

Visualizing and Understanding Patch Interactions in Vision Transformer

Jie Ma^{1,2,3*} Yalong Bai^{3†} Bineng Zhong^{1§} Wei Zhang³ Ting Yao³ Tao Mei³

¹Guangxi Normal University ²Huaqiao University ³JD Explore Academy

majie@stu.hqu.edu.cn ylbai@outlook.com bnzhong@gxnu.edu.cn

{wzhang.cu, tingyao.ustc}@gmail.com tmei@jd.com

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 vision transformer, 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 vision transformer. 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 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. Moreover, the results on downstream fine-grained recognition tasks further validate the generalization of our proposal.

1. Introduction

Transformer architecture has led to the revolutionizing of Natural Language Processing (NLP) filed and inspires the emergence of transformer-type works on learning word [3] and character [14] level representations with self-attention mechanisms [35] for capturing dependency syntax [8] and grammatical [18] relationships. This has also motivated the recent works of vision transformers for vision tasks by using multi-head self-attention and multi-layer perceptrons, which are shown able to perform well on ImageNet clas-

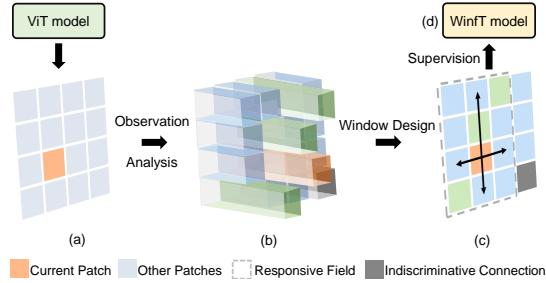


Figure 1. Given a patch (yellow) in (a), we first analyze and quantify the impact of interactions between the current patch and all other patches. According to our patch interaction quantitative results, we observe the distributions of relevant patches for current patch in (b), figure out the boundary of responsive field that containing patches with critical information to current patch in (c). Then, we apply the responsive field for guiding vision transformer model training. The 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.

sification and various downstream tasks, such as object detection [5, 33, 40], semantic segmentation [12, 38, 39], etc.

Different from convolutional neural networks [15, 16] which focus on local receptive fields, transformer-based architecture [11] utilizes patch-wise attention mechanism for full-patch information interactions with dynamic receptive fields [22]. As a general vision transformer backbone, ViT is capable of global feature extraction by dense information aggregation and interactions among full patch tokens. Several previous methods [13, 35] have interpreted how classification outputs are formed in vision transformer 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 [6, 10, 19, 26] of ViT with heuristic configurations demonstrated the restricted attention range/window/region for patches can reduce redundancy but without performance

* This work was done at JD Explore Academy.

† Equal contribution. § Corresponding author.

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 connections, and eventually guiding the visual attention model design.

To this end, we seek to obtain a better understanding of vision transformer models, especially the information interactions between patches. The problem of understanding vision transformer 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, the 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 token often gather information from its 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 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 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 architecture (WinfT) 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:

- We propose a novel explainable visualization schema to analyze and interpret the crucial attention interactions among patches for vision transformer. 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.
- Based on the quantification of the impact of patch interaction, we figure out the existence of indiscrimina-

tive patches in images for ViT model training. Further, dropping these patches also benefit the ViT model.

- We define responsive field for providing a crucial understanding of vision transformer. The statistic analysis shows that both the size and tendency of the informative attention window for each patch are semantic oriented.
- Inspired by above observations and analysis, we propose a novel window-free transformer 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.

2. Related Work

Transformer-based vision models Transformer-based vision models aim to utilize the attention mechanism to learn global dependencies representations. ViT [11] 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. Various transformer-based vision models [6, 7, 10, 19, 37] present the attention mechanism through heuristic patch-wise interactions, in particular, focus on different structures to enhance the patch information interactions for effective attention computation. Beyond that, DynamicViT [26] propose 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.

Explainability for Vision Transformer Given the key role of explainability in deep learning, several works have analyzed the gradients [28, 30–32, 34] and attribute propagation [1, 21] in Convolutional Neural Network to generate an understanding of representations to explain specific assumptions. Recent transformer works mainly [4, 23, 27, 29] focus on analyzing and interpreting attention scores to understand why the model performs so well. Reuse Transformer [4] highlights the relevant relationship between the different layers and captures the similarity to reuse the attention score. Elena Voita et al. [36] 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 to understanding attention mechanisms.

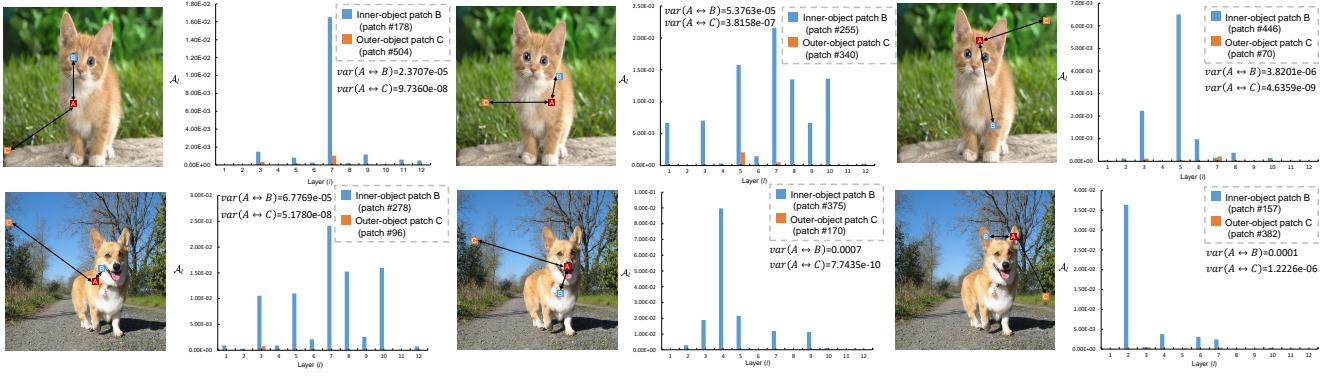


Figure 2. Illustration about quantitative analysis between the inner-object patch and outer-object patch. Zoom in for better visualization.

