information

MDPI

Article

Object Detection of Road Assets Using Transformer-Based YOLOX with Feature Pyramid Decoder on Thai Highway Panorama

Teerapong Panboonyuen [1,*], Sittinun Thongbai [2], Weerachai Wongweeranimit [3], Phisan Santitamnont [3,4], Kittiwan Suphan [2] and Chaiyut Charoenphon [4,*]

1 Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Phayathai Rd, Pathumwan, Bangkok 10330, Thailand
2 InfraPlus Co., Ltd., Phayathai Rd, Pathumwan, Bangkok 10330, Thailand; sittinun2tb@gmail.com (S.T.); kittiwan.sati@gmail.com (K.S.)
3 Center of Excellence in Infrastructure Management, Chulalongkorn University, Phayathai Rd, Pathumwan, Bangkok 10330, Thailand; weerachai.w@chula.ac.th (W.W.); phisan.S@eng.chula.ac.th (P.S.)
4 Department of Survey Engineering, Faculty of Engineering, Chulalongkorn University, Phayathai Rd, Pathumwan, Bangkok 10330, Thailand
* Correspondence: teerapong.panboonyuen@gmail.com (T.P.); chaiyut.c@chula.ac.th (C.C.)

Abstract: Due to the various sizes of each object, such as kilometer stones, detection is still a challenge, and it directly impacts the accuracy of these object counts. Transformers have demonstrated impressive results in various natural language processing (NLP) and image processing tasks due to long-range modeling dependencies. This paper aims to propose an exceeding you only look once (YOLO) series with two contributions: (i) We propose to employ a pre-training objective to gain the original visual tokens based on the image patches on road asset images. By utilizing pre-training Vision Transformer (ViT) as a backbone, we immediately fine-tune the model weights on downstream tasks by joining task layers upon the pre-trained encoder. (ii) We apply Feature Pyramid Network (FPN) decoder designs to our deep learning network to learn the importance of different input features instead of simply summing up or concatenating, which may cause feature mismatch and performance degradation. Conclusively, our proposed method (Transformer-Based YOLOX with FPN) learns very general representations of objects. It significantly outperforms other state-of-the-art (SOTA) detectors, including YOLOv5S, YOLOv5M, and YOLOv5L. We boosted it to 61.5% AP on the Thailand highway corpus, surpassing the current best practice (YOLOv5L) by 2.56% AP for the test-dev data set.

Keywords: deep learning; YOLO; YOLOX; Vision Transformer; highway scenes

Citation: Panboonyuen, T.; Thongbai, S.; Wongweeranimit, W.; Santitamnont, P.; Suphan, K.; Charoenphon, C. Object Detection of Road Assets Using Transformer-Based YOLOX with Feature Pyramid Decoder on Thai Highway Panorama. *Information* **2022**, *13*, 5. https://doi.org/10.3390/info13010005

Academic Editor: Zoran H. Peric

Received: 15 November 2021
Accepted: 22 December 2021
Published: 25 December 2021

1. Introduction

Identifying road asset objects in Thailand highway monitoring image sequences is essential for intelligent traffic monitoring and administration of the highway. With the widespread use of traffic surveillance cameras, an extensive library of traffic video footage has been available for examination. A more distant road surface may usually be evaluated from an eye-observing angle. At this viewing angle, the vehicle's object size varies enormously, and the detection accuracy of a small item far away from the road is low. In the face of complicated camera scenarios, it's critical to address and implement the difficulties listed above successfully. We will focus on the challenges mentioned earlier in this post and provide a suitable solution. This study applies the object detection findings for multi-object tracking and asset object counting, including kilometer signs (marked as KM Sign) and kilometer stones (marked as KM Stone).

Among modern Convolutional Neural Networks (ConvNet/CNNs), there are many techniques, e.g., dynamic heads with attentions [1], dual attention [2], self-attention [3]

have gained increasing attention due to their capability. Still, all suffer from accuracy performance issues.

