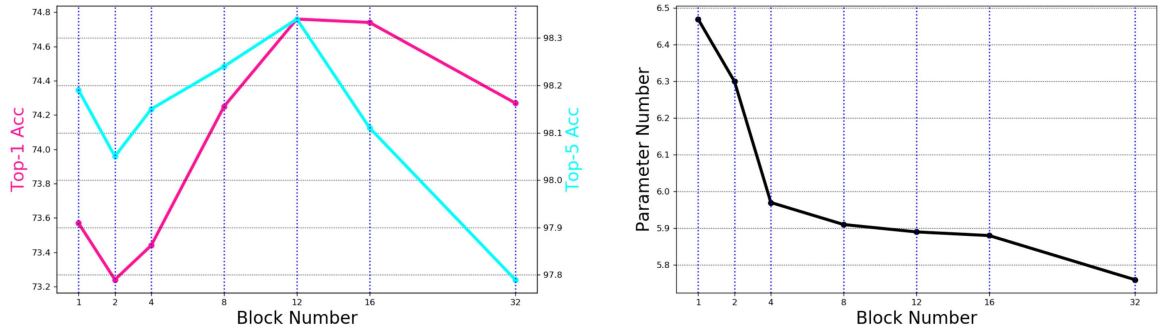

Fig. 4. Comparison of classification accuracy and parameter number on the ACWI-tiny model with different block number settings. The $BlockNumber$ refers to the number of weight sharing blocks for adaptive local block attention mechanism in the adaptive complex wavelet informed operator. Note that assuming $Block\,Number = N(1 < N \leq D)$, the channels share weights in units of (D/N) dimensions. In this case, the memory usage of the parameter matrices $\tau_{b_n}^k$ is $(D/N) \times (D/N)$, the memory usage of the component matrices $c_{b_n}^k$, $x_{b_n}^k$, etc. are $(W/2^k) \times (H/2^k) \times (D/N)$. Therefore, the larger the number N of local weight-sharing blocks in the channel domain is, the fewer model parameters are and the lower the amount of calculation is.

the fps (frames per second) indicator in the tracking task to make a unified comparison. The indicator fps in the table is measured on the OTB dataset. Experimental results show that, compared with the original TrSiam and TrDiMP trackers, the embedding of wavelet operators reduces the fps by 5.3 and 4.2 frames respectively. The overall speed is kept above 20 fps, maintaining the real-time performance of the tracker. The above results show that our adaptive wavelet neural operator can also benefit dynamic video tasks.

D. Analysis and Visualization

We conduct experiments to investigate how different components of ACWI affect performance. Furthermore, we visualize the wavelet operator and the learned features to better understand the learning process.

Adaptive Local Block Attention Structure: Based on the tiny model ACWI-Tiny/16, we analyze the impact of different settings on network performance from the aspect of the adaptive number of diagonal blocks. When we configure the adaptive diagonal block multiplicative structure for the complex wavelet informed operator, the choice of the number of blocks will affect the model's parameter quantity and accuracy performance.

To make the comparison fair, we use the ACWI-Tiny/16 model as a benchmark and adjust the number of blocks between 1 and 32 while keeping other model parameters consistent to fine-tuning the network. The classification accuracy training from scratch with 100 epochs is measured on Cifar-10. The comparison results are shown in Fig. 4. It can be seen that as $Block\,Number$ increases, both Top-1 accuracy and Top-5 accuracy undergo several stages of change as shown in the figure, as follows:

a) When $Block\,Number$ changes from 1 to 2, the accuracy drops significantly. This phenomenon might be related to the layer normalization operation implemented in the wavelet informed operator. When there is no weight block sharing (i.e., $Block\,Number = 1$), the conventional layer normalization is unaffected. However, when weight block sharing is applied in the channel domain (i.e., $Block\,Number > 1$), it may interfere with the globality of layer normalization, thereby impacting classification accuracy.