However, there are two limitations to understand vision transformer models: i) they do not highlight the relevance of patch-to-patch connection. ii) 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.

3. Patch Interactions Visualization

We study the patch interactions on global multihead self-attention of the original visual transformer [11], 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 vision transformers.

3.1. Preliminaries

Given an input image of $H \times W$ resolution, Vision Transformer model (ViT) first splits it into a sequence of non-overlapping patches $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ 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 a multi-head self-attention (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:

$$\mathcal{A} = \text{SoftMax}(QK^T / \sqrt{d}), \quad (1)$$

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 p -th patch token. Multihead self-attention (MSA) is an extension of SA in 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.

3.2. 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 [10, 19, 26, 37] have demonstrated the existence of redundant computations in the patch attention mechanism through fixed-scale window design, pruning, etc. 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 [20, 24, 25, 36]. We believe that 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 from 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 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 mentioned in previous work [22], we can find that, although the irrelevance patch pair (inner-outer object patch pair) consis-

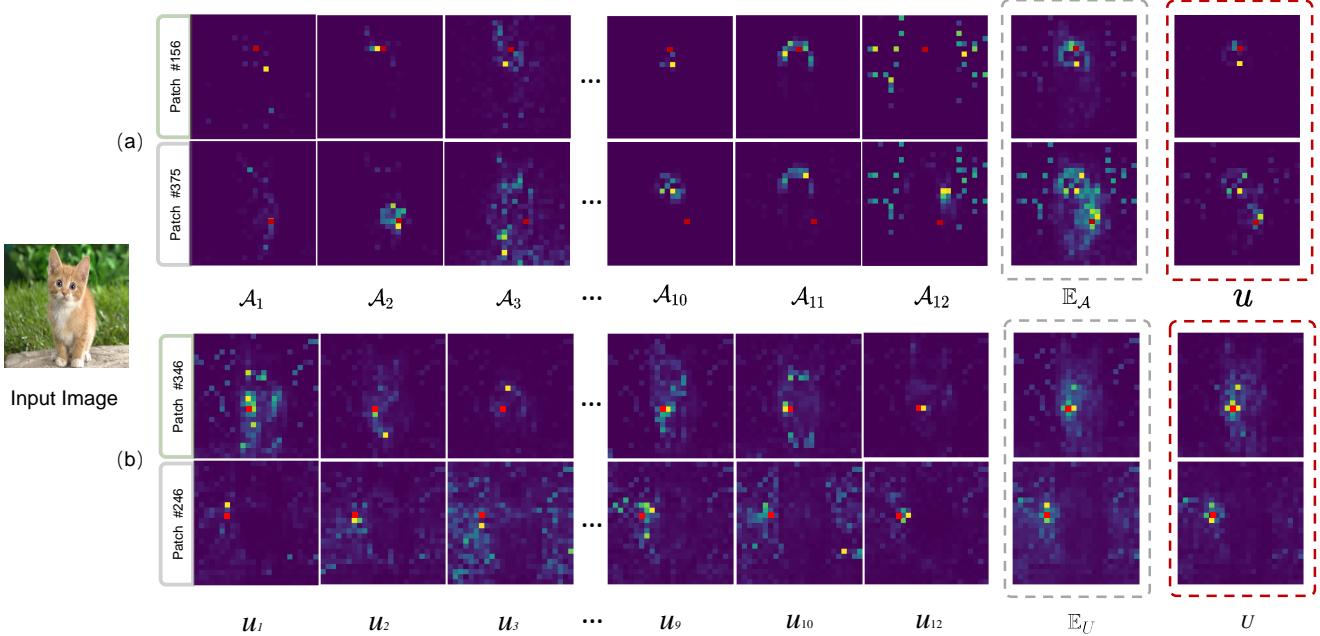


Figure 3. Visualization of attention score maps \mathcal{A} and its uncertainty across different layers \mathcal{U} . One 384×384 resolution images are fed into the well-trained ViT-B/16 [11] with 12 heads, 12 layers and 16×16 patch size. We randomly sampled four patches (in red color) of the input image for visualization.

tently has 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 i -th Transformer block) present an obvious uncertain of “tight” or “loose”. 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 multihead self-attention layers, a patch token can intensely gather representations from other patch tokens of high relevance with it. This results in unstable attention score and high uncertainty among relevance patches across layers.

Thus, for a standard ViT model contains of l multihead self-attention layer, all 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:

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

Thanks to the uniform head size setting of 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 high uncertainty region are gathered

around the corresponding patch, different channel of score map results in various range 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 of uncertainty cross k channel, and highlight the boundary of interactive regions for each patch as

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

where \mathcal{U}_i is the $N \times N$ score map extracted from the i -th 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 most interactive regions of the current patch is in the surrounding area. There are also some connections of distant patches, which are usually relevant to background around the main object in image. These visualizations are also conform to the general knowledge that locality information play as a critical role for object recognition [7, 17, 19]. Inspired by these related works, we also proposed two U guided attention window design methods for ViT in Sec. 4.1 and Sec. 5 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

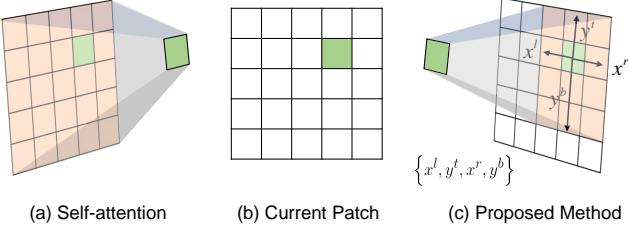


Figure 4. Illustration about global self-attention mechanism (a) and our proposed adaptive window design (c). Green patch represents the current patch for attention operation. Yellow region indicates the range of patch interactions for current patch.

details can be found in the sections below.