Nowadays, many works [4–8] extensively used architectures and applied in road object detection such as You Only Look Once version 3 (YOLOv3) [9], Mask R-CNN [10], BiseNet [11], YOLOv4 [12], YOLOv5 [13], and/or YOLOX [14]. They are created for image recognition, consist of stacked Conv blocks. Due to anxieties about the cost of computation, the purpose of kernel maps is decreased gradually. Furthermore, the encoder network can learn more semantic visual theories with a steadily increased receptive field. Consequently, it also inflates a primary restriction of studying long-range dependency knowledge, significant for image labeling in unconstrained scene images. It matures challenging due to still limited receptive fields. However, the previous architecture has not fully leveraged various feature maps from convolution or attention blocks conducive to image segmentation, and this has become a motivation for this work.

To defeat this limitation, as mentioned earlier, a completely new architecture known as Vision Transformer YOLO (ViT-YOLO) [15] with ViT [16] as major backbone has a tremendous capacity in long-range dependency acquisition and sequence-based picture modeling. It is a vision model-based that is built as closely as possible on the Transformer architecture, which was created for text-based jobs in the first area [17]. Furthermore, it has matured famous in several computer vision tasks, such as hyperspectral image classification [18,19], bounding-box detection [20,21], and image labeling [22,23]. ViT moves the window divider between successive levels of self-attention. The shifted windows provide links between the windows of the last layer, considerably increasing modeling capability. In terms of real-world precision, this method is also effective.

In this work, prompted by the preceding observation, we introduce transformer-based Feature Pyramid Network (FPN) [24] decoder designs. It learns the importance of different input features instead of simply summing up or concatenating, which may cause feature mismatch and performance degradation as demonstrated in Figure 4 This work points to further improving the SOTA on object detection in Thailand highway road images. For better performance, we inject the FPN style of decoder design into Transformer-based YOLOX reasoning. In this article, our main contributions are twofold:

- Utilizing a pre-training ViT to retrieve the virtual visual tokens based on the vision patches on images. We immediately fine-tune the model weights on downstream responsibilities by appropriating pre-training ViT as a backbone of YOLOX [14] by appending responsibility layers superimposing the pre-trained encoder.
- We apply the Feature Pyramid Network (FPN) [24] as decoder designs on our Transformer-Based YOLOX. It adds a different bottom-up path aggregation architecture. Notably, when the deep architecture is relatively shallow, and the feature map is more significant, the transformer layer is used prematurely to enforce regression boundaries which can lose some meaningful context information.

The experimental results on Thailand highway road demonstrate the effectiveness of the proposed scheme. The results proved that our Transformer-Based YOLOX with FPN decoder designs overcomes YOLOv5S, YOLOv5M, and YOLOv5L based architectures [25,26] in terms of AP, $AP50$, and $AP75$ score sequentially.

This article is organized as follows. Section 2 discusses related work, and our data set is detailed in Section 3. Next, Section 4 provides the detail of our methodology, and Section 5 presents our experimental results. Finally, conclusions are drawn in Section 6.

2. Related Work

Most relevant to our methodology is the Vision Transformer (ViT) [16] and their follow-ups [27–31]. Vision Transformer (ViT) [16] is a deep learning architecture that utilizes the mechanism of attention, and there are many works that follow-ups ViT [8,27–31]. Several works of ViT directly employ a Transformer model on non-overlapping medium-sized image patches for image classification. It reaches an exciting speed-performance trade-off on almost all computer vision tasks compared to previous deep learning networks.