b) In the process of increasing $Block\,Number$ from 2 to 12, the accuracy climbs to a maximum value. This improvement occurs because, with an increased number of weight-sharing blocks, the number of free parameters in the adaptive complex wavelet informed operator layer decreases, resulting in a slightly more compact model. After 100 epochs of uniform training, this more streamlined weight-sharing model optimizes more quickly, achieving higher classification performance.

c) Subsequently, as $Block\,Number$ continued to increase from 12, the classification accuracy starts to decline. This decline happens because an excessive number of weight-sharing blocks increases the probability of losing specific feature information between channels, thereby limiting the network's representational capacity and causing a drop in classification accuracy. The model achieves peak accuracy at 12 blocks, which is close to the number of complex wavelet filter banks.

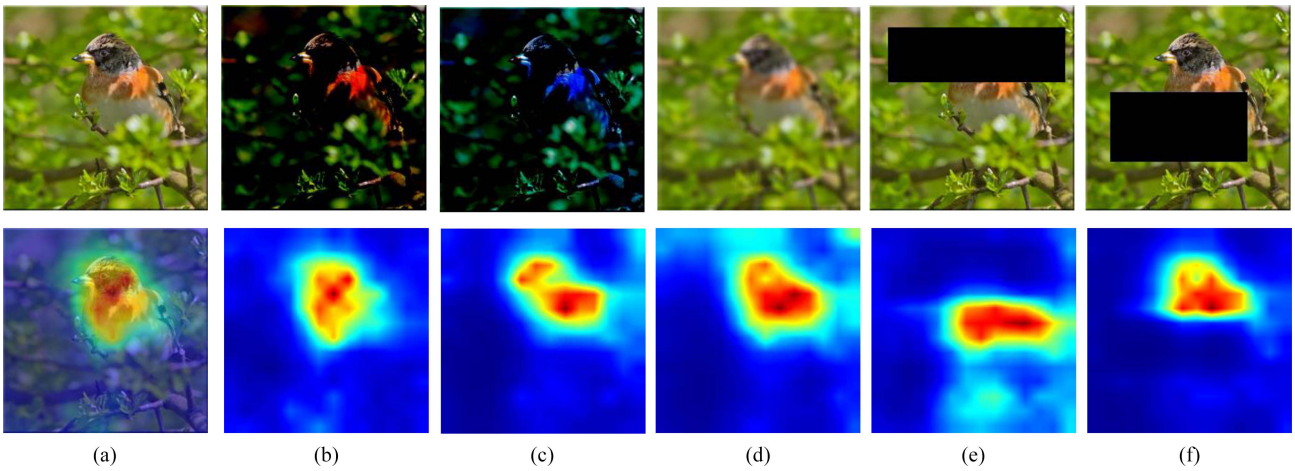

Fig. 5. Visual comparison of images and response maps under various interference transformations. (a) shows the original input image and its corresponding response map. (b) shows the image and its response map after normalization, with the mean and variance of each channel set to 0.5. (c) shows the image and its response map after color jitter, with brightness, contrast, saturation, and hue all set to 0.5. (d) shows the image and its response map after Gaussian blur, with a Gaussian kernel size of 7×7. (e) And (f) show the images and their response maps after random erasure.