4. Patch Interactions Analysis

Based on our quantification of patch-wise interactions, we figure out the existence of potentially indiscriminative patches for ViT model (Sec. 4.2), and proposed an adaptive attention window design method (Sec. 4.1) for further analyzing the redundancy of global attention mechanism and the responsive field of each patch in ViT model (Sec. 4.3).

4.1. 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 [6, 10, 19] and dynamic pruning structures [26] 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 p -th patch whose coordinate is $\langle x_p, y_p \rangle$ ($x_p = p\% \sqrt{N}$, $y_p = p/\sqrt{N}$), we can construct a window boundary candidate set B_p for it:

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

Subsequently, we select the maximum and minimum offset in the x-axis 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

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

Method	α	$\mathbb{E}_{\mathcal{O}}$	Acc. (%)
ViT-B/16 ¹	1.0	38,416	81.20
	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

Table 1. 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.

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 build long-range interaction between all patches. Our proposed method provides adaptive attention window in terms of the effective interactions for each patch. Moreover, adaptive attention window design can reduce the redundant attention operation and decrease the complexity of ViT model.

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 [9] 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 Tab. 1. 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 the nearly 23% of patch connections for global self-attention, with the improvement of 0.7% from the baseline (81.20%).

The experimental results demonstrate that the responsive filed 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 [6, 19, 26] of region/window/local are designed in a reasonable way. Meanwhile, it also validated the rationality of our approach for patch-wise interaction quantification.

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

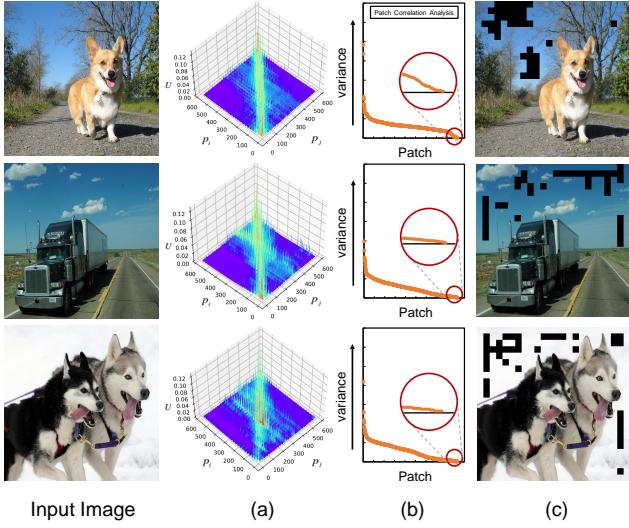


Figure 5. Illustration about analysis for the existence of potentially indiscriminative patches (Black) for ViT model.

4.2. Indiscriminative patch

Noting that U is not a symmetrical matrix. The raw values $\mathbf{u}_p = \{u_{p,1}, \dots, u_{p,N}\}$ in U measure the relevance of all N patch tokens to the p -th patch token, while the column values $\mathbf{u}'_p = \{u_{1,p}, \dots, u_{N,p}\}$ 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 background to the main object in 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.

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

Method	α	β	E_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 <i>w/ DIP</i>	0.10	0.1	7,982	82.09
	0.10	0.2	7,080	82.40
	0.10	0.5	4,420	80.62

Table 2. Top-1 accuracy of Vision Transformer on ImageNet validation set by adapting our proposed AWD and drop indiscriminative patches (DIP) on various settings of α and β .

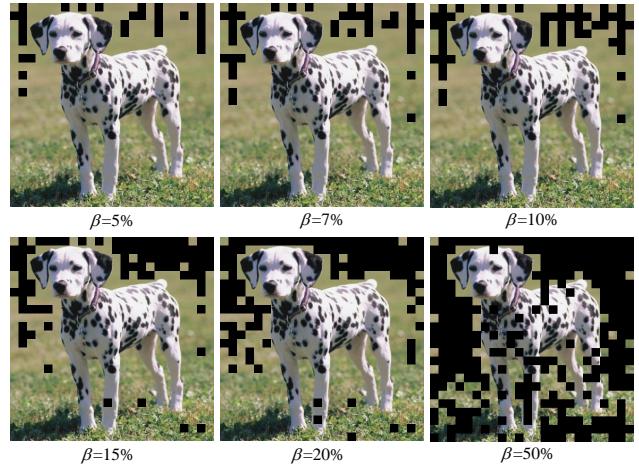


Figure 6. Illustration of dropping indiscriminative patches (Black) on various β rates for ViT model.

experimental results of ViT on ImageNet can be found in Tab. 2. Moreover, 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 the better final performance improvement (82.40 vs. 81.90). Even after masking out half of 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 indiscriminative patches in ViT, less is more.

4.3. Responsive field analysis

We define the adaptive attention window updated by dropping indiscriminative patches as the responsive field for

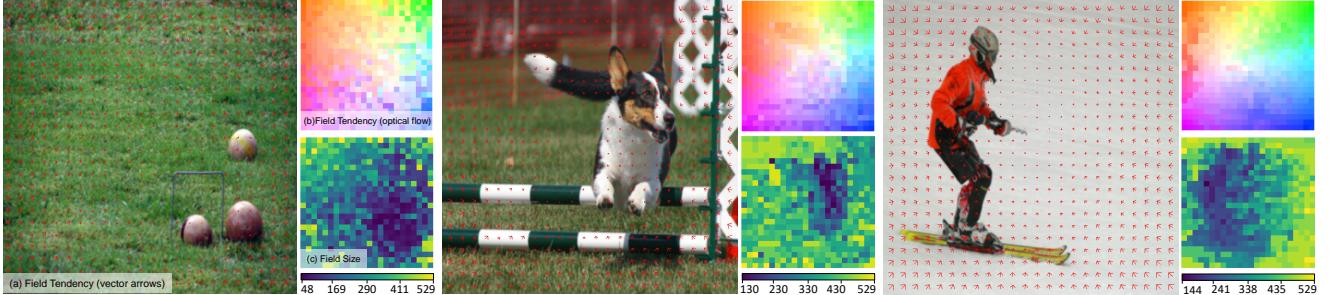


Figure 7. Examples of responsive field analysis from ViT model. We illustrate the responsive field tendency using optical flow tool [2] and vector arrows. Meanwhile, we visualize the size of the responsive field for each patch using a heatmap.

each patch. Here we make statistic analysis on responsive field in the following two aspects:

Field size Following the adaptive window design, given the p -th 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 x_p can be expressed as:

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

Here we visualize the size of responsive field $|S_p|$ for each patch in Fig. 7 (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 patch of the main object usually tend to be smaller than 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 object and the background.

