Visualizing and Understanding Patch Interactions in Vision Transformer

Jie Ma, Yalong Bai, Bineng Zhong, Wei Zhang, *Member, IEEE*, Ting Yao, *Senior Member, IEEE*, and Tao Mei, *Fellow, IEEE*

***Abstract*— Vision transformer (ViT) has become a leading tool in various computer vision tasks, owing to its unique self-attention mechanism that learns visual representations explicitly through cross-patch information interactions. Despite having good success, the literature seldom explores the explainability of ViT, and there is no clear picture of how the attention mechanism with respect to the correlation across comprehensive patches will impact the performance and what is the further potential. In this work, we propose a novel explainable visualization approach to analyze and interpret the crucial attention interactions among patches for ViT. Specifically, we first introduce a quantification indicator to measure the impact of patch interaction and verify such quantification on attention window design and indiscriminative patches removal. Then, we exploit the effective responsive field of each patch in ViT and devise a window-free transformer (WinfT) architecture accordingly. Extensive experiments on ImageNet demonstrate that the exquisitely designed quantitative method is shown able to facilitate ViT model learning, leading the top-1 accuracy by 4.28% at most. More remarkably, the results on downstream fine-grained recognition tasks further validate the generalization of our proposal.**

***Index Terms*— Classification, explainable visualization, patch interaction, vision transformer (ViT).**

I. Introduction

TRANSFORMER architecture has led to the revolutionizing of the natural language processing (NLP) field and inspires the emergence of transformer-type works on learning word [1] and character [2] level representations with self-attention mechanisms [3] for capturing dependency syntax [4] and grammatical [5] relationships. This has also motivated the recent works of vision transformers (ViTs) for vision tasks by using multi-head self-attention (MSA) and multi-layer perceptrons, which are shown able to perform well on ImageNet classification and various downstream tasks, such as object detection [6], [7], [8], [9], [10], semantic segmentation [11], [12], [13], [14], image captioning [15], [16], and so on.

Different from convolutional neural networks [17], [18] which focus on local receptive fields, transformer-based architecture [19] utilizes patch-wise attention mechanism for full-patch information interactions with dynamic receptive fields [20]. As a general ViT backbone, ViT is capable of global feature extraction by dense information aggregation and interactions among full patch tokens. Several previous methods [3], [21] have interpreted how classification outputs are formed in ViT models. Nevertheless, these visualization schemes mainly focus on the analysis of attention mechanism in discriminative patch selection or feature representations visualization, but remains unclear on the actual scope of 1-to-N patch attention. The valid question then emerges as is there redundancy in global self-attention? Recently, there are some variations [22], [23], [24], [25] of ViT with heuristic configurations demonstrated the restricted attention range/window/region for patches can reduce redundant but without performance degradation. Thus, the analysis of information interactions among patches during global self-attention become increasingly important, since it would play a crucial role in specifying the boundary of efficient attention scope, precisely dropping the indiscriminative patches or patch-wise connections, and eventually guiding the visual attention model design.

To this end, we seek to obtain a better understanding of ViT models, especially the information interactions between patches. The problem of understanding ViT presents various challenges due to its inherent architecture complexity. Specifically, the input image embedding features are learned across multiple layers and utilize self-attention mechanism, that expresses an independent image patch output embedding feature as a convex combination of all patches embedding features. Meanwhile, self-attention leverages multiple attention heads that operate independently.

In this work, we propose a novel explainable visualization approach to analyze and interpret patch-wise interactions via quantifying the reliability of patch-to-patch connections, as shown in Fig. 1. Specifically, we propose a method for quantifying the impact of patch-wise attentions, in which patch tokens often gather information from their related high-impact patches. In this way, we can briefly highlight the boundary of interactive regions for each patch. To verify the effectiveness of our quantification, we propose a novel adaptive attention window design schema yielding to the

Manuscript received 13 October 2022; revised 27 February 2023; accepted 19 April 2023. This work was supported in part by the Project of Guangxi Science and Technology under Grant 2022GXNSFDA035079, in part by the National Natural Science Foundation of China under Grant 61972167 and Grant U21A20474, in part by the Guangxi "Bagui Scholar" Teams for Innovation and Research Project, in part by the Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing, in part by the Guangxi Talent Highland Project of Big Data Intelligence and Application, and in part by the Research Project of Guangxi Normal University under Grant 2022TD002. *(Jie Ma and Yalong Bai contributed equally to this work.) (Corresponding author: Bineng Zhong.)*

Jie Ma and Bineng Zhong are with the Key Laboratory of Education Blockchain and Intelligent Technology, Ministry of Education, and the Guangxi Key Laboratory of Multi-Source Information Mining and Security, Guangxi Normal University, Guilin 541004, China (e-mail: bnzhong@gxnu.edu.cn).

Yalong Bai and Wei Zhang are with the JD Explore Academy, Beijing 100010, China.

Ting Yao and Tao Mei are with the HiDream.ai, Beijing 100190, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3270479.

Digital Object Identifier 10.1109/TNNLS.2023.3270479

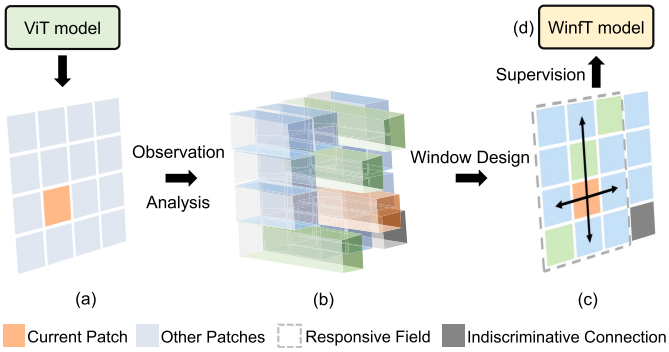

Fig. 1. Given a patch (yellow) in (a), we first analyze and quantify the impact of interactions between the current patch and all other patches. According to our patch interaction quantitative results, we observe the distributions of relevant patches for current patch in (b), figure out the boundary of responsive field that containing patches with critical information to current patch in (c). Then, we apply the responsive field for guiding ViT model training. The significant performance improvement from ViT to WinfT in (d) can be regarded as a strong posteriori proof of the rationality of our proposed explainable visualization schema for ViT.

interactive region boundary for each patch. The experimental results show that this schema can improve ViT performance with reducing a large number of attention operations outside the attention window. Meanwhile, we observe that some indiscriminative patches provide their contextual information to all other patches consistently. Therefore, we design a mining schema for dropping these indiscriminative patches. To further understand the patch interactions, we define the responsive field for the patch and make statistic analysis on it. We find that the responsive field for each patch presents semantic relevance. This further motivated us to propose a window-free transformer (WinfT) architecture by incorporating the supervisions of responsive field. Correspondingly, the stable and significant performance improvements from window-free transformer further demonstrate the effectiveness of our analysis. The main contributions of this work are summarized below as follows.