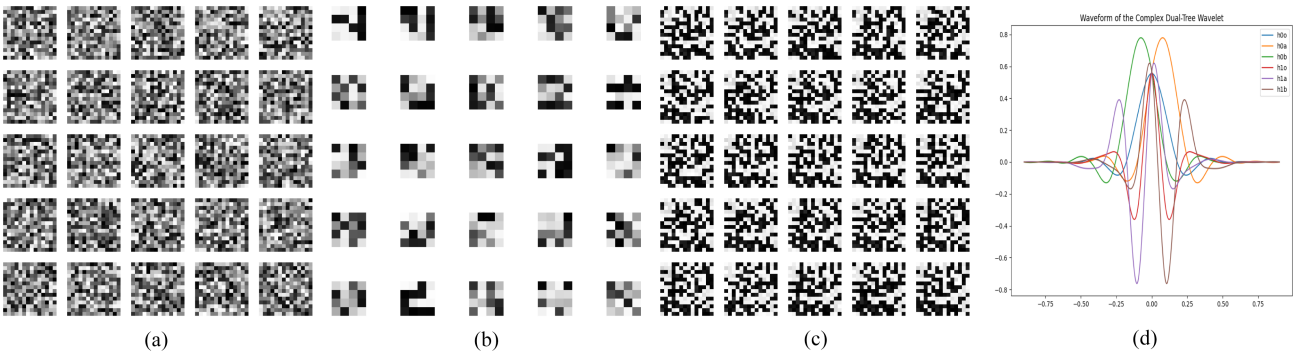

Fig. 6. Visualization of the learning process of adaptive complex wavelet informed transformer operator layer. (a) Input features of the adaptive complex wavelet informed operator layer. (b) Decomposition features of the wavelet domain. (c) Reconstruction features back to the image source domain after operators such as AWB, ABM, and BM. (d) The curve of the initialization wavelet transform basis for the dual-tree complex wavelet.

Moreover, the model memory decreases as the size of the block increases. It can be seen that 12 blocks achieve the best Top-1 accuracy. Therefore, we set the number of blocks to 12 in this work.

Sparsity Threshold: In addition, we explore the range of values for the sparse threshold. We gradually increase the sparsity threshold from 0 to 10. We first weight-initialize the ACWI-small with pre-trained weights, and then fine-tune the sampled to achieve classification. If the sparsity threshold is set to 0, it corresponds to no sparsity. Based on the experimental accuracy scores on the validation set of CIFAR10 data set, it is shown that the Top-1 accuracy peaks around $= 0.01$, demonstrating the effectiveness of sparsity.

Robustness verification: To investigate the robustness of our proposed method, we applied various perturbations such as normalization, color jitter, Gaussian blur, and random erasure to the image (This image demo selected from the ImageNet validation set). After the input samples undergo the aforementioned transformations, the corresponding output response results of the adaptive complex wavelet informed operator network are shown in Fig. 5. The visual response maps show that the input subjected to different disturbances can still focus and locate the target area well, thus robustly capturing key feature information in the image. We believe that the proposed method has a certain robustness to the above different disturbances.

Operator learning visualization: To understand how our wavelet informed operator learns and processes image features, we analyze its learned feature representations through the ACWI. Fig. 6 shows the principal component features learned by the adaptive complex wavelet informed operator network based on the Transformer framework (ACWI-Former). It can be seen that these components are equivalent to the representation of wavelet basis functions in different frequency-phase combination structures in image patches.

Comparing Fig. 6(a) and (c), it can be seen that after passing through the wavelet informed transformer operator network layer, the features have prominent geometric structure characteristics and are more sparse. This also shows that the ACWI

can sparsely capture the features related to classification semantics in the image. Furthermore, compared with the acquired feature maps, the attention of regions in different channels tends to focus on the characteristics of the channel's frequency band. Moreover, we also visualize the initialization wavelet transform basis of the dual-tree complex wavelet, as shown in Fig. 6(d) of this text. h0o and h1o are used in the 1st-level of decomposition. They separate the input signal into low-frequency and high-frequency components. h0a and h1a are used in 2rd-levels of decomposition in one tree of the dual-tree structure. h0b and h1b are used in subsequent levels of decomposition in the other tree of the dual-tree structure.

Limitations: The main limitations of this work are as follows: 1) Since wavelet decomposition and reconstruction are not a parallel process, more levels of decomposition and reconstruction will increase the time cost of network learning. 2) The two-dimensional wavelet transform used in this paper only decomposes and reconstructs along two orthogonal directions, and further exploration is still needed in the decomposition of finer-grained directions.

V. Conclusion

In this paper, we propose a complex wavelet informed transformer operator learning paradigm and build its deep adaptive complex wavelet informed Former network. Combining wavelet transform with attention kernel theoretically improves the transformer's ability to capture sparse multi-resolution and multi-directional structured features for images. Our experiments show that our approach shows good generalization potential on multiple classification benchmarks and video tracking tasks compared to existing state-of-the-art algorithms. It also serves as an important exploration inspiration for future work, such as exploring the combination of multi-wavelet bases in neural networks and using multi-scale geometric transformation to capture singular features more delicately.

