IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 April 2024, accepted 18 April 2024, date of publication 24 April 2024, date of current version 6 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3393152

RESEARCH ARTICLE

One-Stage Detection Model Based on Swin Transformer

TAE YANG KIM, ASIM NIAZ, JUNG SIK CHOI, AND KWANG NAM CHOI
Department of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Kwang Nam Choi (knchoi@cau.ac.kr)

This work was supported in part by the Chung-Ang University Research Scholarship Grants, in 2023; and in part by the Ministry of Science and Information and Communication Technology (ICT) and National IT Industry Promotion Agency (NIPA) through the High Performance Computing (HPC) Support Project.

ABSTRACT Object detection using vision transformers (ViTs) has recently garnered considerable research interest. Vision Transformers execute image classification through a multi-head attention-based MLP head and post-image segmentation into patches. However, conventional models prioritize object classification over predicting bounding boxes crucial for precise object detection. To address this gap, a two-stage detector has been devised based on Transformers, which initially extracts feature maps via a pre-trained CNN model. In contrast, our research introduces a one-stage object detector founded on the Swin-Transformer architecture. This one-stage detector adeptly performs simultaneous object classification and bounding box prediction employing a pure Swin-Transformer Encoder Block, obviating the need for a pre-trained CNN model. Our proposed model is trained, validated, and evaluated on the COCO dataset comprising 82,783 training images, 40,504 validation images, and 40,775 test images. The proposed model showed average precision (AP) 30.2% performance improvement by 5.59% compared to the performance evaluation of the existing ViT-based 1-stage detector.

INDEX TERMS Attention, computer-vision, object detection, transformer network, single-stage detection.

I. INTRODUCTION

The prevailing landscape of object detection techniques [1], [2], [3], [4] in computer vision predominantly revolves around Convolutional Neural Network (CNN) architectures [5], [6]. However, a transformative shift has occurred with the introduction of the Vision Transformer (ViT) [7], [8], which has reimagined the Self-Attention-based Transformer [9], originally devised for natural language processing and ushered in a new era of computer vision research. The fundamental ViT [7] model, designed primarily for image classification tasks, necessitated supplementary components for enabling object detection.

Within the realm of object detection, a delicate balance between prediction accuracy and processing speed dictates the choice of technique. Depending on the specific processing approach, the emphasis is placed on either prediction accuracy or processing efficiency, guiding the adoption of the appropriate strategy. In the two-step detector paradigm, an initial feature map is extracted to define a Region of Interest (RoI), representing the potential location of an object. Subsequently, object classification is performed on this feature map, alongside Bounding Box Prediction for the identified object. Conversely, the one-stage detector methodology encompasses the extraction of RoIs, object classification, and bounding box prediction, all in a single step.

The dichotomy between these approaches lies in the fact that while two-stage detectors offer commendable performance, they often exhibit slower image processing speeds. Recent object detection models utilizing the ViT [7] framework typically involve a two-step detection process, wherein a pre-trained CNN model acts as the backbone for feature map extraction.

The inherent ViT model employs patch tokenization to segment input images into discrete components, subsequently establishing correlations between these patches via self-attention [10]. Classification is executed by identifying

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Sharif.

the patch with the highest Attention Score and related patches, determined through the Sigmoid function. To adapt this mechanism for object detection, the ViT [7] model necessitates the integration of a pre-trained CNN model as the backbone. This backbone's role is to extract regions of interest. The ViT [7] model then handles object classification for these regions and predicts bounding boxes accordingly.

Existing ViT [7]-based object detection models utilize the CNN-based RetinaNet [8], introduced by Facebook AI Research, as the backbone for extracting regions of interest. Subsequent object detection hinges on ViT [8]-based classification for these extracted regions. Notably, the ViT [7]-based object detection model introduced in the ''You Only Look at One Sequence'' [11] proposed by Fang et al. adopts a one-step detection approach, a departure from the previously established two-step detectors. However, it is essential to consider the trade-off between object detection prediction accuracy and processing speed. In this context, the YOLOS [11] model, a form of first-stage detector, demonstrates processing efficiency conducive to real-time object detection. Yet, in contrast to conventional two-stage detectors, its predictive performance falls short.

This paper presents a novel object detection model that improves upon the existing ViT-based first-stage detector by utilizing the Swin-Transformer as the backbone. Notably, this model stands out for its independence from pre-trained CNN models, employing the inherent Multi-head Attention architecture and Multi-layer Perceptron (MLP) for feature extraction and object detection in a unified one-step approach. To be specific, the contributions of this work can be summarized as follows:

1) **Innovative Object Detection Approach:** We introduce a novel object detection model that departs from the conventional two-step detection process and utilizes the Swin Transformer as its backbone, improving upon the previously established ViT-based first-stage detector.
2) **Elimination of Pre-trained CNN Backbone:** Unlike existing ViT-based object detectors that rely on a pre-trained CNN model for feature extraction, our model dispenses with this reliance. Instead, it harnesses the inherent Multi-head Attention architecture and MLP within the Swin Transformer for feature map extraction, classification, and bounding box prediction, resulting in a streamlined one-step detection process.
3) **Unified One-Step Detection:** Our model offers a unified one-step detection approach that integrates all aspects of object detection, including feature extraction, object classification, and bounding box prediction, in a holistic manner. This approach enhances the overall efficiency of object detection tasks.
4) **Balancing Speed and Accuracy:** We address the trade-off between object detection prediction accuracy and processing speed by providing an efficient real-time object detection solution without significant compromises in predictive performance.

These contributions collectively advance the state-of-the-art in object detection by offering a more efficient and streamlined approach guided by the evolving landscape of computer vision research.

II. RELATED WORK

This section provides an overview of the relevant studies and research that serve as the foundation for our investigation.

A. OBJECT DETECTION

Object detection, a fundamental task in computer vision, entails identifying and localizing objects within images or videos. This field is primarily divided into two categories: the first-stage detector and the two-stage detector, with the latter being commonly employed due to its heightened object detection accuracy. Prominent CNN-based models that fall under this category encompass Faster-RCNN [12], EfficientNet [13], and DenseNet [14]. Nevertheless, the two-stage detector methodology is characterized by its division into a backbone model, responsible for extracting regions of interest, and a classification model. This division entails a drawback, as it demands more extensive computational resources than its first-stage counterpart.

On the other hand, first-stage detectors encompass SSD [15], YOLO [16], and RetinaNet [8]. These detectors hold the advantage of expending less time on object detection compared to second-stage detectors, rendering them suitable for real-time detection scenarios. As various papers have iterated, effective real-time detection should operate at a performance level of 30 frames per second (FPS) or higher [16, 17, 18].

Within the framework of this object detection model, the Intersection over Union (IoU) technique is employed to ascertain the validity of diverse predicted bounding boxes. The IoU technique is defined as follows (1): it revolves around adopting the prediction box with the highest value, based on the ratio between the union and intersection of the predicted bounding box value (Prediction) and the actual value (Ground Truth).

$$IoU = \frac{Ground\ Truth \bigcap Prediction}{Ground\ Truth \bigcup Prediction} \tag{1}$$

B. VISION TRANSFORMER

The Transformer model, a significant advancement in the field of natural language processing (NLP), stands in contrast to recurrent models like Long Short-Term Memory (LSTM) [18] or Gated Recurrent Unit (GRU) [19]. Inspired by the Encoder-Decoder structure of Recurrent Neural Networks (RNN) [20], the Transformer model redefines the approach.