1) We propose a novel explainable visualization schema to analyze and interpret the crucial attention interactions among patches for ViT. For verifying the rationality of our visualization schema, we apply it to guide the attention window design and results in indeed performance improvement with significantly reducing the computational complexity.
2) Based on the quantification of the impact of patch interaction, we figure out the existence of indiscriminative patches in images for ViT model training. Further, dropping these patches also benefit the ViT model.
3) We define responsive field for providing a crucial understanding of ViT. The statistic analysis shows that both the size and tendency of the informative attention window for each patch are semantic-oriented.
4) Inspired by the above observations and analysis, we propose a novel WinfT architecture with predictive and adaptive attention window to restrict the patch-wise interactions. The experimental results in ImageNet show that our window-free transformer can improve 2.56% top-1 accuracy while reducing nearly 58.9% of patch-wise attention operations in average across various ViT structures with different input resolutions.

II. Related Work

A. Transformer-Based Vision Models

Transformer-based vision models aim to utilize the attention mechanism to learn global dependencies representations. ViT [19] is the first convolution-free transformer-based vision architecture, which applied attention to a sequence of fixed-size non-overlapping patches. This architecture is beneficial for exploring global contextual information and achieves high performance on downstream tasks. Masked image modeling [26], [27] adopt a random mask on the input tokens to generate latent representations, which lacks interpretable semantic analysis. Various transformer-based vision models [22], [23], [24], [28], [29] present the attention mechanism through heuristic patch-wise interactions, in particular, focusing on different structures to enhance the patch information interactions for effective attention computation. Beyond that, DynamicViT [25] proposes a dynamic token sparsification to prune the tokens. These models highlight the importance of information interactions, typically through heuristic patch-wise interactions. Different from the above approaches for patch-wise interactions, our approach aims to propose adaptive patch-wise interactions for attention mechanism.

1) Explainability for ViT: Given the key role of explainability in deep learning, several works have analyzed the gradients [30], [31], [32], [33], [34] and attribute propagation [35], [36] in a convolutional neural network to generate an understanding of representations to explain specific assumptions. Recent transformer works mainly [37], [38], [39], [40] focus on analyzing and interpreting attention scores to understand why the model performs so well. Reuse transformer [37] highlights the relevant relationship between the different layers and captures the similarity to reuse the attention score. Voita et al. [41] apply layer-wise relevance propagation to consider the different relevance of multihead attention block. These works are focused on visually understanding individual attention scores, and provide an explanation for understanding attention mechanisms. However, there are two limitations to understanding ViT models: 1) they do not highlight the relevance of patch-to-patch connection and 2) focus on the individual model output, or attention scores, does not directly interpret the patch-wise interactions well. We aim to provide an explainable visualization schema for ViT, therefore it's possible to analyze and interpret the interactions among patches.

III. Patch Interactions Visualization

We study the patch interactions on global MSA of the original visual transformer [19], so we first present a brief background of ViT and describe our proposed visualization method for quantifying the magnitudes of patch-to-patch connections in ViTs.

A. Preliminaries

Given an input image of $H \times W$ resolution, ViT model (ViT) first splits it into a sequence of non-overlapping patches $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of fixed size P, and then transforms them into tokens by linear projection. For capturing the long-range dependencies among patches, the patch tokens are fed to stacked transformer encoders, each of which contains an

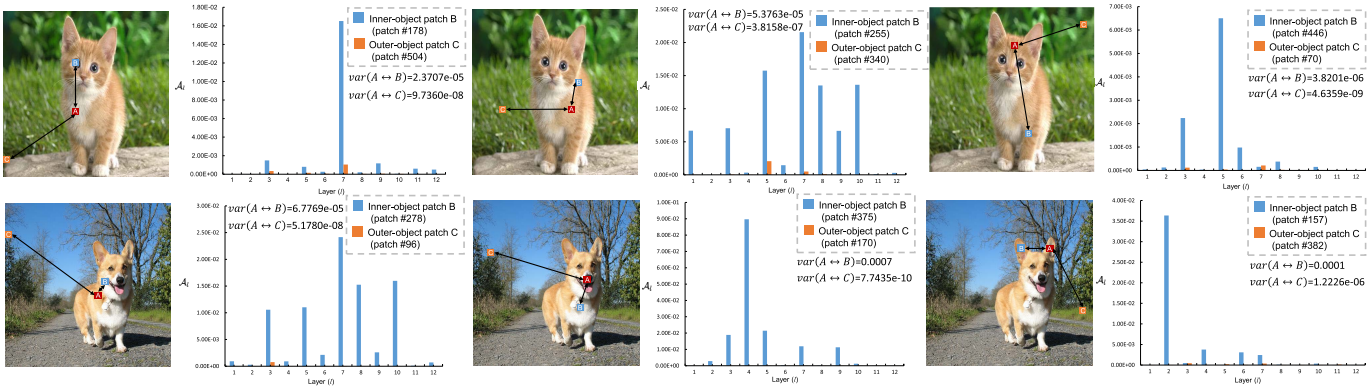

Fig. 2. Illustration about quantitative analysis between the inner-object patch and outer-object patch. Zoomed-in view for better visualization.

MSA mechanism, which concatenates multiple scaled dot-product attention modules. Specifically, the scaled attention modules (SA) first linearly projects the patch tokens to a query matrix Q, key matrix K, and value matrix V, and then computes the attention weight matrix $\mathcal{A}$ according to the patch-wise similarity between the query and key matrices

$$\mathcal{A} = \text{SoftMax}\left(QK^T/\sqrt{d}\right) \tag{1}$$

where d is the channel dimension of the query or key. By computing the sum over V weighted by row values $\mathbf{a}_p \in \mathbb{R}^{1\times N}$ in $\mathcal{A}$, information originating from different patch tokens get mixed for updating the representation of pth patch token. MSA is an extension of SA which concatenates k self-attention operations. We denote the k-head attention weight matrix in MSA as $\mathcal{A} \in \mathbb{R}^{k\times N\times N}$. The interactions or contextual information exchanging among different patches mainly depends on the attention weight matrix $\mathcal{A}$. Thus, in this work, we focus on the theoretical and empirical analysis of attention weight matrix $\mathcal{A}$, rather than the representation of each patch or image region that widely used in previous visual feature analysis methods.

B. Patch Interactions Quantification

The transformer-based architectures are ideally capable of learning global contextual information by leveraging a fully self-attention mechanism among all patches. Extensive works [23], [24], [25], [29] have demonstrated the existence of redundant computations in the patch attention mechanism through fixed-scale window design, pruning, and so on. However, these works usually focus on the spatial division or local region structuring on patches, while overlooked the reliability of the patch connections for self-attention mechanism. Meanwhile, through an analysis of similarity of attention scores by different layers and heads, the ability of interaction representation is not consistent [41], [42], [43], [44]. We believe that the reliability measure of patch connections by both structure and features would lead to a better understanding of the interactions among patches and further guide the attention mechanism design for ViT.

Many methods were suggested for generating a heatmap that indicates discriminative regions, given an input image and a CNN or transformer model. However, there are not many studies that explore the effectiveness of connections across image patches. Here, we start with visualization and statistic analysis on patch-wise interactions in ViT. As shown in Fig. 2, we randomly sampled target patch A (in red color) inner object for analysis of the patch interactions between the inner-object patch B (in blue color) and outer-object patch C (in orange color). Similar to the observation of "ViT model's highly dynamic receptive field" that was mentioned in previous work [20], we can find that, although the irrelevance patch pair (inner-outer object patch pair) consistently has a low response, the patch-to-patch attention score between relevance pair varies a lot across different layer, as the statistical results shown in Fig. 2. More intuitive examples can be found in Fig. 3(a), that the high attention areas cross all layers ($\mathcal{A}_i$ denotes the attention weight matrix in the ith transformer block) present an obvious uncertain of "tight" or "loose." The patch interaction strength generally presents a periodic "enhancing-fading" cross transformer blocks. This would be the potential cause of the flexible and dynamic receptive field of ViTs. Owing to the global information propagation among patch tokens in stacked MSA layers, a patch token can intensely gather representations from other patch tokens of high relevance to it. This results in an unstable attention score and high uncertainty among relevant patches across layers.