References

[1] K. Han et al., "Transformer in transformer," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 15908–15919.

[2] Y. Liu et al., "A survey of visual transformers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7478–7498, Jun. 2024.

[3] Q. Wen et al., "Transformers in time series: A survey," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, 2023, pp. 6778–6786.

[4] R. Girdhar and K. Grauman, "Anticipative video transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13485–13495.

[5] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–12.

[6] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 280–296.

[7] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 22419–22430.

[8] H. Zhou et al., "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 11106–11115.

[9] T. Zhou et al., "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting," in *Proc. Int. Conf. Mach. Learn.*. 2022, pp. 27268–27286.

[10] S. Liu et al., "Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–20.

[11] T. Xiao et al., "Early convolutions help transformers see better," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 30392–30400.

[12] Y. Rao et al., "DynamicViT: Efficient vision transformers with dynamic token sparsification," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 13937–13949.

[13] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 980–993.

[14] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontanon, "FNet: Mixing tokens with Fourier transforms," in *Proc. 2022 Conf. North Amer. Chapter Ass. Comput. Linguistics: Human Lang. Technol.*, 2022, pp. 4296–4313.

[15] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian, "A Fourier-based framework for domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14378–14387.

[16] C. Mayer et al., "Transforming model prediction for tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8721–8730.

[17] H. K. Cheng and A. G. Schwing, "XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 640–658.

[18] M. Stoiber, M. Sundermeyer, and R. Triebel, "Iterative corresponding geometry: Fusing region and depth for highly efficient 3D tracking of textureless objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 6845–6855.

[19] A. Anandkumar et al., "Neural operator: Graph kernel network for partial differential equations," in *Proc. ICLR 2020 Workshop Integration Deep Neural Models Differ. Equ.*, 2020, pp. 1–21.

[20] L. Lu, P. Jin, and G. E. Karniadakis, "DeepONet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators," 2019, *arXiv:1910.03193*.

[21] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis, "Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators," *Nat. Mach. Intell.*, vol. 3, no. 3, pp. 218–229, 2021.

[22] V. May, Y. Keller, N. Sharon, and Y. Shkolnisky, "An algorithm for improving non-local means operators via low-rank approximation," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1340–1353, Mar. 2016.

[23] B. Lütjens, C. H. Crawford, C. D. Watson, C. Hill, and D. Newman, "Multiscale neural operator: Learning fast and grid-independent PDE solvers," in *Proc. Int. Conf. Mach. Learn. 2022 2nd AI Sci. Workshop*, 2022, pp. 1–20.

[24] Y.-H. H. Tsai, S. Bai, M. Yamada, L.-P. Morency, and R. Salakhutdinov, "Transformer dissection: A unified understanding of transformer's attention via the lens of kernel," in *Proc. Conf. Emp. Methods Natural Lang. Process.*, 2019, pp. 1–10.

[25] Z. Li et al., "Fourier neural operator for parametric partial differential equations," 2020, *arXiv:2010.08895*.

[26] S. Cao, "Choose a transformer: Fourier or Galerkin," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24924–24940.

[27] L. Chi, B. Jiang, and Y. Mu, "Fast Fourier convolution," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 4479–4488.

[28] N. Kovachki et al., "Neural operator: Learning maps between function spaces," *J. Mach. Learn. Res.*, vol. 24, no. 89, pp. 1–97, 2023.

[29] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[30] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 548–558.

[31] X. Lin et al., "EAPT: Efficient attention pyramid transformer for image processing," *IEEE Trans. Multimedia*, vol. 25, pp. 50–61, 2023.

[32] P. Liu et al., "Spectrum-driven mixed-frequency network for hyperspectral salient object detection," *IEEE Trans. Multimedia*, vol. 26, pp. 5296–5310, 2024.