In this context, the term ''encoder'' denotes a model responsible for converting a word representation into a latent vector, while a ''decoder'' converts this latent vector into a different representation. The ViT model marked a pivotal moment by extending this concept to images, applying it to a multi-head attention classification task, all without resorting

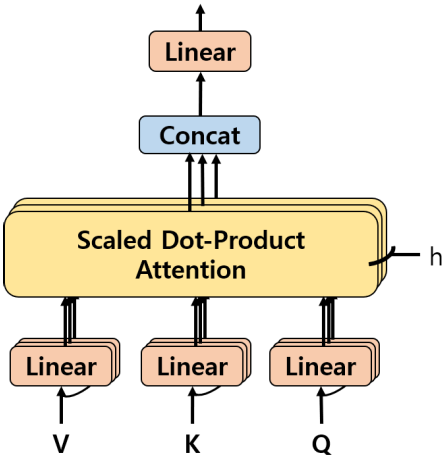

FIGURE 1. Multi-head attention is a key component of the transformer model that enhances parallelism and efficiency by dividing the input sequence into multiple sub-sequences and performing attention on each of them separately. Each attention head considers the input sequence from a different perspective, and the final attention weight is obtained by combining the attention weights from all the heads.

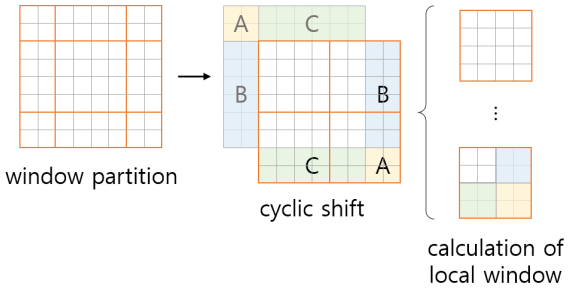

FIGURE 2. The operation of cyclic shift used in swin-transformer is explained. Through cyclic shift, the unequally divided local boundary was logically changed to be the same so that only units of the same size were used to calculate attention.

to a CNN. Figure 1 provides a visual representation of the multi-head attention mechanism.

Since then, numerous models have emerged that rival or even surpass the performance of existing CNN models. Examples include the Swin-Transformer [3], ViViT [21], DeiT [22], and DETR [23], all rooted in the ViT model, and all contributing to the ongoing exploration of enhanced image analysis methodologies.

C. SWIN TRANSFORMER

The existing ViT [7] model performs image classification utilizing an encoder that adheres to a consistent patch size. In contrast, the novel Swin-Transformer [24] model has introduced a versatile mechanism capable of classifying images of varying dimensions. This is achieved through the application of diverse patch sizes and the merging of patches using multiple ratios. To facilitate this process, a local window method was devised for the computation of these varied patch sizes. This local window serves as a boundary that segregates patches into four equitably sized quadrants. However, this approach introduces a limitation in the calculation of attention scores between patches that overlap the boundary.

To overcome this limitation, a dynamic approach is adopted. The local window is redefined to encompass sizes such as (1×1), (1×2), (2×1), and (2×2). This recalibration is facilitated using the cyclic shift technique [24], which effectively changes the sizes during calculations. This innovative strategy reduces computational costs by shifting and logically aligning patches to the same dimensions. Figure 2 offers a visual depiction of this process.

Empowered by this strategy, the Swin-Transformer [24] model is capable of detecting objects by hierarchically combining ViT [7] elements of various sizes. This fusion of multiple patches enables a more accurate assessment of inter-patch relationships. Notably, the Swin-Transformer [24] can identify both small and large objects due to the gradual increment of patch size from (4, 4) to (32, 32), facilitating a comprehensive image classification process.

Furthermore, this approach offers a more efficient means of attending to the relationships between merged patches, as opposed to the existing method that involves segmenting objects based on patch size and subsequently obtaining self-attention by cyclically altering the local window's shape [24].

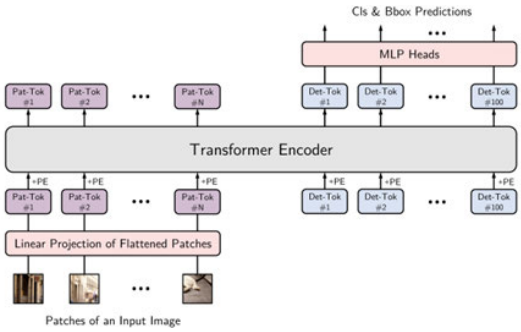

FIGURE 3. This figure shows the encoder structure of the YOLOS [11] model. YOLOS [11] inherited the DETR [23] model, but by adding the [DET] token, it operates as a 1-stage detector, unlike the existing DETR [23].

D. YOLOS

Object detectors employing the established ViT [7] model typically function as two-step detectors, incorporating a pre-trained CNN model as their backbone. Diverging from this convention, the YOLOS [11] model takes a distinct approach by serving as a first-stage detector, utilizing the DeiT [7] model as its backbone.

In the training phase, the DeiT [7] model addresses the challenge of learning from a substantial pre-training dataset. Notably, the class token featured in the original ViT [7] model is eliminated, making way for a detection token specifically introduced for object detection tasks. This detection token traverses through the Transformer Encoder Block. Subsequently, the detection token is subjected to classification

using a MLP, followed by bounding box prediction via a bipartite matching loss function. To further enhance accuracy, the YOLOS [11] model leverages Hungarian Loss [25] to improve the prediction of boxes and their corresponding correct answer sets. The structural layout of the YOLOS [11] model can be visualized in Figure 3.

In YOLOS [11], the existing ViT [2] model undergoes pre-training using the JFT-300M [26] dataset, followed by Fine Tuning using the ImageNet-1K [20] dataset. The model's efficacy is subsequently gauged using the COCO dataset [28]. While YOLOS [11] does exhibit a drawback in terms of detection rates, falling short of existing CNN-based models and two-step object detectors based on ViT [7], it underscores the potential of achieving satisfactory object detection performance exclusively through a pure transformer architecture.

III. PROPOSED METHOD

A. YOLOS

Extensive research in the realm of image analysis networks has yielded a multitude of investigations, each striving to apply network models grounded in object detection, recognition, segmentation [29], [30], [31], and ViT [7]. Of particular significance are the endeavors focused on replacing established CNN models commonly utilized in object detection tasks. This evolution has culminated in the development of a two-stage detector variant employing both existing CNN models and the Swin-Transformer [24] as backbones. Notably, these models have recently exhibited superior performance outcomes. Concurrently, the exploration of a real-time image processing solution takes form in the shape of a one-stage detector, with the YOLOS [11] model representing this approach.

Building on this foundation, the present paper introduces an innovative departure from the conventional practice of employing pre-trained CNN models for region of interest extraction. Instead, our approach involves the Swin-Transformer Encoder [11] configuration, seamlessly integrating multi-head attention-based feature map extraction and classification processes within a unified framework. This amalgamation enables simultaneous execution, yielding a streamlined and effective solution to object detection challenges.

B. MODEL ARCHITECTURE

Achieving superior accuracy compared to the two-stage detector is a formidable challenge for the single-stage detector. However, the single-stage detector boasts a distinct advantage in terms of image processing speed, rendering it a fitting choice for real-time applications. The pioneering YOLOS [11] model presented an innovative approach by proposing a first-stage detector, a departure from the conventional two-stage paradigm. Utilizing the DeiT [22] model as its backbone; this model exhibited modest performance while maintaining comparable accuracy across the dataset.