Thus, for a standard ViT model contains l MSA layer, all the above observations and analysis motivated us to measure the impact of patch interactions by estimating the uncertainty of attention score among them across all transformer blocks

$$\mathcal{U} = \frac{1}{l}\sum_{i=1}^{l}(\mathcal{A}_i - \mathbb{E}_{\mathcal{A}})^2. \tag{2}$$

Thanks to the uniform head size setting of the ViT model, we can directly compute $\mathbb{E}_{\mathcal{A}} \in \mathbb{R}^{k\times N\times N}$ as the mean of attention scores across all k-head self-attention layers. $\mathbb{E}_{\mathcal{A}}$ can be regarded as a k-channel score map depicting the uncertainty of interactions among patches. As we show in Fig. 3(a), although most of the high uncertainty region are gathered around the corresponding patch, different channels of score map results in various ranges of uncertainty. Moreover, averaging the score map $\mathcal{U}$ cross channel (denoted as $\mathbb{E}_U$) results in signal attenuation of border areas. Thus, we apply another uncertainty estimation on $\mathcal{U}$ to quantify the various uncertainty cross k channel, and highlight the boundary of

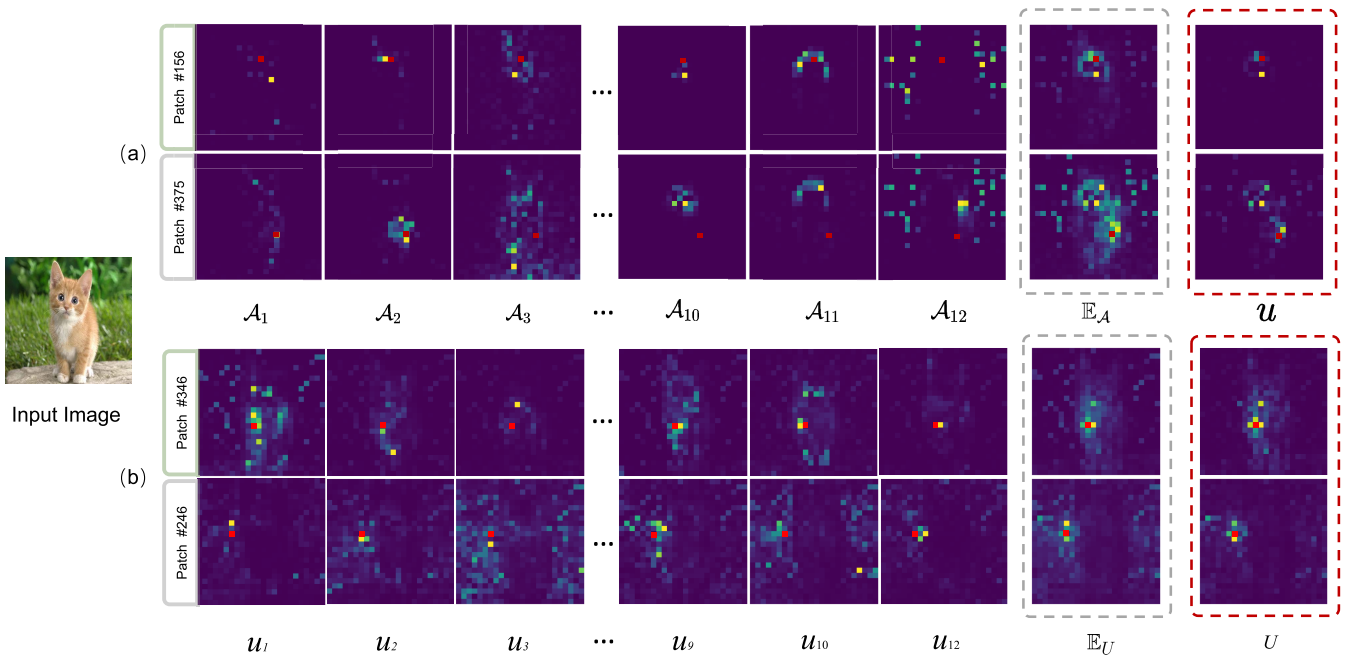

Fig. 3. Visualization of attention score maps $\mathcal{A}$ and its uncertainty across different layers $\mathcal{U}$. One 384 × 384 resolution images are fed into the well-trained ViT-B/16 [19] with 12 heads, 12 layers, and 16 × 16 patch size. We randomly sampled four patches (in red color) of the input image for visualization. (a) Visualization of attention score maps $\mathcal{A}$. (b) Visualization of uncertainty across layers $\mathcal{U}$.

interactive regions for each patch as follows:

$$U = \frac{1}{k}\sum_{i=1}^{k}(\mathcal{U}_i - \mathbb{E}_U)^2 \tag{3}$$

where $\mathcal{U}_i$ is the $N \times N$ score map extracted from the ith channel of $\mathcal{U}$. We visualize the row values $\mathbf{u}_p$ in U for the given image patch $\mathbf{x}_p$ in Fig. 3(b). It can be found that the most interactive regions of the current patch are in the surrounding area. There are also some connections of distant patches, which are usually relevant to the background around the main object in the image. These visualizations are also conform to the general knowledge that locality information play as a critical role for object recognition [24], [28], [45]. Inspired by these related works, we also proposed two U guided attention window design methods for ViT in Sections IV-A and V, respectively. In special, we restricted the interaction range for each patch during 1-to-N attention operations, by ignoring the patches outside the boundary of interactive regions in U. The experimental results show that U can well guide the ViT model training to focus on essential attention operation and lead to performance improvement while significantly decreasing the computational complexity. More details can be found in the sections below.

IV. Patch Interactions Analysis

Based on our quantification of patch-wise interactions, we figure out the existence of potentially indiscriminative patches for ViT model (Section IV-B), and proposed an adaptive attention window design method (Section IV-A) for further analyzing the redundancy of global attention mechanism and the responsive field of each patch in ViT model (Section IV-C).

A. Adaptive Attention Window Design

Attention mechanism is one of the core computations of the transformer-based model, requiring expensive quadratic calculations. Fixed-scale window/region design [22], [23], [24] and dynamic pruning structures [25] are approaches to address redundancy in attention computation. However, these approaches are extremely restrictive in the scale of the information interactions that need to be predefined artificially. Therefore, we propose a novel adaptive attention window design schema guided by the quantification result of patch-wise interactions for attention computation.