[33] T. Tripura and S. Chakraborty, "Wavelet neural operator: A neural operator for parametric partial differential equations," 2022, *arXiv:2205.02191*.

[34] A. Grinsted, J. C. Moore, and S. Jevrejeva, "Application of the cross wavelet transform and wavelet coherence to geophysical time series," *Nonlinear Processes Geophys.*, vol. 11, no. 5/6, pp. 561–566, 2004.

[35] L. Zhang et al., "Wavelet knowledge distillation: Towards efficient image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12454–12464.

[36] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 17864–17875.
[37] K. Xian, J. Peng, Z. Cao, J. Zhang, and G. Lin, "ViTA: Video transformer adaptor for robust video depth estimation," *IEEE Trans. Multimedia*, vol. 26, pp. 3302–3316, 2024.
[38] N. Carion et al., "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*. Springer, 2020, pp. 213–229.
[39] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[40] Z. Zhang, Y. Liu, X. Wang, B. Li, and W. Hu, "Learn to match: Automatic matching network design for visual tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13319–13328.
[41] B. Yu et al., "High-performance discriminative tracking with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9836–9845.
[42] T. Yao, Y. Pan, Y. Li, C.-W. Ngo, and T. Mei, "Wave-ViT: Unifying wavelet and transformers for visual representation learning," in *Proc. Eur. Conf. Comput. Vis.*. Springer, 2022, pp. 328–345.
[43] N. Wang, W. Zhou, J. Wang, and H. Li, "Transformer meets tracker: Exploiting temporal context for robust visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1571–1580.
[44] S. Sun, S. Zhi, J. Heikkilä, and L. Liu, "Evidential uncertainty and diversity guided active learning for scene graph generation," in *Proc. 11th Int. Conf. Learn. Representations*, 2023, pp. 1–22.
[45] S. Sun, S. Zhi, Q. Liao, J. Heikkilä, and L. Liu, "Unbiased scene graph generation via two-stage causal modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12562–12580, Oct. 2023.
[46] G. Gupta, X. Xiao, and P. Bogdan, "Multiwavelet-based operator learning for differential equations," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24048–24062.
[47] Z. Liu et al., "Render unto numerics: Orthogonal polynomial neural operator for PDEs with non-periodic boundary conditions," 2022, *arXiv:2206.12698*.
[48] Q. Li et al., "Transparent operator network: A fully interpretable network incorporating learnable wavelet operator for intelligent fault diagnosis," *IEEE Trans. Ind. Inform.*, vol. 20, no. 6, pp. 8628–8638, Jun. 2024.
[49] H. Touvron et al., "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*. 2021, pp. 10347–10357.
[50] J. Guibas et al., "Adaptive fourier neural operators: Efficient token mixers for transformers," 2021, *arXiv:2111.13587*.
[51] Y. You, Z. Zhang, C.-J. Hsieh, J. Demmel, and K. Keutzer, "ImageNet training in minutes," in *Proc. 47th Int. Conf. Parallel Process.*, 2018, Art. no. 1.
[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[53] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 2008 6th Indian Conf. Comput. Vis., Graph. Image Process.*. 2008, pp. 722–729.
[54] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2013, pp. 554–561.
[55] R. C. Çalik and M. F. Demirci, "Cifar-10 image classification with convolutional neural networks for embedded systems," in *Proc. 2018 IEEE/ACS 15th Int. Conf. Comput. Syst. Appl.*. 2018, pp. 1–2.
[56] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.
[57] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2411–2418.
[58] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 445–461.
[59] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 850–865.
[60] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6181–6190.
[61] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *Proc. 2009 IEEE Conf. Comput. Vis. Pattern Recognit.* 2009, pp. 248–255.
[62] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 843–852.
[63] Z. Lu, H. Xie, C. Liu, and Y. Zhang, "Bridging the gap between vision transformers and convolutional neural networks on small datasets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 14663–14677.
[64] H. Touvron et al., "ResMLP: Feedforward networks for image classification with data-efficient training," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 5314–5321, Apr. 2023.
[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[66] X. Li et al., "A complex-former tracker with dynamic polar spatio-temporal encoding," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 12, pp. 17344–17358, Dec. 2024.
[67] B. Ye, H. Chang, B. Ma, S. Shan, and X. Chen, "Joint feature learning and relation modeling for tracking: A one-stream framework," in *Proc. Eur. Conf. Comput. Vis.*. Springer, 2022, pp. 341–357.