In this study, we have devised a network to enhance the performance of the YOLOS [11] model. This was achieved by integrating the encoder structure from the Swin-Transformer [24] into the YOLOS [11] model's architecture, leading to an improved design. The holistic network architecture is vividly depicted in Figure 4. Please see Table 1 for the proposed model architecture overview.

The process commences with the segmentation of input data into patches of predetermined dimensions, a step referred to as Patch Partition. Pose embedding and detection tokens are subsequently introduced to embed the patch order and encapsulate object-related information. This amalgamated data is then channeled through the encoder block. Notably, the encoder block leverages the Swin-Transformer [11] format, thus harnessing patch merging and cyclic shift mechanisms to compute attention scores between individual patches.

Following this, object classification insights and bounding box predictions are linked through MLP. The resulting information is then employed to ascertain object types and bounding box coordinates, achieved through the innovative application of the bipartite matching loss function. This multi-step process culminates in a robust and efficient object detection methodology.

C. TRANSFORMER ENCODER BLOCK

As a means to enhance model performance, the conventional transformer encoder block is substituted with the encoder block originating from the Swin-Transformer [24] structure, within the framework of the existing YOLOS [11] model. The Swin-Transformer [24] introduces a unique approach, delimiting local attributes by delineating boundaries using a local window. This methodology paves the way for the creation of a hierarchical model that progressively augments patch dimensions through patch merging. Consequently, this architecture facilitates the computation of locality attention scores across patches of varying sizes, thereby yielding model enhancement [24]. Notably, the Swin-Transformer [24] serves as the backbone of the proposed model.

Figure 5 elucidates the mechanics of the proposed model. The process commences with the segmentation of the input image into patches of size (4×4) through patch segmentation. An additional detection token tailored for object detection and classification is integrated into each split patch, followed by the implementation of pose embedding to ascertain the sequence of patches and tokens. Subsequently, the amalgamated data traverses the Swin Transformer Encoder.

During this phase, the Swin Transformer Encoder establishes four local windows, each maintaining a consistent (2:2) ratio relative to the received image patch. These local windows serve as the foundation for computing the locality attention score. Moreover, through a process involving the transformation of local window boundaries, the size ratios are adapted to (1:1), (1:2), (2:1), and (2:2), effectively converting the local window to a masked configuration with a 2:2 ratio. Following this, the region window size is reconstructed post-patch merging, contributing to an augmentation in

TABLE 1. Proposed object detection model architecture: A summary of our single-stage object detection model that combines YOLOS with a swin-transformer encoder. It includes patch handling, pose embedding, an encoder block, MLPs for classification and bounding box prediction, and a bipartite matching loss for improved efficiency and accuracy.

Component	Description
Object Detection Paradigm	Single-stage detection
Backbone	YOLOS with Swin-Transformer encoder
Patch Partition	Segmentation of input data into patches
Pose Embedding	Introduction of pose embedding and detection tokens
Encoder Block	Swin-Transformer format for patch merging and cyclic shift mechanisms
Object Classification and Bounding Box Prediction	Multi-Layer Perceptrons (MLP)
Loss Function	Bipartite matching loss function

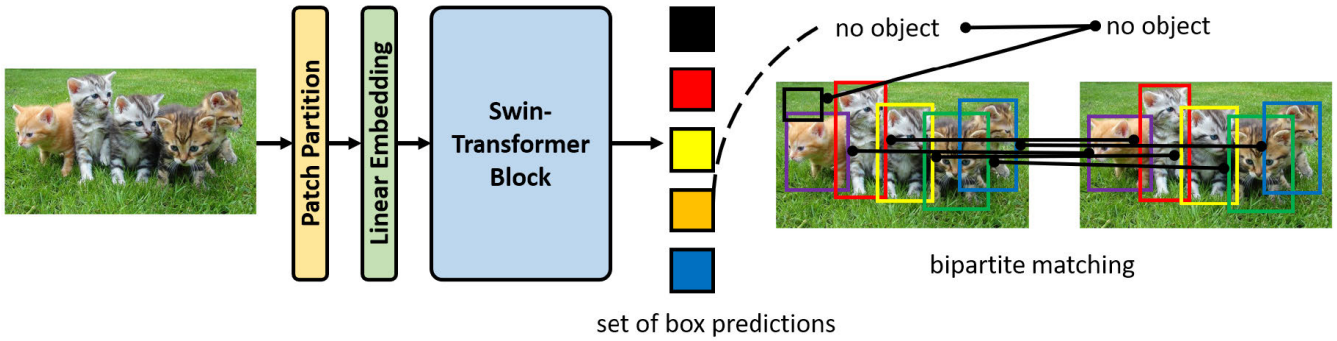

FIGURE 4. Overall network structure. It is based on YOLOS and has been changed to a swin-transformer type encoder. Also, based on the added [DET] Token, a pair predicted by (class, bbox) is output, and the result is output by bipartite matching. Afterwards, the training of the model proceeds based on the Hungarian loss function.

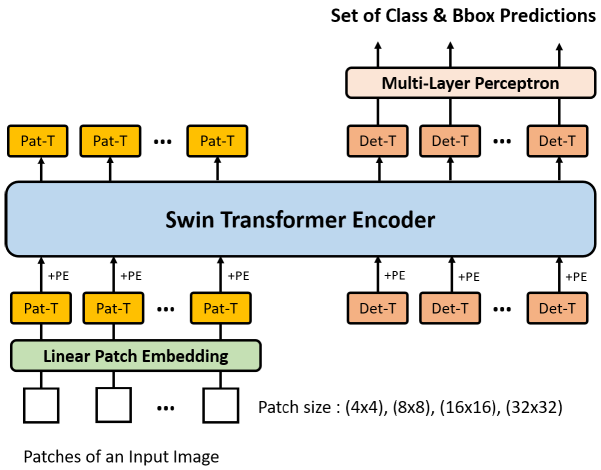

FIGURE 5. This is an encoder block based on swin-transformer. Before passing through the encoder, a position patch and [DET] token are added through linear patch embedding. After passing through the encoder, the [DET] Token is used to predict the class and bbox through the multi-layer perceptron.

size. Capitalizing on the configured regional window, the aforementioned sequence is reiterated, culminating in the computation of attention scores.

Subsequent to this intricate process, the detection token progresses through an MLP for parallel object classification and bounding box prediction. The resultant predictions, encompassing both bounding box coordinates and object type classifications, are conjoined as a pair. A judicious selection of optimal prediction pairs is orchestrated via a sophisticated bipartite matching loss function. This comprehensive methodology enables refined object detection outcomes with enhanced accuracy and performance.

D. COST FUNCTION

The Attention Function constitutes a process that maps a given query and a set of key-value pairs into an output value. Each component—query, key, value, and output—has a vector format. The output value is determined by computing the weighted summation of the values, where each value's weight is established via the compatibility function between the query and its corresponding key.