Considering U can not only measure the informative of patch, but also highlight the boundary of interactive region for patch, we rank all values in U, and select the top T elements with high value to construct a subset U'. After that, given the pth patch whose coordinate is $\langle x_p, y_p\rangle$ ($x_p = p\%(N)^{1/2}$, $y_p = p/(N)^{1/2}$), we can construct a window boundary candidate set B_p for it

$$B_p = \{\langle x_p, y_p\rangle\} \cup \{\langle x_i, y_i\rangle : u_{p,i} \in U'\}. \tag{4}$$

Subsequently, we select the maximum and minimum offset in the x- and y-axis in B_p, respectively, to finalize the attention window for x_p, denoted as $\{x_p^l, y_p^t, x_p^r, y_p^b\}$. Specifically, for the situation of $|B_p| = 1$, $B_p = \{\langle x_p, y_p\rangle\}$ that there is no relevant patch hitting in U', the global self-attention on x_p degenerates into an identical operation. As a result, there are

$$\mathcal{O} = \sum_{p=1}^{N}(x_p^r - x_p^l) \times (y_p^b - y_p^t) \tag{5}$$

patch-wise interactions for each head in each self-attention layer. We denote $\mathbb{E}_{\mathcal{O}}$ as the averaged number of patch interactions per head over all images in dataset.

As shown in Fig. 4, the original self-attention mechanism leverages non-overlapping image patches and then builds long-range interaction between all patches. Our proposed method provides the adaptive attention window in terms of the effective interactions for each patch. Moreover, adaptive

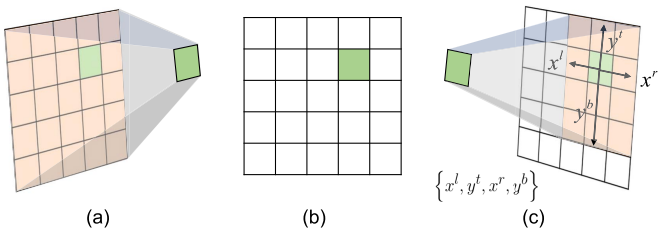

Fig. 4. Illustration about global self-attention mechanism (a) and our proposed adaptive window design (c). Green patch represents the current patch for attention operation. Yellow region indicates the range of patch interactions for current patch. (a) Self-attention. (b) Current Patch. (c) Proposed Method.

TABLE I

TOP-1 ACCURACY OF VIT ON IMAGENET BY ADAPTING ADAPTIVE WINDOW DESIGN (AWD) ON VARIOUS SETTINGS OF α. VIT-B/16 MODEL ARE PRETRAINED ON IMAGENET-21K AND FINE-TUNED ON IMAGENET-1K AT 224 × 224 RESOLUTION

Method	α	$\mathbb{E}_{\mathcal{O}}$	Acc. (%)
ViT-B/16[1]	1.0	38,416	81.20
AWD-ViT-B/16	0.50	38,334	81.62
	0.20	10,283	81.25
	0.10	8,866	81.90
	0.05	6,801	80.60
	0.025	4,971	78.28

attention window design can reduce the redundant attention operation and decrease the complexity of ViT model.

1) Justification: Naturally, the quality of adaptive attention window design can directly reflect the rationality of our proposed uncertainty-aware quantification of patch-wise interactions. Thus, we trained ViT models under various settings on ImageNet-1K dataset [46] for justifying effectiveness of U. First, we computed the patch interaction score map U for each image based on the well-trained ViT model, and then get the attention window based on B_p for each patch. After that, we incorporate this priori attention window range into all self-attention operations for re-training (not finetuning) ViT model. We selected $T = \alpha N^2$ with $\alpha = \{2.5\%, 5\%, 10\%, 20\%, 50\%\}$ to generate window at different scales. The experimental results are shown in Table I. It can be found that best Top-1 accuracy for adaptive attention window design is achieved when α is set to 10%. In this case, we only use nearly 23% of patch connections for global self-attention, with improvement of 0.7% from the baseline (81.20%).

The experimental results demonstrate that the responsive field for each patch is unique and data-dependent. Without the full-patch global attention, there is no decrease in the performance of the model. Based on the observation and analysis, we further prove that these approaches [22], [24], [25] of region/window/local are designed in a reasonable way. Meanwhile, it also validated the rationality of our approach for patch-wise interaction quantification.

B. Indiscriminative Patch

Noting that U is not a symmetrical matrix. The raw values $\mathbf{u}_p = \{u_{p,1}, \ldots, u_{p,N}\}$ in U measure the relevance of all

[1]The reported result in the official ViT implementation: https://github.com/google-research/vision_transformer

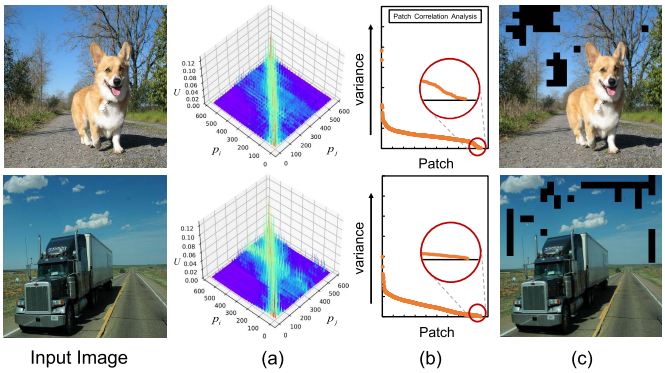

Fig. 5. Illustration about analysis for the existence of potentially indiscriminative patches (black) for ViT model. (a) Visualization of score map. (b) Calculate and sort the variance. (c) Visualization of indiscriminative patches.

TABLE II

TOP-1 ACCURACY OF VIT ON IMAGENET VALIDATION SET BY ADAPTING OUR PROPOSED AWD AND DROP INDISCRIMINATIVE PATCHES (DIP) ON VARIOUS SETTINGS OF α AND β

Method	α	β	$\mathbb{E}_{\mathcal{O}}$	Acc. (%)
ViT-B/16	1.0	0	38,416	81.20
AWD-ViT-B/16	0.10	0	8,866	81.90
	0.05	0	6,801	80.60
AWD-ViT-B/16 ***w/ DIP***	0.10	0.1	7,982	82.09
	0.10	0.2	7,080	82.40
	0.10	0.5	4,420	80.62

N patch tokens to the pth patch token, while the column values $\mathbf{u}'_p = \{u_{1,p}, \ldots, u_{N,p}\}$ in U reflects how informative x_p is. Here, we visualize the $N \times N$ score map of U for a given image in Fig. 5(a), and observe some anomalous patches with constant high column values of U (in the red box). It means that these patches indiscriminately provide their information to all other patches from the background to the main object in the image. For a more intuitive explanation, we compute the variance of column values $\mathbf{u}'$ for each patch and sort the results in Fig. 5(b). We define the patch with low variance in $\mathbf{u}'$ as indiscriminative patches and visualize the indiscriminative patches for three different images [Fig. 5(c)]. Obviously, these indiscriminative patches mainly located at empty information area in background of the key objects in image. More visualization results can be found in the Appendix.

Since the indiscriminative patches are data-dependent, but provide their contextual information to all other patch tokens consistently, they can be also regarded as the data-dependent bias for ViT model training.