Field tendency Meanwhile, the responsive field of each patch is constrained with four directions offsets in x-axis and y-axis. The patch-wise interactions can calculate the current patch tendency of responsive field as shown in Fig. 8. Thus, we compute the p -th patch responsive field \vec{R} :

$$\vec{R}_p = (x_p^{tr} - x_p^{tl}, y_p^{tb} - y_p^{tt}), \quad (7)$$

where $x_p^{tr}, x_p^{tl}, y_p^{tb}, y_p^{tt}$ are the maximum and minimum offset in the x-axis and y-axis of responsive field S_p respectively. After that, we normalize the \vec{R}_p to represent the tendency of p -th patch responsive field, and then visualize the tendency of each patch responsive field. As shown in Fig. 7, similar to the visualization of field size, the tendency of responsive field are also semantically relevant. Overall, the directional field of S_p is object-centric.

5. Window-free Transformer

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

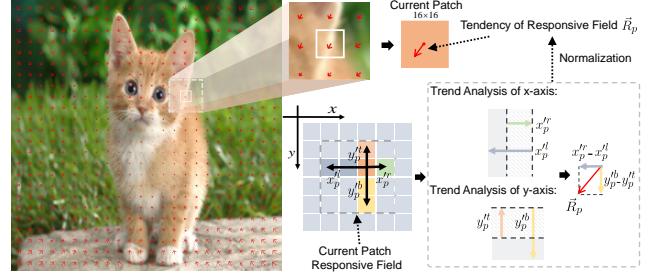


Figure 8. Illustration of the tendency and size analysis of responsive field for given patch.

Meanwhile, this architecture can further validate the effectiveness of our observations.

5.1. 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. 9, 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:

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

$$W_{i,j} = \begin{cases} 1, & w'_{i,j} > \overline{w'_{i,:}} \\ 0, & w'_{i,j} \leq \overline{w'_{i,:}} \end{cases} \quad (8)$$

where i and j are the horizontal and vertical, respectively.

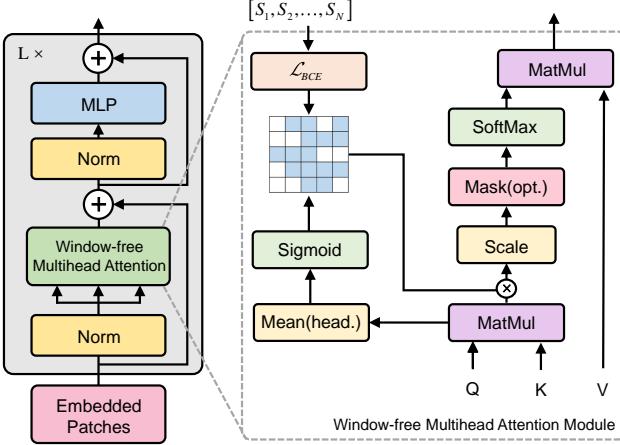


Figure 9. Illustration of a Transformer block with our Window-free Multihead Attention module. This attention mechanism is trained with external supervisions of the responsive field for presenting an adaptive patch-wise interaction window. No external knowledge need during inference.

$(1 \leq i, j \leq N)$. $\overline{w'_{i,:}}$ represents as a dynamic threshold to get binary mask, which is the average of all values $w'_{i,1:N}$ from the i -th 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:

$$\mathcal{A}' = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}} * W\right), \quad (9)$$

where d is embedding dimension of Q and K .

5.2. Window ground-truth and loss

Window ground-truth We use the patch-wise interactions analysis tools to evaluate all training dataset images. The outputs of image I patch-wise interactions window groundtruth w_{gt} can be written as:

$$w_{gt}(I) = \Theta_{(0,1)}([S_1, S_2, \dots, S_N]), \quad (10)$$

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.

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} :

$$\mathcal{L}_{BCE} = -\frac{1}{N^2} \sum (w_{gt} \log(w') + (1 - w_{gt}) \log(1 - w')). \quad (11)$$

Method	image size	Acc. (%)	\mathbb{E}_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

Table 3. Top-1 accuracy comparison with ViT methods on ImageNet. The \mathbb{E}_O of WinfT measures the sum of predicted binary patch-wise attention mask W averaged over all images.

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

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

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{CE} + \lambda_2 \mathcal{L}_{BCE}, \quad (12)$$

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

5.3. Experimental Results

Results on ImageNet Following the settings in ViT [11], 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 the 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 Tab. 3 that adaptive window learning to restrict the patch-wise interactions consistently lead to better performance. Specifically, compared with different input image resolutions and patch sizes, the window-free transformer architecture can achieve better performance with strong correlation patch connections for interactions. An interesting point is that after fitting 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% vs. 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 vision transformer models. In general, these experimental results validate the effectiveness of the supervision of responsive field and our proposed quantification method for the impacts of patch-wise interactions.