Within the framework of Scaled Dot-Product Attention, the input encompasses a query and key vectors of dimension "dk," along with a value vector of dimension "d2." To derive the attention value, the dot product of all input keys and queries is computed. Subsequently, this product is divided by the square root of the input dimension "root-d." The obtained result undergoes further processing through the softmax function, yielding the weight attributed to the value. This weight is then employed to scale the value. In the context of concurrently calculating the attention function, the queries are grouped as a matrix denoted as Q, while the keys and values are grouped into matrices K and V, respectively. The calculation formula governing the output matrix is defined

by (2).

$$Attention = SoftMax(QK^T/\sqrt{d} + B)V \tag{2}$$

The process of Scaled Dot-Product Attention operates on an individual patch, yielding an attention value. This value is then organized into a single head. To create Multi-Head Self Attention (MSA), these individual heads are amalgamated using a concatenation function, followed by multiplication with an output weight. This concerted process optimizes the attention mechanism, enabling it to capture and comprehend complex relationships within the data.

$$Head = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

represents the individual head of the Multi-Head Self Attention (MSA) mechanism. It involves the attention calculation for a specific set of queries (Q), keys (K), and values (V), with each being transformed by learnable weight matrices (W_i^Q, W_i^K, W_i^V). This process captures intricate relationships within the data, providing a focused understanding of the input's contextual information.

$$MultiHead(Q, K, V) = Concat(head_i)W^O \tag{4}$$

describes the formation of Multi-Head Self Attention (MSA) by concatenating individual heads ($head_i$) and multiplying the result by an output weight matrix (W^O). This operation is crucial for capturing diverse features and relationships within the input data, enhancing the model's ability to comprehend complex patterns. The concatenation ensures a comprehensive representation of attention across multiple heads, contributing to the overall effectiveness of the attention mechanism in the Swin Transformer Encoder.

To effectively model the Swin Transformer Encoder, we introduced an innovative approach to computing Self-Attention within a confined local window. The initial equation delineates the lower limit time complexity for a fundamental Multi-Head Self Attention (MSA) calculation, as depicted in (5). This calculation considers variables such as the input's height and width, as well as the number of dimensions represented by C.

The Swin-Transformer employs a localized window-based Multi-Head Self Attention (W-MSA) mechanism to optimize efficiency. This W-MSA is encapsulated by the second formula within (5), accounting for the (M x M) patches encompassed by the window. By incorporating these equations, our methodology adeptly addresses the challenges associated with modelling the Swin Transformer Encoder while maintaining computational efficiency.

$$\begin{aligned} \Omega(MSA) &= 4hwC^2 + 2(hw)^2C, \\ \Omega(W-MSA) &= 4hwC^2 + 2M^2hwC \end{aligned} \tag{5}$$

The conventional ViT [7] model operates by transforming input data into a one-dimensional token format for processing. However, to effectively handle two-dimensional input images, YOLOS [11] adopts an alternate approach, modifying the image shape from "X" to "X-P" [4]. In this notation, (P x P) signifies the size of the patch, and "X-P" represents a patch of size p. By partitioning the two-dimensional image into "X-P" patches and subsequently flattening it, the dimensionality is shaped into D through a learnable linear projection.

To further refine this sequence, YOLOS [11] introduces 100 detection tokens labeled "X-det," collectively constituting a single sequence. Added to this sequence is the patch embedding denoted as "E-pe," thereby culminating in the final sequence structure as outlined in (6). This transformation methodology, implemented in ViT and YOLOS, demonstrates a strategic approach to processing images and sequences, facilitating enhanced efficiency and accuracy in image analysis.

$$Z_0 = [X_p^1 E; X_p^2 E; \ldots; X_p^n E; X_{DET}^1; \ldots; X_{DET}^{100}] + E_{PE} \tag{6}$$

The process of calculating attention unfolds through the manipulation of sequences, initiated by the outcome of (6) as the initial input. This sequence-driven attention derivation is expressed by (7). The initial equation within (7) commences with the sequential application of Layer Normalization (LN) to the input sequence. This is succeeded by Window Multi-head Self Attention (W-MSA) executed within a localized window, and the result is augmented by the inclusion of existing input values to yield "Z-l."

Subsequently, "Z-l" serves as the input for the second equation within (7). This equation entails a successive application of Layer Normalization (LN) and MLP, and once again, the outcome is fused with "Z-l." During this phase, an additional computation is undertaken involving the shifted window Multi-Head Self Attention (SW-MSA). This mechanism is responsible for modifying the local window boundary for the previously calculated values. The third equation within (7) captures this process, which is executed similarly to the previous steps. These computations culminate in the derivation of the subsequent sequence denoted as "Z-l+1." This intricate sequence-based approach effectively realizes attention calculation, facilitating a comprehensive transformation of data for enhanced analysis and understanding.

$$\begin{aligned} \hat{z}_l &= W - MSA(LN(z_{l-1})) + z_{l-1}, \\ z_l &= MLP(LN(\hat{z}_l)) + \hat{z}_l, \\ \hat{z}_{l+1} &= SW - MSA(LN(z_l)) + z_l, \\ z_{l+1} &= MLP(LN(\hat{z}_{l+1})) + \hat{z}_{l+1}, \end{aligned} \tag{7}$$

The prevalent one-step detector model relies on a non-maximum suppression (NMS) loss function for refinement. The NMS mechanism revolves around calculating the Intersection over Union (IoU) for all projected bounding boxes in relation to ground truth boxes. Subsequently, it retains solely the bounding box with the highest score while discarding others. Despite its efficacy, this approach demands substantial computational resources due to its complex nature.

To address this concern, our paper introduces an innovative solution. We incorporate the bipartite matching loss

function drawn from DETR [23], a method that mitigates computational expenses and elevates model prediction speed. This strategy functions by juxtaposing a prediction pair, encompassing both object type classification and bounding box prediction—yielded through an MLP-driven forecast of the detection token—with the corresponding true values. The process hinges on the selection of a prediction pair that minimizes the prediction error, ultimately optimizing model accuracy and efficiency. This transformational approach is encapsulated by (8), offering a comprehensive understanding of the mechanism's operation and effectiveness.

$$\hat{\sigma} = \underset{\sigma \in R}{\operatorname{argmin}} \sum_{i}^{N} L_{match}(y_i, \hat{y}_{\sigma(i)}) \tag{8}$$

In (8), the term "L-match" denotes the object type in (9) and encapsulates the degree of loss incurred due to the bounding box's dissimilarity. This loss value serves as a vital component in our methodology.

In (9), the introduced permutation "sigma" defines an arrangement that optimizes the pairing of object type predictions and their corresponding bounding box predictions. This permutation is instrumental in training the model through the application of a Hungarian loss function, as delineated in (9). The utilization of the Hungarian loss function facilitates the division of the object type prediction and bounding box prediction values, aiding the convergence towards an optimal solution.

In this context, the variable "c" denotes the class label, while "b" is composed of the image size represented within the range of 0 to 1, in addition to the relative bounding box's center, height, and width. The parameter "p" signifies the prediction probability, while "Pi" corresponds to padding introduced to account for non-existent objects. By incorporating these elements, our approach demonstrates a meticulous consideration of object characteristics and their predictions, thereby fostering a robust and accurate learning process.

$$L_{Hungarian}(y, \hat{y}) = \sum_{i=1}^{N} [-log\hat{p}_{\hat{\sigma}(i)}(c_i) + log_{\{c_i \neq \pi\}} L_{bbox}(b_i, \hat{b}_{\hat{\sigma}}(i))] \tag{9}$$

Table 2 gives a brief view of each of the equations in the proposed model.