1) Justification: To understand how the indiscriminative patches impact the ViT model training, we retrained the adaptive attention window designed ViT models of $\alpha = 0.10$ by erasing the indiscriminative patches. In special, we ranked the variance of $\mathbf{u}'$ for each patch, and generate a mask matrix M, where the βN patches which has the lowest variance values are masked as 0, while other patches are masked as 1. We multiply M for all patch tokens across all layers in ViT model during training and inference. The experimental results of ViT on ImageNet can be found in Table II. Moreover,

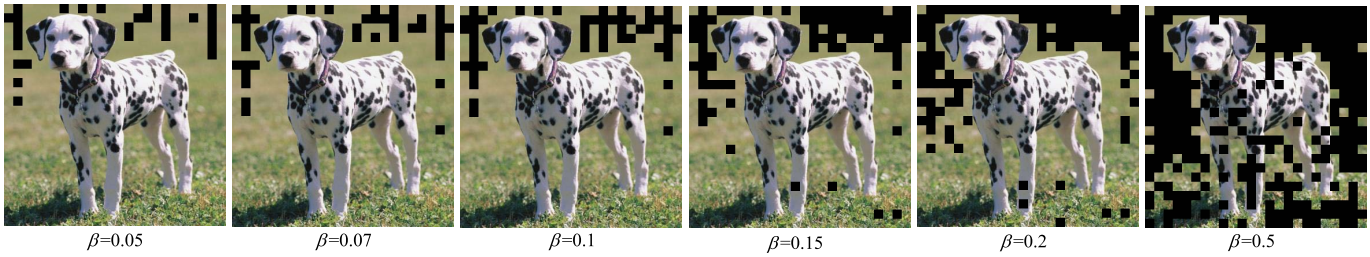

Fig. 6. Illustration of dropping indiscriminative patches (black) on various β rates for ViT model.

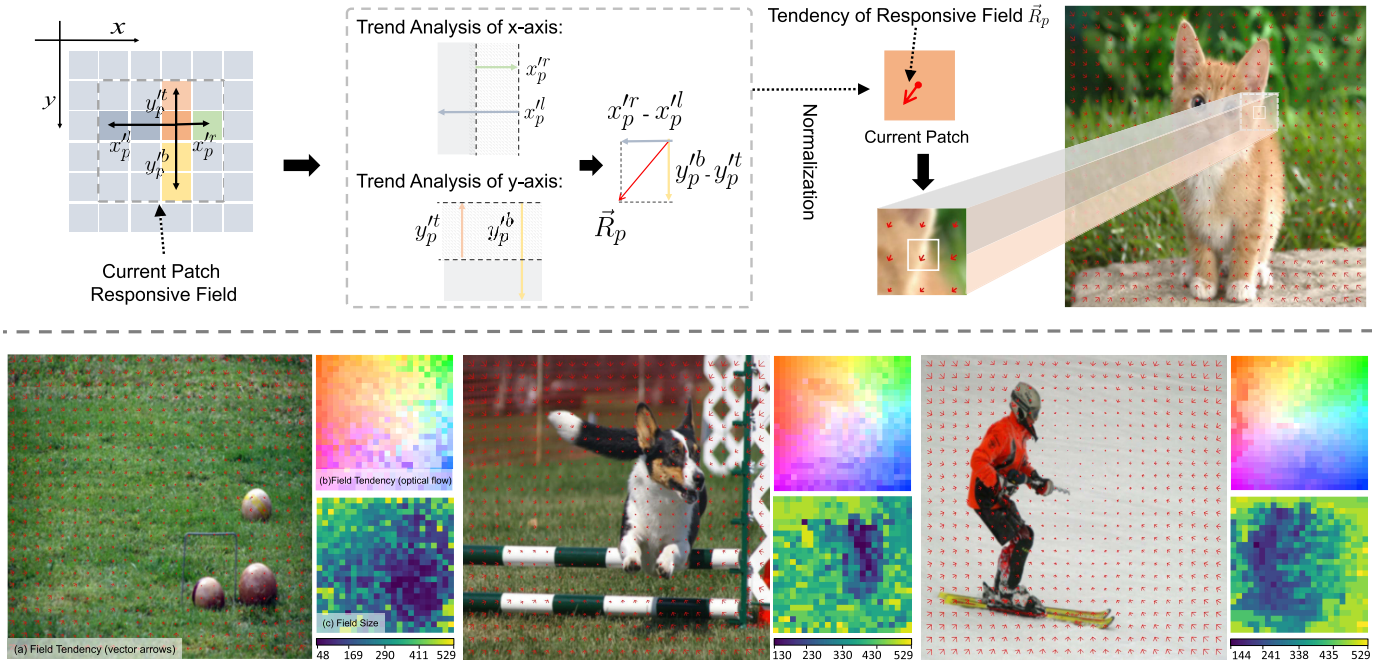

Fig. 7. Up row: Illustration of the tendency and size analysis of responsive field for given patch. Bottom row: Examples of responsive field analysis from ViT model. We illustrate the responsive field tendency using optical flow tool [47] and vector arrows. Meanwhile, we visualize the size of the responsive field for each patch using a heatmap.

Fig. 6 provides visualization about the influence of β rates for dropping indiscriminative patches.

It shows that dropping indiscriminative patches during ViT model training and inference results in better final performance improvement (82.40 versus 81.90). Even after masking out half of the patches ($\beta = 0.5$), the results of AWD-ViT-B/16 *w/ DIP* still have comparable performance with the original ViT. Considering patches with low variance in $\mathbf{u}'$ fairly provide information to all patches during global attention operation, and they are also data-dependent (different image results in different indiscriminative patches distributions), we can treat these indiscriminative patches as the image-specific bias during model training. Such bias would mislead the ViT model to learn image identification rather than the general discriminative patterns. In general, for such indiscirminative patches in ViT, less is more.

C. Responsive Field Analysis

We define the adaptive attention window updated by dropping indiscriminative patches as the responsive field for each patch. Here we make statistic analysis on responsive field in the following two aspects.

1) Field Size: Following the adaptive window design, given the pth patch coordinate $\langle x_p, y_p \rangle$, its attention window offsets $\{x_p^l, y_p^t, x_p^r, y_p^b\}$, indiscriminative patch set D of current image, the responsive field of $\mathbf{x}_p$ can be expressed as follows:

$$S_p = \{\mathbf{x}_p\} \cup \{\mathbf{x}_i : \langle x_i, y_i \rangle \in B_p, \mathbf{x}_i \notin D\}. \tag{6}$$

Here we visualize the size of responsive field $|S_p|$ for each patch in Fig. 7 [bottom row(c)]. We observe that the distribution of patch's responsive field size is relevant to the semantic information of each patch, i.e., responsive field for a patch of the main object usually tends to be smaller that the patch of background. A consequence is that smaller responsive fields are more focused on the local texture or structure learning, while big responsive field aims to learn the correlation between the object and the background.

a) Field Tendency: Meanwhile, the responsive field of each patch is constrained with four directions offsets in x- and y-axis. The patch-wise interactions can calculate the current patch tendency of a responsive field as shown in Fig. 7 (up row). Thus, we compute the pth patch responsive field $\vec{R}$

$$\vec{R}_p = \left(x_p'^r - x_p'^l, y_p'^b - y_p'^t\right) \tag{7}$$

where $x_p'^r, x_p'^l, y_p'^b, y_p'^t$ are the maximum and minimum offset in the x- and y-axis of responsive field S_p, respectively. After that, we normalize the $\vec{R}_p$ to represent the tendency of pth patch responsive field, and then visualize the tendency of each

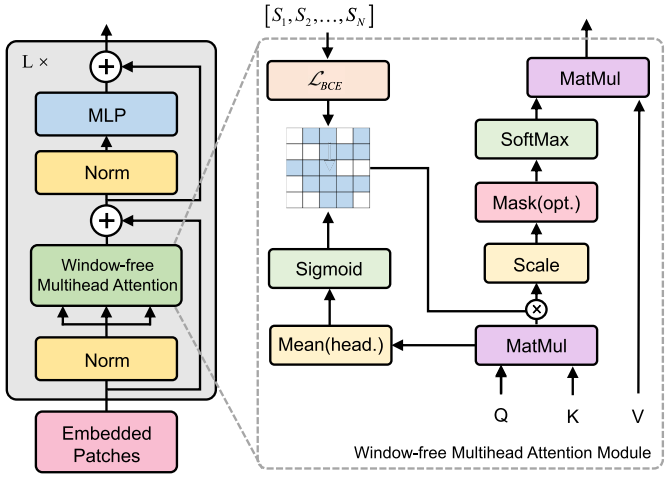

Fig. 8. Illustration of the window-free multihead attention mechanism. We propose a window-free modified attention mechanism to present an adaptive window for patch-wise interaction. This attention mechanism is trained with external supervisions of the responsive field. No external knowledge need during inference.

patch responsive field. As shown in Fig. 7, similar to the visualization of field size, the tendency of responsive field is also semantically relevant. Overall, the directional field of S_p is object-centric.