Results of transfer learning To further verify the ef-

Method	Acc. (%)	$\mathbb{E}_{\mathcal{O}}$
ViT-B/16	90.30	614,656
WinfT-B/16	90.58	207,532

Table 4. Performance comparison on CUB dataset. All models are trained and evaluated at 448×448 resolution.

fectiveness 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 Tab. 4. Our WinfT achieves 0.28% improvement on Top-1 accuracy with reducing 66.24% patch-wise interactions compared with the results of the original ViT. It further demonstrated the generalization of our proposed patch interaction analysis method.

6. 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 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. More importantly, we hope that our analytical methodology can provide some new insights for future transformer-based model design.

7. Conclusions

In this paper, we proposed a novel explainable visualization schema to analyze and interpret the patch-wise interactions for vision transformer. Concretely, we first investigate the interaction between patches and then propose the uncertainty-aware 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 window-free transformer architecture by adaptively restricting the patch interactions. Experimental results demonstrated both effectiveness and efficient of our architecture.

References

- [1] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. [2](#)
- [2] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International journal of computer vision*, 92(1):1–31, 2011. [7](#)
- [3] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in Neural Information Processing Systems*, 13, 2000. [1](#)
- [4] Srinadh Bhojanapalli, Ayan Chakrabarti, Andreas Veit, Michal Lukasik, Himanshu Jain, Frederick Liu, Yin-Wen Chang, and Sanjiv Kumar. Leveraging redundancy in attention with reuse transformers. *arXiv preprint arXiv:2110.06821*, 2021. [2](#)
- [5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. [1](#)
- [6] Chun-Fu Chen, Rameswar Panda, and Quanfu Fan. Regionvit: Regional-to-local attention for vision transformers. *arXiv preprint arXiv:2106.02689*, 2021. [1, 2, 5](#)
- [7] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting spatial attention design in vision transformers. *arXiv e-prints*, pages arXiv–2104, 2021. [2, 4](#)
- [8] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert’s attention. *arXiv preprint arXiv:1906.04341*, 2019. [1](#)
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009. [5](#)
- [10] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baineng Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. *arXiv preprint arXiv:2107.00652*, 2021. [1, 2, 3, 5](#)
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. [1, 2, 3, 4, 8](#)
- [12] Ruohao Guo, Dantong Niu, Liao Qu, and Zhenbo Li. Sotr: Segmenting objects with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7157–7166, 2021. [1](#)
- [13] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, Changhu Wang, and Alan Yuille. Transfg: A transformer architecture for fine-grained recognition. *arXiv preprint arXiv:2103.07976*, 2021. [1](#)

- [14] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. Character-aware neural language models. In *Thirtieth AAAI conference on artificial intelligence*, 2016. 1
- [15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1
- [16] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 1
- [17] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*, 2021. 4
- [18] Yongjie Lin, Yi Chern Tan, and Robert Frank. Open sesame: getting inside bert’s linguistic knowledge. *arXiv preprint arXiv:1906.01698*, 2019. 1
- [19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1, 2, 3, 4, 5
- [20] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019. 3
- [21] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65:211–222, 2017. 2
- [22] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 3
- [23] Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, and Xuezhi Wang. Understanding and improving robustness of vision transformers through patch-based negative augmentation. *arXiv preprint arXiv:2110.07858*, 2021. 2
- [24] Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. Fixed encoder self-attention patterns in transformer-based machine translation. *arXiv preprint arXiv:2002.10260*, 2020. 3
- [25] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34, 2021. 3
- [26] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamiccvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34, 2021. 1, 2, 3, 5
- [27] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. 2
- [28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2
- [29] Sofia Serrano and Noah A Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019. 2
- [30] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016. 2
- [31] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 2
- [32] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. *Advances in neural information processing systems*, 32, 2019. 2
- [33] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris M Kitani. Rethinking transformer-based set prediction for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3611–3620, 2021. 1
- [34] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 2
- [35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1
- [36] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019. 2, 3
- [37] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 2, 3
- [38] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 1
- [39] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 1
- [40] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1

A. Supplementary Material

This supplementary material provides more visualization analysis samples to verify our visualization and understanding of patch interactions in vision transformer. First, we present details of our experimental setup. Then, the responsive field's tendency and size distributions are also visualized. Together, we provide more extended visualization results by analyzing the patch-wise interactions to prove the efficacy of our proposed methods further. After that, we compared the supervision of the responsive field from ViT and the predicted attention mask in WinfT.

1. Experimental Details

As shown in Fig. 10, our explainable visualization schema aims to analyze and interpret the patch-wise interactions, and then extract valuable patch interaction supervisions for guiding WinfT model training. It is worth noting that the WinfT model does not require any supervision information during the inference phase.

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 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. All these settings stay the same for our re-implemented ViT and our WinfT.

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 initial learning rate of 0.03. Cosine annealing is adapted as the learning rate scheduler of optimizer.