TABLE 2. Comprehensive overview of equations: Each equation is briefly explained to highlight its role in model architecture and efficiency.

Equation	Description
(2)	Scaled Dot-Product Attention for calculating matrix attention values based on input keys, queries, and SoftMax-scaled values.
(3)	Individual head of Multi-Head Self Attention (MSA) calculating attention for specific queries, keys, and values with learnable weight matrices.
(4)	Formation of MSA by concatenating individual heads and multiplying by an output weight matrix, crucial for capturing diverse input features.
(5)	Lower Limit Time Complexity for MSA and Window MSA, considering input dimensions and represented by C.
(6)	Transformation process in YOLOS model, converting a 2D image into a sequence with patches, detection tokens, and patch embedding.
(7)	Sequence-driven attention process involving LN, W-MSA, MLP, and SW-MSA.
(8)	Bipartite matching loss function for object detection inspired by DETR, optimizing efficiency by minimizing prediction error.
(9)	Details of the Hungarian loss function within bipartite matching loss, involving class labels, image size, bounding box information, prediction probability, and padding for non-existent objects.

E. ENHANCED MODEL WITH SWIN-TRANSFORMER INTEGRATION

This section provides a more detailed explanation and step-by-step breakdown of the modifications made to the YOLOS model when integrating the Swin-Transformer encoder:

- **Introduction of Swin-Transformer Encoder:** The initial modification involves replacing the conventional YOLOS model's encoder with the Swin-Transformer encoder. This transition aims to leverage the unique capabilities of the Swin-Transformer, such as its attention mechanisms and hierarchical patch merging, to enhance feature extraction and representation.
- **Patch Partition and Token Embedding:** Before passing through the Swin-Transformer encoder, the input image undergoes patch partitioning, breaking it down into smaller segments. Additionally, a specialized detection token ([DET]) is introduced to each patch, serving as a key element for subsequent object detection. This process is followed by embedding the positional information and [DET] token into the patches through linear patch embedding.
- **Swin-Transformer Encoder Operation:** The Swin-Transformer encoder operates on the patch-embedded data, establishing local windows and computing attention scores between individual patches. This process involves a unique approach to attention computation within confined windows, adapting local window boundaries and employing a masked configuration. This mechanism facilitates the extraction of complex relationships within the data, contributing to improved feature representation.
- **Multi-Head Self Attention and MLP:** Within the Swin-Transformer encoder, Multi-Head Self Attention (MSA) is applied to the individual patches, capturing intricate dependencies. The attention values from multiple heads are then concatenated and multiplied with an output weight, creating a comprehensive attention mechanism. Following this, the detection token progresses through the MLP for parallel object classification and bounding box prediction.
- **Bipartite Matching Loss Function:** The proposed method introduces a novel bipartite matching loss function for refined object detection. This loss function is inspired by DETR and involves pairing object type

predictions and bounding box predictions, minimizing the prediction error. The Hungarian loss function is employed to select optimal prediction pairs that optimize both object type classification and bounding box coordinates.

In summary, the integration of the Swin-Transformer into the YOLOS model brings about significant improvements in object detection. From replacing the traditional encoder to introducing patch partitioning, token embedding, and unique Swin-Transformer operations, each step enhances the model's feature extraction capabilities. The use of Multi-Head Self Attention and MLP adds finesse to capturing intricate dependencies, while the innovative bipartite matching loss function optimizes both object type classification and bounding box coordinates. This holistic approach not only boosts the YOLOS model's performance but also sets the stage for future advancements in computer vision.

IV. EXPERIMENTS

The conducted experiment in this study centres on object detection, employing the Microsoft-provided COCO dataset [28]. This dataset encompasses an expansive collection of 82,783 training samples, along with 40,504 validation samples and 40,775 test samples. The defining characteristic of the COCO dataset [28] lies in its diverse assortment of objects across varying sizes within the training data. Each image within the dataset is distinctly categorized, rendering it an ideal resource for object detector learning and validation [28]. Notably, the training and validation datasets incorporate annotations that include object bounding box coordinates, object areas, and classifications of 80 object types.

The experiment was conducted utilizing three RTX 3090 GPUs, with the model being trained to employ a batch size of eight over the course of 100 epochs. To ensure a comprehensive comparative analysis, the experiment commenced by fine-tuning the established YOLOS [11] model using the COCO dataset [28]. Subsequently, the performance of the proposed model was evaluated within the same experimental environment, providing a direct benchmark against the established baseline. This systematic approach yields a robust assessment of the proposed model's capabilities and improvements within the context of object detection tasks.

A. SETUP

Efforts to apply the ViT [7] model to the domain of object detection have primarily focused on two-stage detectors that employ pre-trained CNN models as feature map extractors. While the Swin Transformer [24] model combined with RetinaNet [17] has demonstrated superior average precision (AP) compared to conventional CNN models, the ViT [7] model faces challenges due to its extensive parameter count resulting from performing Multi-Head Self Attention (MSA) across the entire CNN-extracted feature map. As parameters increase, computational costs escalate, impeding its viability as a real-time detector.

TABLE 3. Setup for experimentation of YOLOS and the proposed model.

Model	backbone	Layers (Depth)	Params (M)	Heads
YOLOS [11]	Deit-Ti [22]	12	6.9	3
Proposed model	Swin-Ti [7]	12	8.3	3

To address these issues, the YOLOS [11] model introduced a first-stage detector paradigm, diverging from the conventional two-stage approach. This strategy aimed to counteract parameter growth and associated processing speed reductions. However, this shift brought to light class imbalance concerns within the training dataset. YOLOS [11] tackled this by utilizing the DeiT [22] model as a backbone, successfully achieving balanced model learning outcomes.

Our experiments aimed to enhance the performance of the model introduced in this paper, employing the Swin Transformer Encoder Block as a first-stage detector. The experimentation followed a configuration akin to YOLOS [4], featuring a batch size of 8, Adam optimizer learning rate of $2.5 \times 10^{-}5$, and a weight decay of $1 \times 10^{-}4$. We enhance data preprocessing by resizing images to 224×224 pixels, applying random horizontal flipping with a probability of 0.5, normalization with a mean and standard deviation of 0.5, adding random Gaussian noise with a probability of 0.1, and color jitter. At the same time, other model hyperparameters, including cosine learning rate schedule, dropout rate of 0.1, patch size of 7 in transformer encoder block, and 100 detection tokens, are detailed to improve reproducibility and understanding of the training process. Table 3 is the tabular illustration of the setup for experimentation of YOLOS and the proposed model.

B. EXPERIMENTS RESULT

This section presents a comprehensive comparison between the experimental outcomes of the YOLOS [11] model and the proposed model within the realm of object detection, utilizing the COCO dataset [28]. All conducted experiments were consistently executed on the same RTX 3090 GPU, ensuring an equitable evaluation platform.

The experimental findings of the YOLOS [11] model revealed an achieved Average Precision (AP) of 28.6, coupled with a Frames Per Second (FPS) rate of 103. Contrastingly, the proposed model demonstrated remarkable improvement, yielding an AP of 30.2 and an FPS of 89. These results underscore the enhanced performance of the proposed model, as it not only surpassed the YOLOS [11] model in terms of AP but also maintained a commendable FPS rate despite the heightened precision. This comparison emphasizes the viability of our approach in achieving superior accuracy while upholding satisfactory processing speed, thus contributing to the advancement of object detection capabilities.