V. Window-Free Transformer

The patch-wise interactions analysis provides a novel complementary view to understanding the ViT model. Based on our observation and analysis, we propose a simple yet transformer architecture by incorporating the supervision of responsive fields during training. Meanwhile, this architecture can further validate the effectiveness of our observations.

A. Window-Free Multihead Attention

Multihead attention mechanism is the core of transformer architecture. Although much work has focused on the attention scores and layers to explore the representational ability in transformer, understanding which interactions are most effective that may influence effectiveness is critical to achieve improvement. Therefore, following our understanding and visual analysis, we design a data-driven multihead attention mechanism by incorporating the supervision of responsive fields during training.

As shown in Fig. 8, the window-free module first linearly projects the patch tokens to Q, K, and V as inputs. we compute the dot products of the Q with all K. After that, we apply an average computation and a sigmoid function to obtain the weights w' for window design on the values QK^T. Then, we can generate a binary mask W through the weights w' for restricting the patch-wise interactions.

Specifically, we formulate this process as follows:

$$w' = \text{Sigmoid}\left(\frac{1}{k}\sum^{k}(QK^T)\right) \tag{8}$$

$$W_{i,j} = \begin{cases} 1, & w'_{i,j} > \overline{w'_{i,:}} \\ 0, & w'_{i,j} \leqslant \overline{w'_{i,:}} \end{cases} \tag{9}$$

where the index k represents the number of heads in MSA. i and j are horizontal and vertical, respectively. $(1 \leqslant i, j \leqslant N)$. $\overline{w'_{i,:}}$ represents a dynamic threshold to get a binary mask, which is the average of all values $w'_{i,1:N}$ from the ith patch in the weights w'.

Therefore, the window-free multihead attention module computes the element-wise products between the QK^T and W, scales to stabilize training, and then softmax normalizes the result. The final attention results are obtained by computing dot production of value matrix V with masked attention score matrix

$$\mathcal{A}' = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}} * W\right) \tag{10}$$

where d is embedding dimension of Q and K.

B. Window Ground-Truth and Loss

1) Window Ground-Truth: We use patch-wise interaction analysis tools to evaluate all training dataset images. The outputs of image I patch-wise interactions window groundtruth w_{gt} can be written as follows:

$$w_{\text{gt}}(I) = \Theta_{(0,1)}([S_1, S_2, \ldots, S_N]) \tag{11}$$

where $\Theta_{(0,1)}$ represents the conversion of each patch's responsive field S_p into a binary mask. Note that we convert all window offsets to binary masks for more efficient attention computation.

2) Loss Function: To give a clear hint, we introduce the patch-wise interactions mask to guide the adaptive window design via adding a binary cross-entropy (BCE) loss between the window-free module output w' and corresponding interactions window groundtruth binary mask w_{gt}

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N^2}\sum(w_{\text{gt}}\log(w') + (1 - w_{\text{gt}})\log(1 - w')). \tag{12}$$

$\mathcal{L}_{\text{BCE}}$ provides a learnable adaptive window design representation of patch-wise interactions. By doing so, our WinfT architecture adaptively captures the responsive field.

We adopt cross-entropy loss $\mathcal{L}_{\text{CE}}$ as classification loss. Therefore, our model is trained with the sum of $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{BCE}}$ together which can be formulated as follows:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{CE}} + \lambda_2\mathcal{L}_{\text{BCE}} \tag{13}$$

where the λ_1 and λ_2 are hyper-parameters, which are respectively set to 1 and 1 in our experiments.

C. Implementation Details

1) Implementation on ImageNet: We use the intermediate weights from official ViT-B/16 and ViT-B/32 models pretrained on ImageNet-21K. The training is conducted with 4 GPUs (gradient accumulation is applied owing to the limited GPU memory) with mini-batch size of 512 and an initial learning rate of 0.01. We use SGD as the optimizer, the momentum of SGD is set as 0.9. The model is trained for 15 epochs. We use 500 warmup steps and adopt cosine annealing as the scheduler of the optimizer. Random cropping is employed as data augmentation during training. The backbone is designed to have 12 transformer layers with 12 attention heads. All these settings stay the same for our re-implemented ViT and our WinfT.

TABLE III

TOP-1 ACCURACY COMPARISON WITH VIT METHODS ON IMAGENET. THE $\mathbb{E}_{\mathcal{O}}$ OF WINFT MEASURES THE SUM OF PREDICTED BINARY PATCH-WISE ATTENTION MASK W AVERAGED OVER ALL IMAGES

Method	image size	Acc. (%)	$\mathbb{E}_{\mathcal{O}}$
ViT-B/32	224^2	74.46	2,401
ViT-B/32	384^2	81.28	20,736
ViT-B/16	224^2	81.20	38,416
ViT-B/16	384^2	83.61	331,776
WinfT-B/32	224^2	78.74	906
WinfT-B/32	384^2	84.33	8,638
WinfT-B/16	224^2	83.11	16,938
WinfT-B/16	384^2	84.62	136,327

a) Implementation on CUB-200-2011: We finetuned the ViT and WinfT model on CUB dataset from the official ViT-B/16 model pretrained on ImageNet-21K. These models are trained for 20 000 steps with a batch size of 16 and an initial learning rate of 0.03. Cosine annealing is adapted as the learning rate scheduler of an optimizer.

D. Experimental Results

1) Results on ImageNet: Following the settings in ViT [19], we adopt ViT-B as our backbone, which contains 12 transformer layers in total and pretrained on ImageNet-21K with 16^2 or 32^2 patch size. The batch size is set to 512. And we train all models using a mini-batch Stochastic Gradient Descent optimizer with a momentum of 0.9. The learning rate is initialized as 0.01 for ImageNet-1K. We then apply cosine annealing as the scheduler for the optimizer.

We notice in Table III that adaptive window learning to restrict the patch-wise interactions consistently leads to better performance. Specifically, compared with different input image resolutions and patch sizes, the WinfT architecture can achieve better performance with strong correlation patch connections for interactions. An interesting point is that after fitting a suitable attention window, the ViT model with 32×32 patch size can achieve similar performance with the settings of 16×16 patch size (84.33% versus 84.62%). Bigger patch size with less patch amount results in substantially reducing computation complexity for self-attention operation. It benefits the practical application of ViT models. In general, these experimental results validate the effectiveness of the supervision of responsive fields and our proposed quantification method for the impacts of patch-wise interactions.