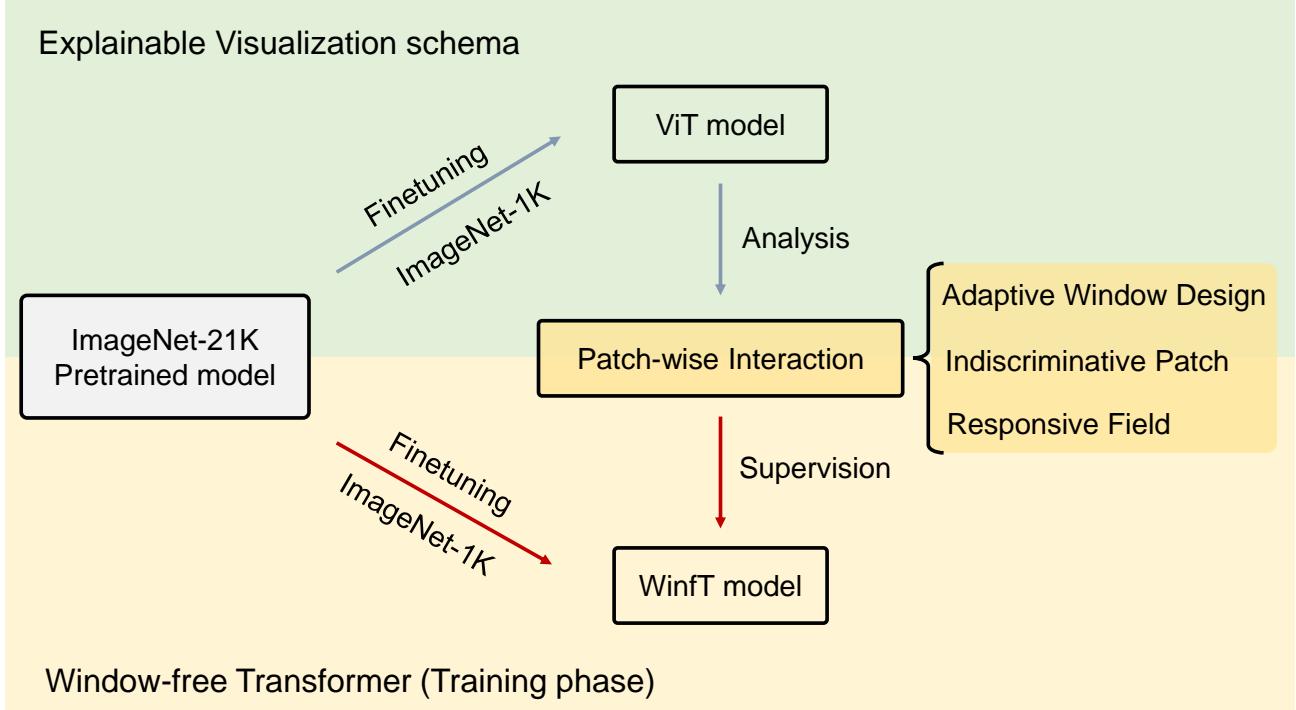


Figure 10. Schematic illustration of the experimental procedure. Notably, **the WinfT model do NOT need the supervision of responsive fields during inference**.

2. Visualization Analysis of Responsive Field

We adopt the adaptive window design schema and indiscriminative patches to define the responsive field (Sec. 4.3), and further analyze and interpret the patch-wise interaction. Then, we utilize the statistic analysis of responsive field tendency and size for studying the patch interactions in ViT. Fig. 11 and Fig. 12 are the higher resolution versions of Fig. 6. We visualize the tendency by directly drawing the direction vector of each patch's responsive field as a pixel in the optical flow.

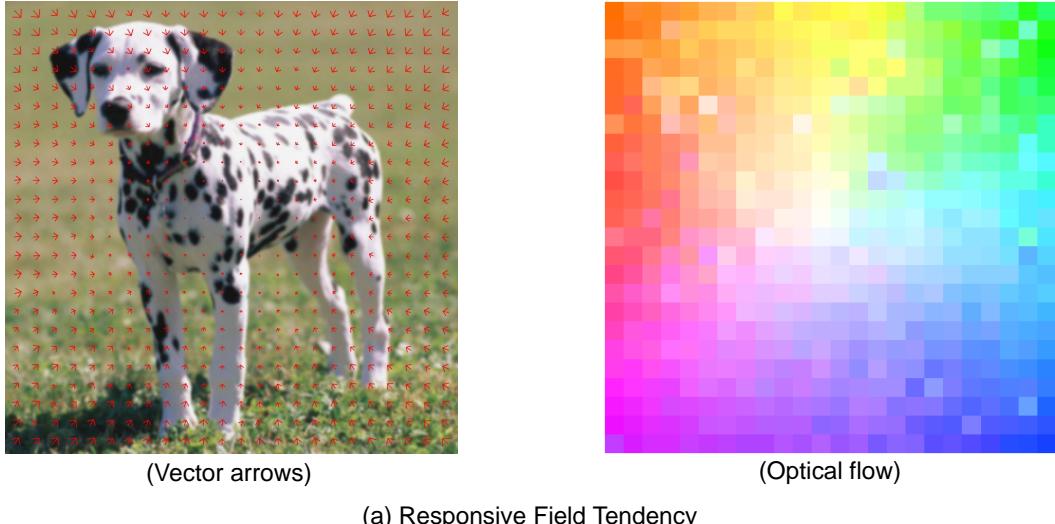


Figure 11. The visualization interpretation of the responsive field tendency by using vector arrows and optical flow.