Table 4 presents a comprehensive performance comparison of various object detection models based on key metrics. Each row corresponds to a specific model, providing details

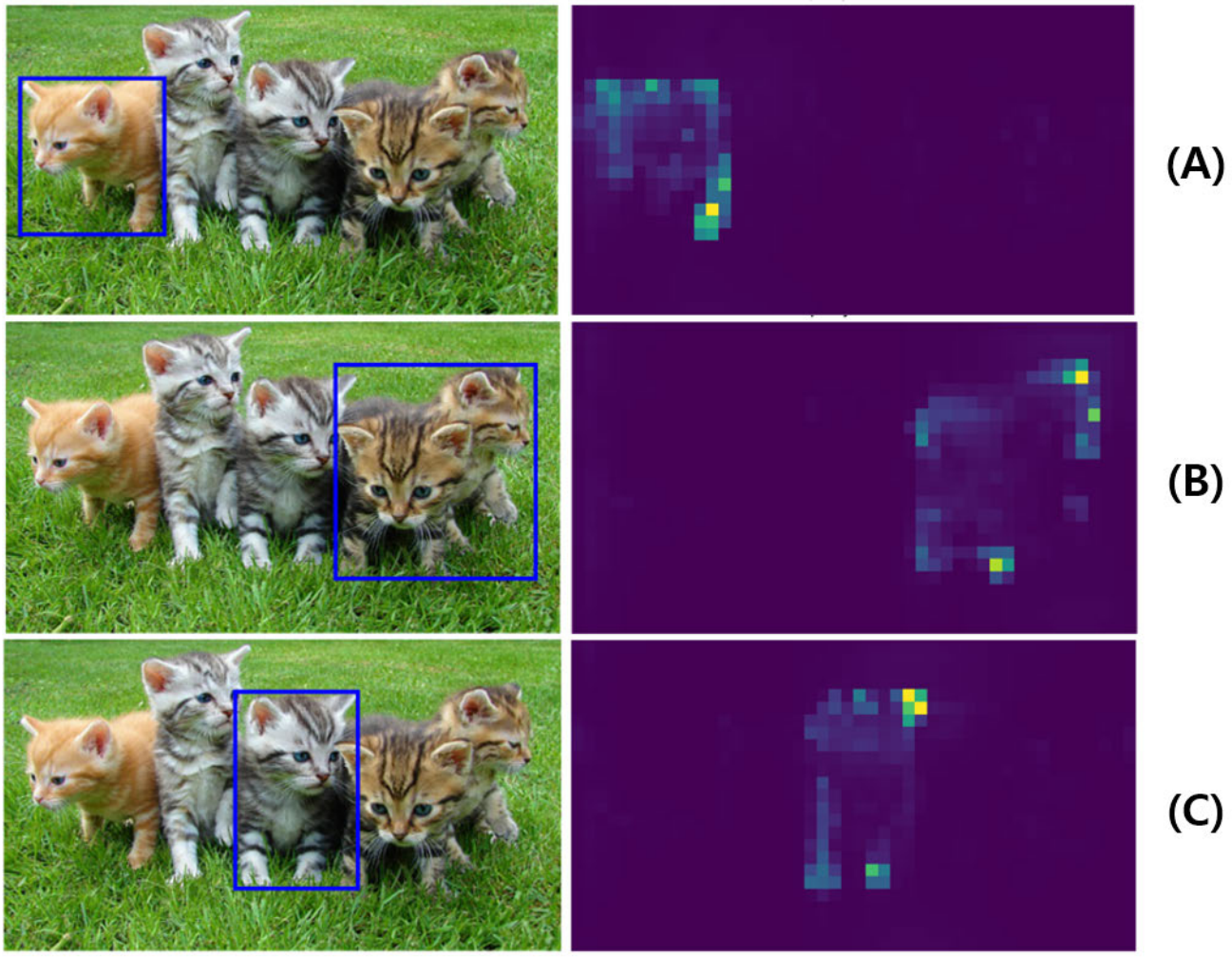

FIGURE 6. Experimental outcomes and attention map illustrating our model's object detection performance, with a focus on the impact near object boundaries.

TABLE 4. Performance comparison of different Object detection models including the proposed method.

Model	Backbone	Input Resolution	Params (M)	FLOPS (G)	Model Size (MBs)	AP	FPS
YOLOv4-Ti	DarkNet53 [16]	224^2	8.9	5.6	33.9	16.6	330
CenterNet	ResNet50 [6]	224^2	15.8	22.7	60.2	28.1	142
DETR \cite{b24}	ResNet50 [6]	224^2	41.1	86.1	205.5	35.3	12
Def. DETR	ResNet18 [6]	224^2	40.4	173	167.0	43.8	19
YOLOS-Ti \cite{b12}	Deit-Ti [22]	224^2	6.5	3.4	27.2	23.1	114
YOLOX-Ti	DarkNet53 [16]	224^2	5.1	6.9	21.3	32.8	90
Proposed model	Swin-Ti [24]	224^2	8.3	3.7	34.8	30.2	89

such as the backbone architecture, number of parameters (in millions), floating-point operations per second (FLOPS) in gigaflops, average precision (AP), and frames per second (FPS). Notable models included in the comparison are YOLOv4-Ti utilizing DarkNet53, CenterNet with ResNet50, DETR with ResNet50, Def. DETR with ResNet18, YOLOS-Ti incorporating Deit-Ti, YOLOX-Ti using DarkNet53, and a proposed model implementing Swin-Ti [24]. The metrics reveal variations in model efficiency and accuracy, allowing for a detailed evaluation and informed selection of object detection architectures based on specific requirements.[Figure 9] is a graphical chart presentation to enhance data visualization, facilitating easier comprehension and interpretation of the tabular data from the Table 4.

The proposed model, which incorporates the Swin-Ti [24] backbone, demonstrates compelling performance across various metrics, enhancing its applicability in practical object detection scenarios. With a relatively moderate number of parameters (8.3 million) and low computational requirements (3.7 gigaflops), the model achieves a competitive average