Meanwhile, we visualize more samples in the ImageNet dataset to verify the efficacy of our proposed explainable visualization schema in Fig. 9.

2) Results of Transfer Learning: To further verify the effectiveness of our proposed visualization analysis and window design method, we conduct a comprehensive study of fine-grained classification. Note that fine-grained classification aims at classifying the sub-classes to find subtle differences in similar classes, and the model needs to focus more on discriminative feature learning.

We show the experimental results of transfer learning on fine-grained benchmarks in Table IV. Our WinfT achieves 0.48% improvement on Top-1 accuracy with reducing 66.27% patch-wise interactions compared with the results of the original ViT. It further demonstrated the generalization of our proposed patch interaction analysis method.

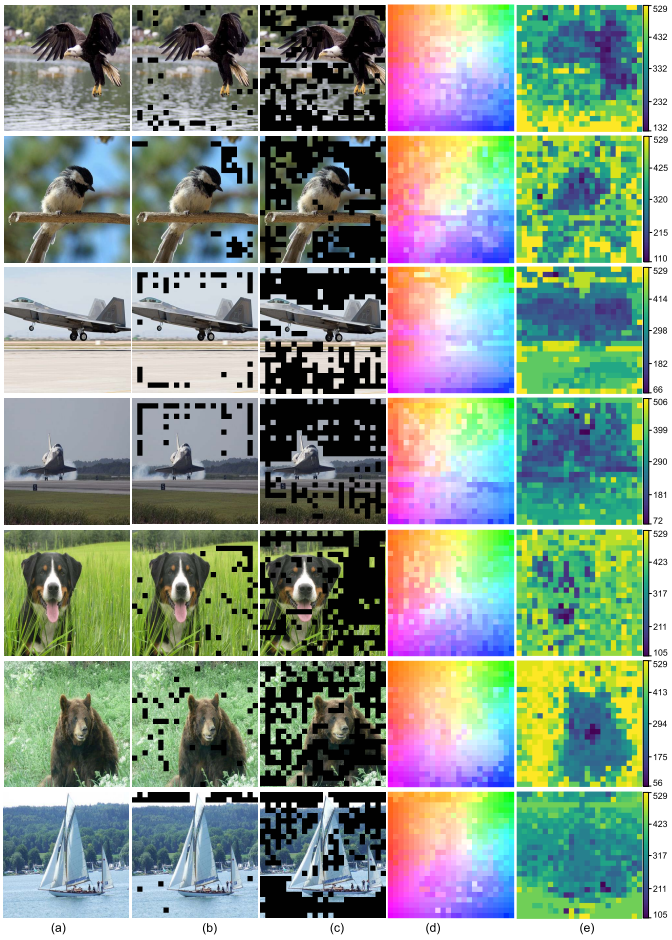

Fig. 9. Illustration of the extended visualization results. (a) Input image. (b) *DIP* ($\beta = 7\%$). (c) *DIP* ($\beta = 50\%$). (d) Field tendency (optical flow). (e) Field size.

TABLE IV

PERFORMANCE COMPARISON ON CUB DATASET. ALL MODELS ARE TRAINED AND EVALUATED AT 448×448 RESOLUTION. *DENOTES THE AVERAGE OF MULTIPLE RUNS RESULTS

Method	Acc. (%)	$\mathbb{E}_{\mathcal{O}}$
ViT-B/16	90.30	614,656
WinfT-B/16	90.78*	207,325*

3) Visualization of Predicted Attention Windows: We select two samples (large and small objects) to visualize the ground-truth attention windows (b) and the predicted attention windows of WinfT (c) in Fig. 10. Since the predicted attention windows are different among all layers, we visualize the sum of predicted attention masks matrix $\sum_l^{12} W_l$ of window-free module across all 12 layers in WinfT (lighter block means more layers voting mask value of 1), where W_l denotes the predicted binary mask in the lth layer of WinfT.

VI. LIMITATIONS

The visual analysis and understanding provide a more interpretive understanding of the patch-wise interactions and further guide the design of effective the transformer architectures, but there still exist some limitations. Inevitably, we need

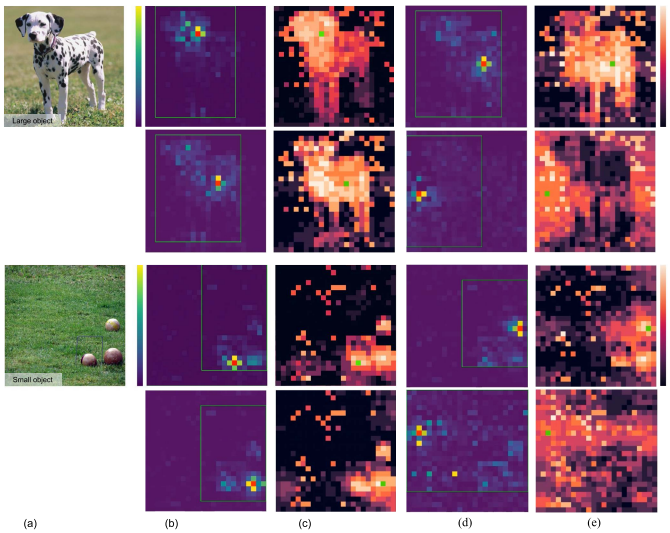

Fig. 10. Illustration of the responsive field from ViT (b) and (d) predicted attention mask in WinfT (c) and (e). (a) Input image. (b) AWD-ViT-B/16. (c) WinfT-B/16. (d) AWD-ViT-B/16. (e) WinfT-B/16.

to retrain the model through the adaptive window design to restrict the patch-wise interactions. Notably, we need to state that the proposed method is aimed at verifying the reasonableness of the analysis of the patch-wise interactions instead of proposing a state-of-the-art transformer architecture. The computation reduction in $\mathbb{E}_{\mathcal{O}}$ reflects the potentiality of higher efficient ViT models, but WinfT still needs to compute the attention mask. Our window-free transformer model can be treated as distilling a model using the less attention operations. WinfT is designed as a posteriori proof of the rationality of our proposed explainable visualization schema for ViT. More importantly, we hope that our analytical methodology can provide some new insights for future transformer-based model design.

VII. Conclusion

In this article, we proposed a novel explainable visualization schema to analyze and interpret the patch-wise interactions for ViT. Concretely, we first investigate the interaction between patches and then propose a quantification schema for measuring the impact of patch interactions. Based on the quantification results, we make a series of experimental verification and statistic analysis on the responsive field of patch. Motivated by our observations on responsive field, we propose a WinfT architecture by adaptively restricting the patch interactions. Experimental results demonstrated both the effectiveness and efficiency of our architecture.

References

[1] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, 2000, pp. 1137–1155.

[2] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-aware neural language models," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 2741–2749.

[3] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[4] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does BERT look at? An analysis of BERT's attention," 2019, *arXiv:1906.04341*.

[5] Y. Lin, Y. C. Tan, and R. Frank, "Open sesame: Getting inside BERT's linguistic knowledge," 2019, *arXiv:1906.01698*.

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 213–229.

[7] T. Ma et al., "Oriented object detection with transformer," 2021, *arXiv:2106.03146*.

[8] Z. Sun, S. Cao, Y. Yang, and K. Kitani, "Rethinking transformer-based set prediction for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3611–3620.