Xiaotong Li received the B.S. degree in electronic information engineering from the Harbin Engineering University, Harbin, China, in 2017, and the Ph.D. degree in computer science and technology from Xidian University, Xi'an, China, in 2024. She is currently a postdoctoral Researcher with the School of Artificial Intelligence, Xidian University. Her main research interests include deep learning, computer vision and remote sensing image analysis and understanding.

Licheng Jiao (Fellow, IEEE) received the B.S. degree from Shanghai Jiao Tong University, Shanghai, China, in 1982, and the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively. Since 1992, he has been a Professor with Xidian University, Xi'an, China. Now, he is a distinguished Professor with the School of Artificial Intelligence, Xidian University, Xi'an. He is currently the Director of the Key Laboratory of Intelligent Perception and Image Understanding, which is affiliated with the Ministry of Education of China. He has been a member of the Academia Europaea. His research interests include machine learning, deep learning, natural computation, remote sensing, image processing, and intelligent information processing. Prof. Jiao is the chairperson of the Awards and Recognition Committee in IEEE Xi'an Branch, the 6th and 7th Vice Chairperson of the Chinese Association of Artificial Intelligence, the chairperson of the Asian Society for Computational Intelligence, the Fellow of IET/CAAI/CIE/CCF/CAA/CSIG/AAIA/AIIA/ACIS, a councilor of the Chinese Institute of Electronics, a committee member of the Chinese Committee of Neural Networks, an expert of the Academic Degrees Committee of the State Council, and an ESI highly cited scientist.

Fang Liu (Senior Member, IEEE) received the B.S. degree in computer science and technology from Xi'an Jiaotong University, Xi'an, China, in 1984, and the M.S. degree in computer science and technology from Xidian University, Xi'an, in 1995. She is currently a Professor with Xidian University. She has authored or coauthored five books and more than 80 papers. Her research interests include image perception and pattern recognition, machine learning, evolutionary computation. Prof. Liu was the recipient the second prize of the National Natural Science Award in 2013.

Shuyuan Yang (Senior Member, IEEE) received the B.A. degree in electrical engineering and the M.S. and Ph.D. degrees in circuit and system from Xidian University, Xi'an, China, in 2000, 2003, and 2005, respectively. She has been a Professor of artificial intelligence with Xidian University. Her research interests include machine learning and image processing.

Hao Zhu (Member, IEEE) received the B.S. degree in physics and photoelectricity engineering and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, China, in 2013 and 2019, respectively. He is currently an Associate Professor with the school of Artificial Intelligence and the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University. His current research interests include deep learning, remote sensing image interpretation, and evolutionary computation.

Xu Liu (Senior Member, IEEE) received the B.S. degree in mathematics and applied mathematics from North University of China, Taiyuan, China in 2013. He received the Ph.D. degree from Xidian University, Xi'an, China, in 2019. He is currently an Associate Professor of Huashan elite and postdoctoral Researcher of Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University. His current research interests include machine learning and image processing.

Lingling Li (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2011 and 2017, respectively. From 2013 to 2014, she was an exchange Ph.D. student with the Intelligent Systems Group, Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, Bilbao, Spain. She is an Associate Professor with the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education, School of Artificial Intelligence, Xidian University. Her research interests include image processing, deep learning, and pattern recognition.

Wenping Ma (Senior Member, IEEE) received the B.S. degree in computer science and technology and the Ph.D. degree in pattern recognition and intelligent systems from Xidian University, Xi'an, China, in 2003 and 2008, respectively. Since 2006, she has been with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, Xidian University, where she is currently an Associate Professor. Her research interests include intelligent computing and image processing.