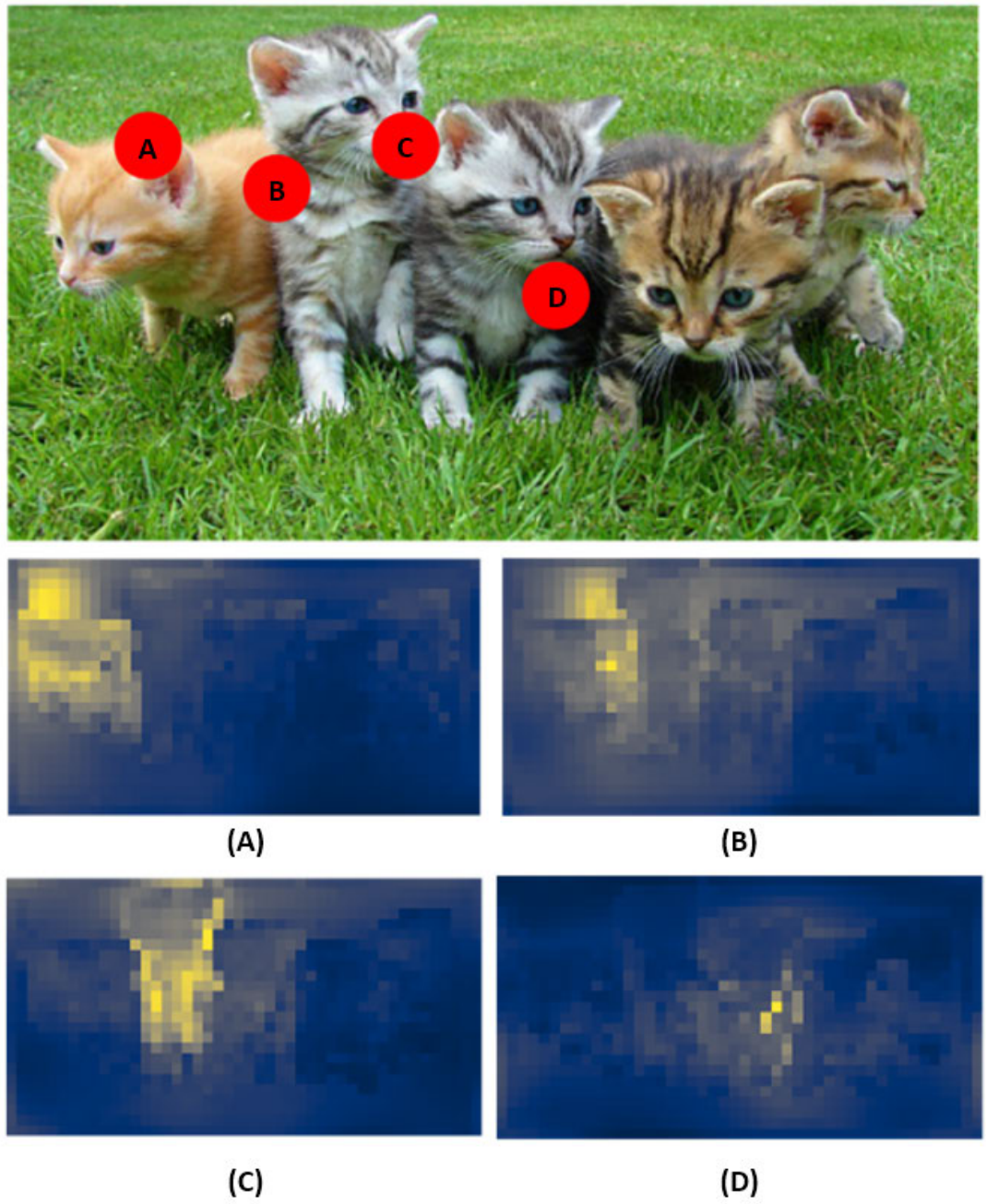

FIGURE 7. Self-attention map showing patch placement and attention scores, revealing patch-object relationships and composition insights.

precision (AP) of 30.2. This combination of efficiency and accuracy positions the proposed model as an attractive choice for real-world applications where computational resources may be constrained. Moreover, the achieved frames per second (FPS) rate of 89 suggests that the model can efficiently process input data in real-time, making it well-suited for dynamic environments where rapid and accurate object detection is crucial. Overall, the proposed model's favorable trade-off between performance and computational efficiency enhances its potential for widespread applicability in diverse scenarios, ranging from surveillance systems to autonomous vehicles.

[Figure 6] showcases the experimental outcomes of the proposed model on real images. The Attention Scores utilized in generating these outcomes are transformed into an Attention Map, allowing for a visual representation that confirms the model's performance. This map effectively dissects the attention associated with each predicted object's

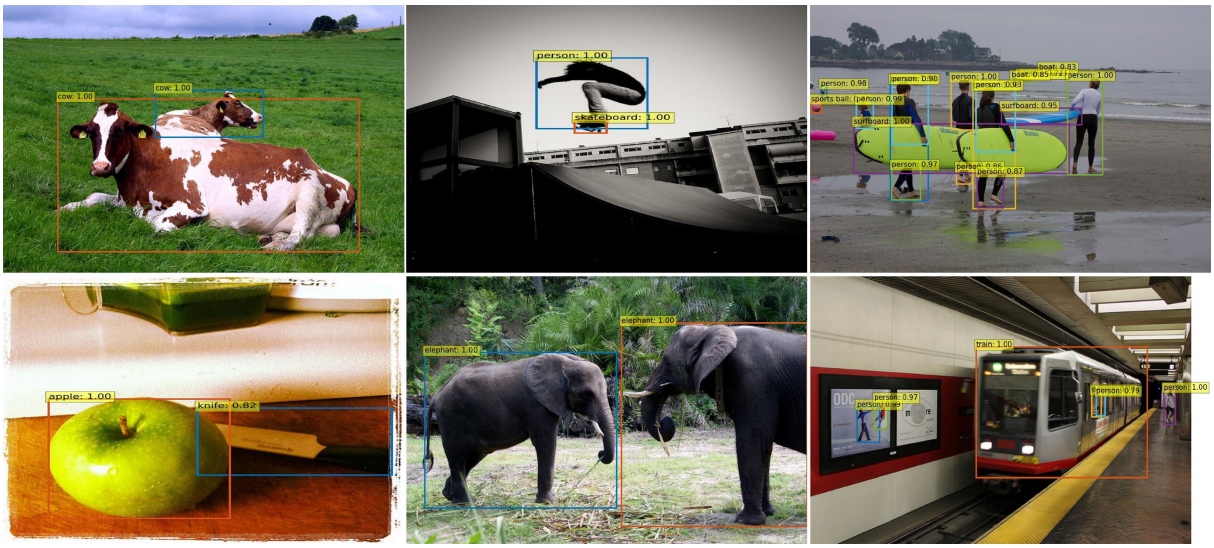

FIGURE 8. Detection results produced by the proposed method.

bounding box. While the figure underscores the efficacy of the proposed model in object detection, it is notable that the model's performance diminishes for objects at the boundary of the local window, attributable to the operating mechanism inherent in the Swin Transformer [24].

Meanwhile, [Figure 7] provides a visual representation of the Self-Attention Map, correlating with the model's patch placement. Circular markers denote attention scores at the centre of patches. High attention scores are illustrated in vivid hues, whereas low attention scores are depicted in darker shades. This visualization unveils the correspondence between each patch and its associated object, offering insights into the relationships that underlie the objects' composition.

Although the experimental results may appear comparatively underwhelming against traditional CNN-based object detectors, it's important to note that the model proposed in this paper intentionally adopts a pure transformer architecture. This design choice not only demonstrates the potential to enhance performance in object detection based on ViT [7], but also opens avenues for diverse applications of existing transformer models, shedding light on the trajectory of research for generating high-performance models. [Figure 8] shows some more results that our method generated on images from the MS COCO dataset.