[9] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[10] M. Zhuge, D.-P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao, "Salient object detection via integrity learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3738–3752, Mar. 2023.

[11] R. Guo, D. Niu, L. Qu, and Z. Li, "SOTR: Segmenting objects with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7157–7166.

[12] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12077–12090.

[13] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.

[14] H. Du, J. Wang, M. Liu, Y. Wang, and E. Meijering, "SwinPA-Net: Swin transformer-based multiscale feature pyramid aggregation network for medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 19, 2022, doi: 10.1109/TNNLS.2022.3204090.

[15] J. Zhang, Z. Fang, H. Sun, and Z. Wang, "Adaptive semantic-enhanced transformer for image captioning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 29, 2022, doi: 10.1109/TNNLS.2022.3185320.

[16] Z. Shao, J. Han, D. Marnerides, and K. Debattista, "Region-object relation-aware dense captioning via transformer," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 11, 2022, doi: 10.1109/TNNLS.2022.3152990.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1106–1114.

[18] Y. LeCun et al., "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[19] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[20] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing properties of vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 23296–23308.

[21] J. He et al., "TransFG: A transformer architecture for fine-grained recognition," 2021, *arXiv:2103.07976*.

[22] C.-F. Chen, R. Panda, and Q. Fan, "RegionViT: Regional-to-local attention for vision transformers," 2021, *arXiv:2106.02689*.

[23] X. Dong et al., "CSWin transformer: A general vision transformer backbone with cross-shaped windows," 2021, *arXiv:2107.00652*.

[24] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[25] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh, "DynamicViT: Efficient vision transformers with dynamic token sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 13937–13949.

[26] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15979–15988.

[27] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: BERT pre-training of image transformers," in *Proc. 10th Int. Conf. Learn. Represent.*, 2022, pp. 1–18.

[28] X. Chu et al., "Twins: Revisiting spatial attention design in vision transformers," 2021, *arXiv:2104.13840*.

[29] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.

[30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[31] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," 2016, *arXiv:1605.01713*.

[32] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.
[33] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 4126–4135.
[34] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.
[35] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.
[36] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognit.*, vol. 65, pp. 211–222, May 2017.
[37] S. Bhojanapalli et al., "Leveraging redundancy in attention with reuse transformers," 2021, *arXiv:2110.06821*.
[38] Y. Qin, C. Zhang, T. Chen, B. Lakshminarayanan, A. Beutel, and X. Wang, "Understanding and improving robustness of vision transformers through patch-based negative augmentation," 2021, *arXiv:2110.07858*.
[39] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how BERT works," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 842–866, Dec. 2020.
[40] S. Serrano and N. A. Smith, "Is attention interpretable?" 2019, *arXiv:1906.03731*.
[41] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," 2019, *arXiv:1905.09418*.
[42] P. Michel, O. Levy, and G. Neubig, "Are sixteen heads really better than one?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 14014–14024.
[43] A. Raganato, Y. Scherrer, and J. Tiedemann, "Fixed encoder self-attention patterns in transformer-based machine translation," 2020, *arXiv:2002.10260*.
[44] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12116–12128.
[45] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. Van Gool, "LocalViT: Bringing locality to vision transformers," 2021, *arXiv:2104.05707*.
[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[47] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *Int. J. Comput. Vis.*, vol. 92, no. 1, pp. 1–31, Mar. 2011.

Jie Ma received the M.S. degree in computer science from Huaqiao University, Xiamen, China, in 2022.

He is currently a Visiting Scholar at Guangxi Normal University, Guilin, China. His current research interests include computer vision and machine learning.

Yalong Bai received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 2018, in Microsoft Research Asia Joint Ph.D. Education Program.

He is a Senior Researcher at JD.com., Beijing, China. He has authored or coauthored 25 academic papers, including the Computer Vision and Pattern Recognition Conference, the International Conference on Computer Vision, the European Conference on Computer Vision, the International Conference on Learning Representations, and the Association for the Advancement of Artificial Intelligence. His current research interests include representation learning, multimodal retrieval, visual question answering, and visual commonsense reasoning.

Dr. Bai has won the first prize in several international challenges such as FGVC at CVPR2019, MSR Image Recognition at ICME2016, and MSR-Bing Image Retrieval at MM2014.

Bineng Zhong received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2004, 2006, and 2010, respectively.

From 2007 to 2008, he was a Research Fellow with the Institute of Automation and Institute of Computing Technology, Chinese Academy of Science, Beijing, China. From September 2017 to September 2018, he was a Visiting Scholar with Northeastern University, Boston, MA, USA. From November 2010 to October 2020, he was a Professor with the School of Computer Science and Technology, Huaqiao University, Xiamen, China. Currently, he is a Professor with the School of Computer Science and Engineering, Guangxi Normal University, Guilin, China. His current research interests include pattern recognition, machine learning, and computer vision.

Wei Zhang (Member, IEEE) received the Ph.D. degree from the Department of Computer Science, City University of Hong Kong, Hong Kong, China, in 2015.

He is now a Senior Researcher at JD.com., Beijing, China. His current research interests include computer vision and multimedia, especially visual recognition and generation.

Dr. Zhang has won the Best Demo Awards in ACM MM 2021 and ACM-HK Openday 2013. He served as the Area Chair for ICME, ICASSP, VCIP, and Technical Program Chair for ACM MM Asia 2023. He also served as a Guest Editors for *ACM Transactions on Multimedia Computing, Communications, and Applications* and *Advances in Multimedia*.

Ting Yao (Senior Member, IEEE) was a Principal Researcher with JD AI Research, Beijing, China and a Researcher with Microsoft Research Asia, Beijing. Currently, he is the CTO of HiDream.ai, a high-tech startup company focusing on generative intelligence for creativity. He has coauthored or coauthored more than 80 peer-reviewed papers in top-notch conferences/journals, with more than 12 000 citations. He has developed one standard 3-D convolutional neural network, i.e., Pseudo-3-D Residual Net, for video understanding, and his video-to-text dataset of MSR-VTT has been used by more than 400 institutes worldwide.

Dr. Yao serves as an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA, *Pattern Recognition Letters*, and *Multimedia Systems*. His works have led to many awards, including the 2015 ACM-SIGMM Outstanding Ph.D. Thesis Award, the 2019 ACM-SIGMM Rising Star Award, the 2019 IEEE-TCMC Rising Star Award, the 2022 IEEE ICME Multimedia Star Innovator Award, and the winning of more than 10 championships in worldwide competitions.

Tao Mei (Fellow, IEEE) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively.

He was the Vice President of JD.com, Beijing, China, and a Senior Research Manager of Microsoft Research. He is the Founder and the CEO of HiDream.ai. He has authored or coauthored more than 200 publications (with 12 best paper awards) in journals and conferences, ten book chapters, and edited five books. He holds more than 25 U.S. and international patents.

Dr. Mei is a fellow of IAPR (2016), a Distinguished Scientist of ACM (2016), and a Distinguished Industry Speaker of IEEE Signal Processing Society (2017). He is the General Co-Chair of IEEE ICME 2019, the Program Co-Chair of ACM Multimedia 2018, IEEE ICME 2015, and IEEE MMSP 2015. He has been an Editorial Board Member of IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, *ACM Transactions on Multimedia Computing, Communications, and Applications*, and *Pattern Recognition*.