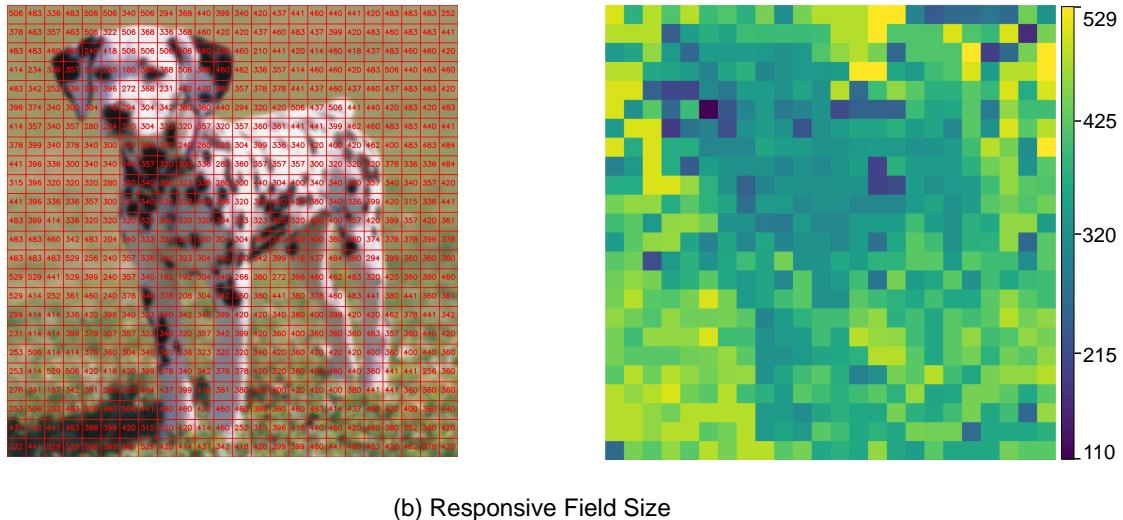
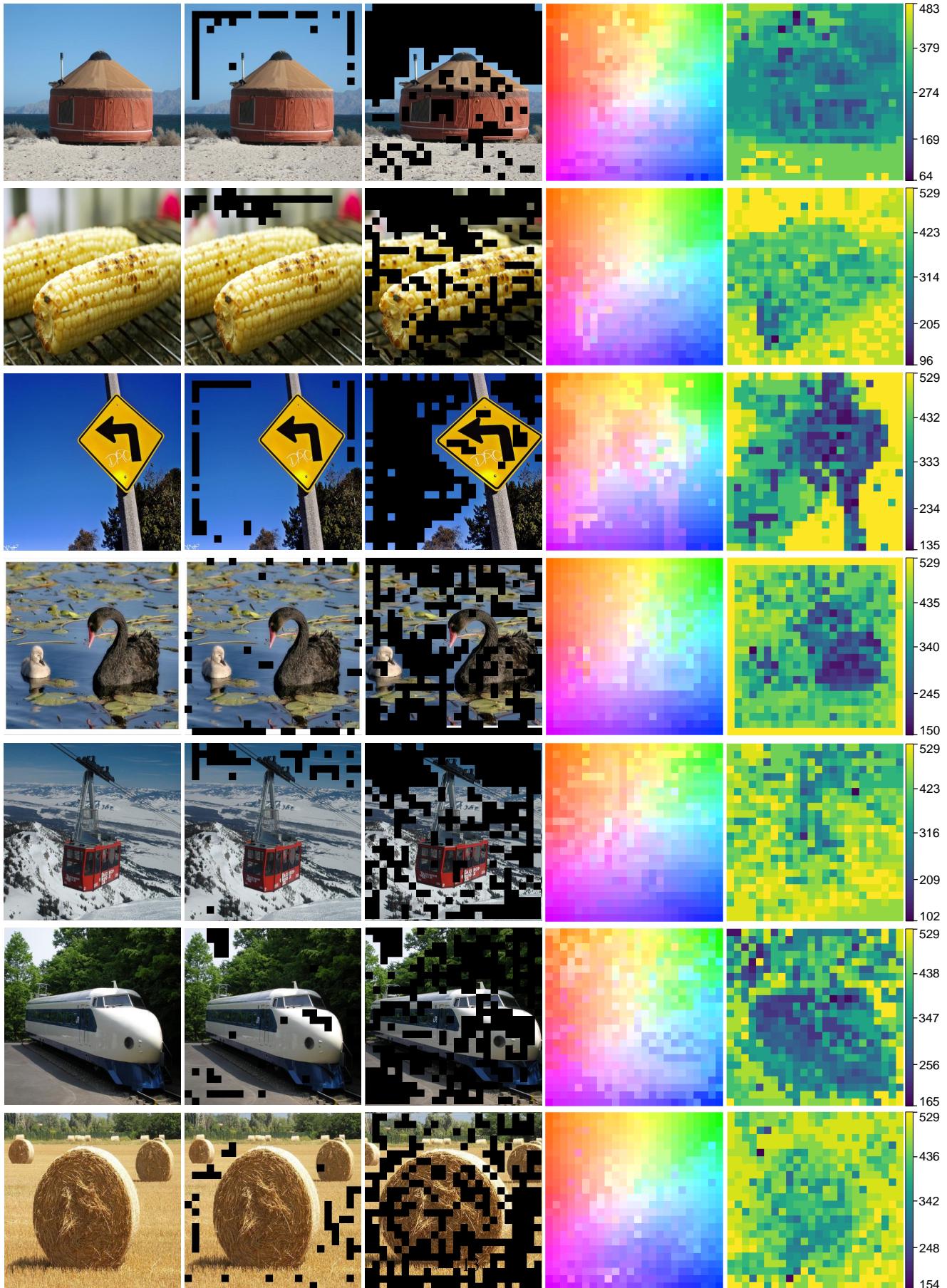


Figure 12. The visualization interpretation of the size of responsive field.

3.Extended Visualization Results

Combined, we visualize more samples in ImageNet dataset to verify the efficacy of our proposed explainable visualization schema below.