V. DISCUSSION RELATED TO ATTENTION MECHANISM IN PROPOSED METHOD

In our proposed method, the integration of the Swin-Transformer encoder introduces crucial enhancements to object detection accuracy. The attention mechanism dynamically allocates resources to relevant image patches, fostering object-attention relationships pivotal for precise detection amidst varying backgrounds. This mechanism also enables contextual comprehension, aiding in disambiguating objects with similar appearances. Additionally, the patch merging process within the Swin-Transformer encoder enhances

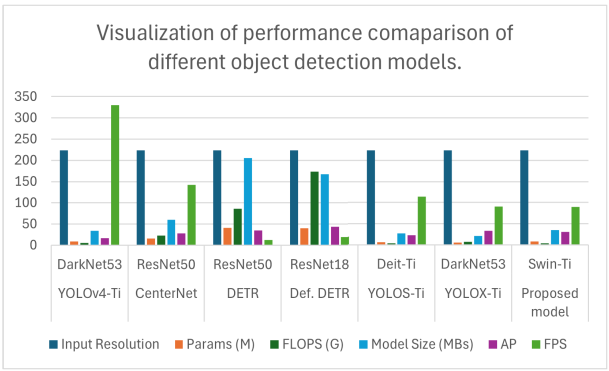

FIGURE 9. Visual illustration of tabular data in Table 4.

feature representation by aggregating information from smaller patches into larger ones. This facilitates the capture of global context and long-range dependencies within the input image, crucial for accurate localization. Hierarchical patch merging further enables the extraction of multi-scale features, essential for handling diverse datasets. By elucidating these technical aspects, we aim to provide readers with a deeper understanding of the mechanisms driving our approach and its potential for advancing object detection in real-world applications.

VI. CONCLUSION AND FUTURE DIRECTIONS

This paper presents a recent study that delves into enhancing object detection performance within the Vision Transformer [7] framework. Specifically, the investigation is centered around a novel first-stage detector, which seamlessly combines feature extraction and bounding box prediction using a transformer block. This deviates from the prevalent approach of employing a two-stage detector rooted in CNN architectures.

The proposed model leverages the Swin-Transformer [24] as its backbone, systematically amalgamating it with the existing Vision Transformer [2] structure. The process involves segmenting input images into patches of predetermined dimensions and calculating attention scores between each patch. This technique is augmented by setting regional windows to grasp inter-patch spatial dynamics. Repeatedly, the regional window is readjusted through patch merging, leading to superior performance compared to the conventional ViT [7] model. As evidenced by this outcome, this paper outlines the construction of a Swin-Transformer-based encoder [24] as the foundation of our proposed first-stage detector. By benchmarking against the existing ViT [7] model and YOLOS [11], a substantial performance improvement of 5.59% to 30.2% is observed in the object detection domain.

The implications of this achievement extend to steering the trajectory of transformer-based first-stage detectors. Such models exhibit potential as real-time object detectors that surpass the object detection speed of established transformer-based two-stage detectors. This, in turn, could serve as a bedrock for future research in the domain of real-time object detection models.

Looking ahead, our research avenue contemplates diversifying into various fields. Beyond elevating object detection rates, we aspire to explore methodologies applicable to diverse domains, including area detection. Additionally, our interests extend to investigating models tailored for object tracking, an endeavor parallel to object detection in image-based datasets. We will incorporate this model into our future research for video anomaly detection to localize objects that cause anomalies in given frames. This multifaceted exploration promises to contribute to the ongoing evolution of computer vision research.

ACKNOWLEDGMENT

(Tae Yang Kim and Asim Niaz are co-first authors.)

REFERENCES

[1] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[2] A. B. Amjoud and M. Amrouch, "Object detection using deep learning, CNNs and vision transformers: A review," *IEEE Access*, vol. 11, pp. 35479–35516, 2023.

[3] J. Liu, S. Liu, S. Xu, and C. Zhou, "Two-stage underwater object detection network using Swin transformer," *IEEE Access*, vol. 10, pp. 117235–117247, 2022.

[4] C. Yao, Y. Kong, L. Feng, B. Jin, and H. Si, "Contour-aware recurrent cross constraint network for salient object detection," *IEEE Access*, vol. 8, pp. 218739–218751, 2020.

[5] H. J. Kim, D. H. Lee, A. Niaz, C. Y. Kim, A. A. Memon, and K. N. Choi, "Multiple-clothing detection and fashion landmark estimation using a single-stage detector," *IEEE Access*, vol. 9, pp. 11694–11704, 2021.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[8] Y. Fang, X. Wang, R. Wu, and W. Liu, "What makes for hierarchical vision transformer?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12714–12720, Oct. 2023.

[9] A. Vaswani, A. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[10] 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*.

[11] Y. Fang, B. Liao, X. Wang, J. Fang, J. Qi, R. Wu, J. Niu, and W. Liu, "You only look at one sequence: Rethinking transformer in vision through object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 26183–26197.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[13] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, 2016, pp. 21–37.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[19] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[20] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," 2014, *arXiv:1409.2329*.

[21] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6836–6846.

[22] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[23] 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.* Piscataway, NJ, USA: IEEE Press, 2020, pp. 213–229.

[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[25] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, 1955.

[26] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 843–852.

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, Apr. 2015.

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis.* Zurich, Switzerland: Springer, 2014, pp. 740–755.

[29] J. Amin, M. Sharif, M. A. Anjum, H. U. Khan, M. S. A. Malik, and S. Kadry, "An integrated design for classification and localization of diabetic foot ulcer based on CNN and YOLOv2-DFU models," *IEEE Access*, vol. 8, pp. 228586–228597, 2020.

[30] M. Sharif, J. Amin, A. Siddiqa, H. U. Khan, M. S. Arshad Malik, M. A. Anjum, and S. Kadry, "Recognition of different types of leukocytes using YOLOv2 and optimized bag-of-features," *IEEE Access*, vol. 8, pp. 167448–167459, 2020.
[31] M. A. Khan, M. I. U. Lali, M. Sharif, K. Javed, K. Aurangzeb, S. I. Haider, A. S. Altamrah, and T. Akram, "An optimized method for segmentation and classification of apple diseases based on strong correlation and genetic algorithm based feature selection," *IEEE Access*, vol. 7, pp. 46261–46277, 2019.

TAE YANG KIM received the B.S. degree in computer engineering from Sahmyook University, South Korea, in 2021. He is currently pursuing the M.S. degree with the Department of Computer Science and Engineering, Chung-Ang University, Seoul, South Korea.

Since 2021, he has been a Research Assistant with the AI Vision Laboratory, Chung-Ang University, under the supervision of Dr. Choi. His current research interests include object detection, vision transformer, generative adversarial networks, and explainable AI.

ASIM NIAZ received the B.S. degree in electrical (computer) engineering from the COMSATS Institute of Information Technology, Islamabad, Pakistan, in 2016, and the M.S. degree in computer science and engineering from Chung-Ang University, Seoul, South Korea, in 2020.

He was a Research Assistant with the Visual Image Media Laboratory. Later, he visited INRIA Sophia Antipolis, France, followed by his research internship at Ericsson Canada, Montreal, Canada. He is currently a Research Assistant with the AI Vision Laboratory, Chung-Ang University. His current research interests include action recognition, video understanding, anomaly detection in images and videos, remote sensing, medical image analysis, and image segmentation.

JUNG SIK CHOI received the B.S. degree in software from Chung-Ang University, Seoul, South Korea, in 2023, where he is currently pursuing the M.S. degree with the Department of Computer Science and Engineering.

Since 2023, he has been a Research Assistant with the AI Vision Laboratory, Chung-Ang University, under the supervision of Dr. Choi. His current research interests include pose estimation and style transfer.

KWANG NAM CHOI received the B.S. and M.S. degrees from the Department of Computer Science, Chung-Ang University, Seoul, South Korea, in 1988 and 1990, respectively, and the Ph.D. degree in computer science from the University of York, U.K., in 2002.

He is currently a Professor with the School of Computer Science and Engineering, Chung-Ang University. His current research interests include motion tracking, object categorization, and 3D image recognition.