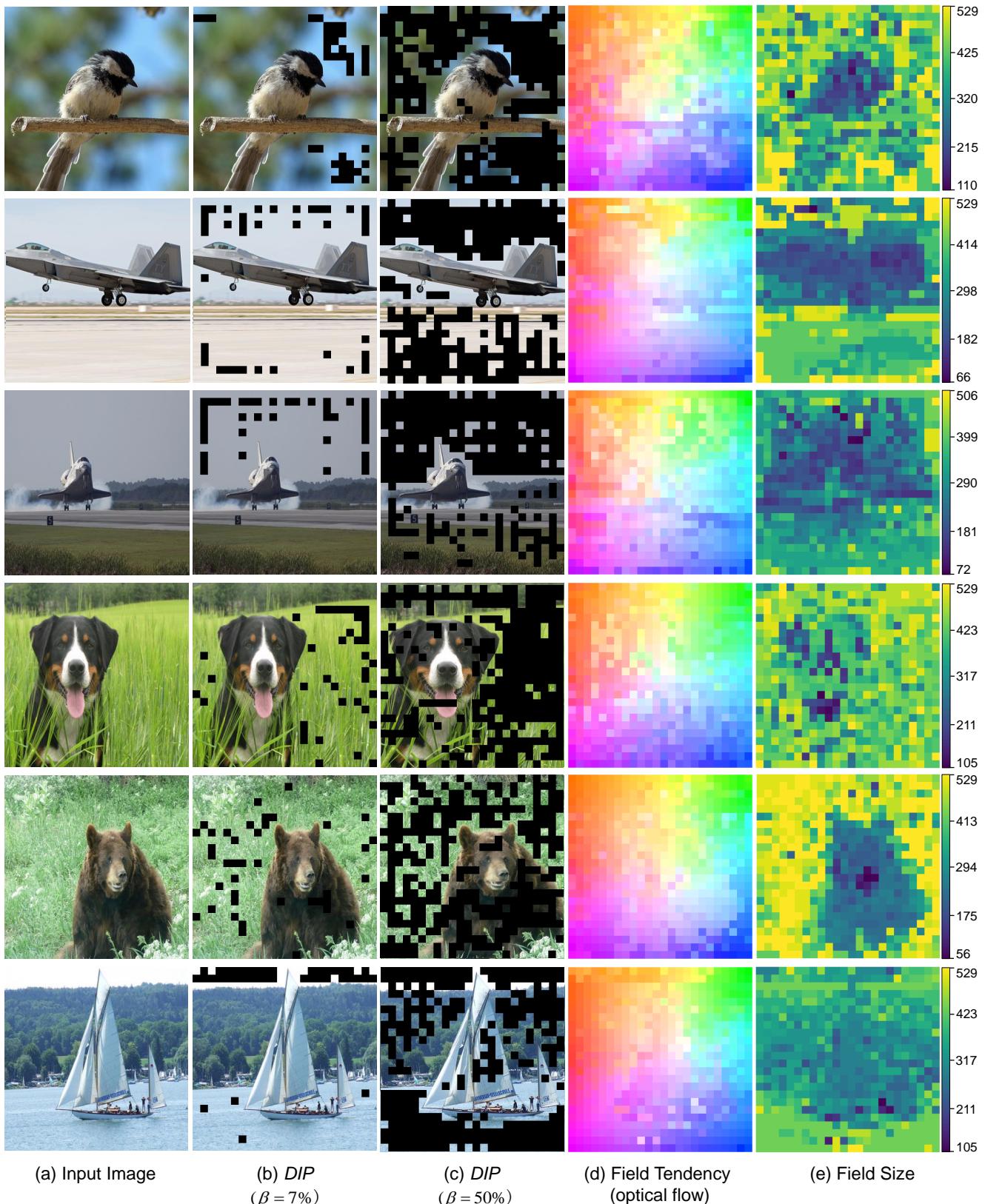
(a) Input Image

(b) DIP
($\beta = 7\%$)

(c) DIP
($\beta = 50\%$)

(d) Field Tendency
(optical flow)

(e) Field Size



4. Visualization of Predicted Attention Windows

We select two samples (large & small objects) to visualize the ground-truth attention windows (b) and the predicted attention windows of WinfT (c) in Fig. 13. 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 l -th layer of WinfT.

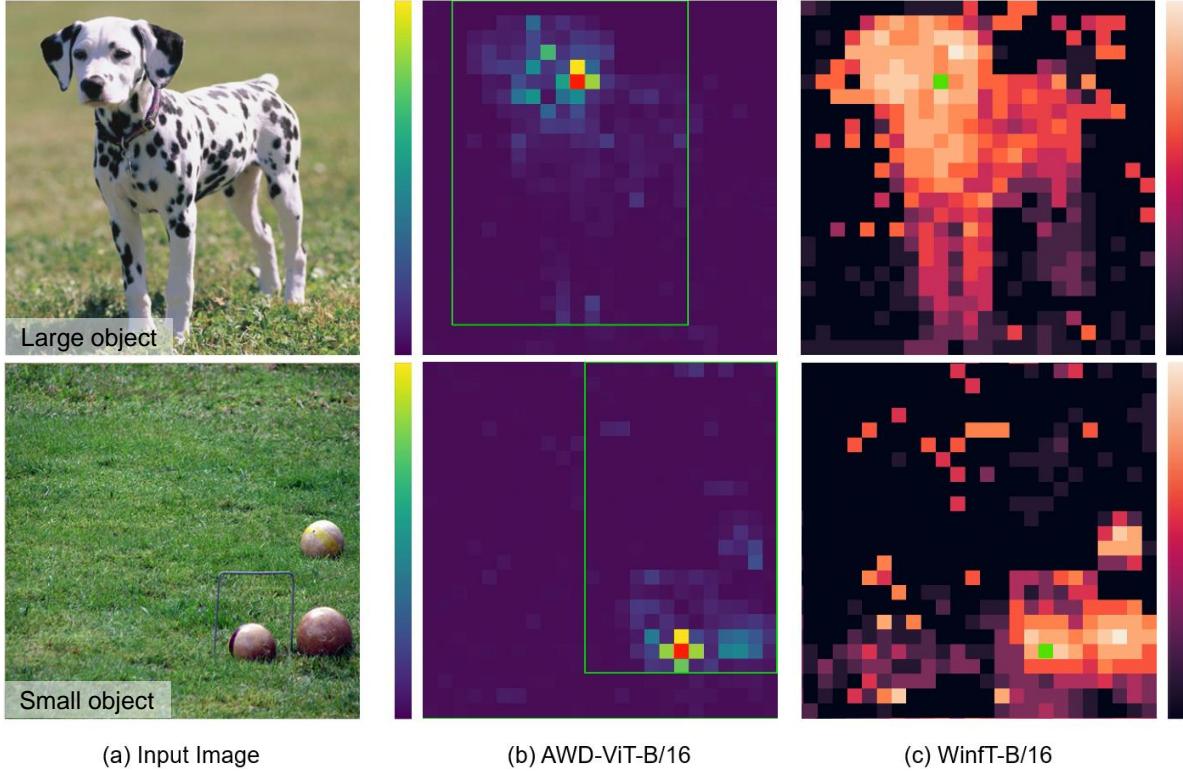


Figure 13. Illustration of the responsive field from ViT (b) and the predicted attention mask in WinfT (c).