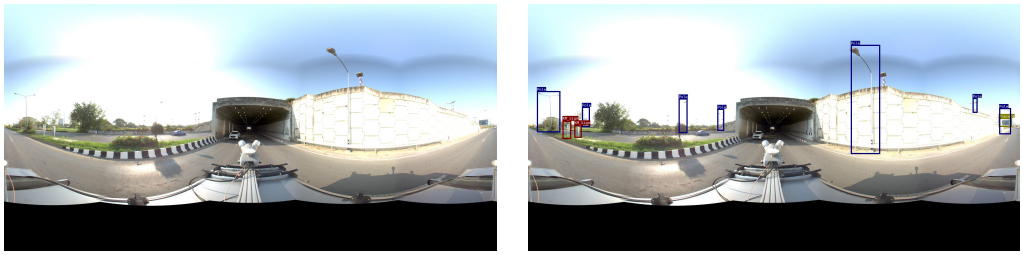
DeiT [32] introduces several training policies that allow it also to be efficient using the extra modest ImageNet-1K corpus. The effects of ViT on computer vision tasks are encouraging. Still, its model is inappropriate for profit as a general-purpose backbone network on dense image tasks due to its low-resolution kernel filter and the quadratic improvement in complexity with the image size. Some works utilize ViT models for the dense image tasks of image labeling and detection through transpose or upsampling layers yet comparatively lower precision. Unsurprisingly, we find ViT [33,34] models perform the best performance-accuracy trade-off among these methods on computer vision tasks, even though this work concentrates on general-purpose performance relatively than particularly on segmentation (labeling). It investigates a comparable line of studying to produce multi-resolution kernel features on ViT. Moreover, its complexity is still quadratic to the size of an image. At the same time, ours is linear and operates regionally, which has shown advantages in modeling the significant correlation in visual signals. Furthermore, it is efficient and effective, achieving SOTA performance, e.g., $MeanIoU$, $AveragePrecision(AP)$ on COCO object detection, and ADE20K image labeling.

3. Road Asset Data Set

Department of Highway (Thailand) has a highway road network of more than 52,000 km. To acquire information about road assets and roadside categories within the highway road network would take a lot of resource equipment and human resources. To solve data collection problems. A Mobile Mapping System (MMS.) and Artificial Intelligence (AI) were implemented. Its results from the current information, complete details, and highly efficient applications.

We have the process of collecting geospatial data from mobile vehicles (cars). Vehicles could be equipped with a range of sensors such as positioning (GNSS, GPS) and cameras.

The panoramic image type of The Ladybug-5 is a 360-degree spherical camera producing an image with a resolution of 8000 × 4000 pixels. It is required to calculate the position of objects on the pictures. Figure 1 depicts an example of the data set used in this study. A 360-degree spherical camera that can capture 8k30 or 4k60 footage is used to obtain the data set. The Ladybug5+ produces high-quality photographs with a 2 mm accuracy level at 10 m because of its proprietary calibration and better global shutter sensors. The Ladybug SDK offers a wide range of features that make it simple to capture, analyze, and spherical export material (shown as Figure 2). Furthermore, Figure 3 shows the sample size required for the detection task (number of images per class).



(**a**) (**b**)

Figure 1. The challenges in the Road Asset corpus. Sample of input image (**a**) and target image (**b**).

To survey and collect information on various types of highway assets. For use in the management of highway works in three main areas: (i) road asset management and maintenance, (ii) in terms of road safety to analyze the location, and (iii) in the planning of highway development projects. Therefore, a complete survey of the number of kilometer digits and the position is correct; accordingly, it is necessary to solve the duplication of construction work. The problem of calculating the amount is their construction work.

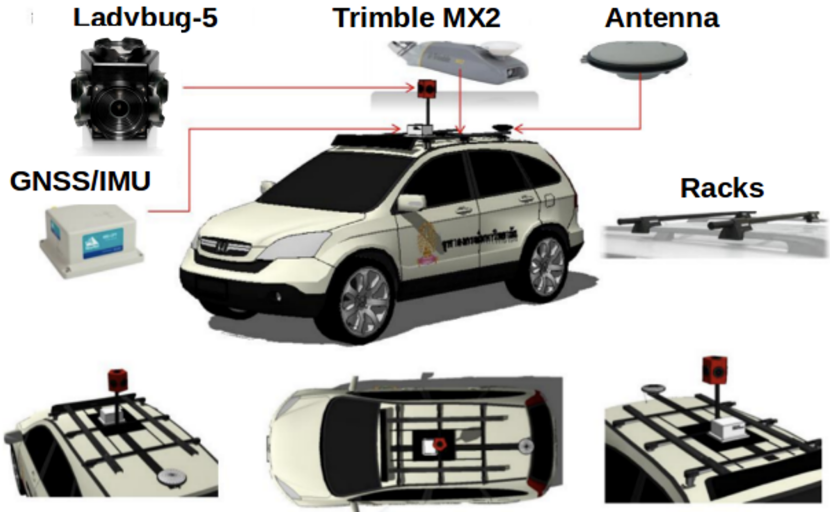
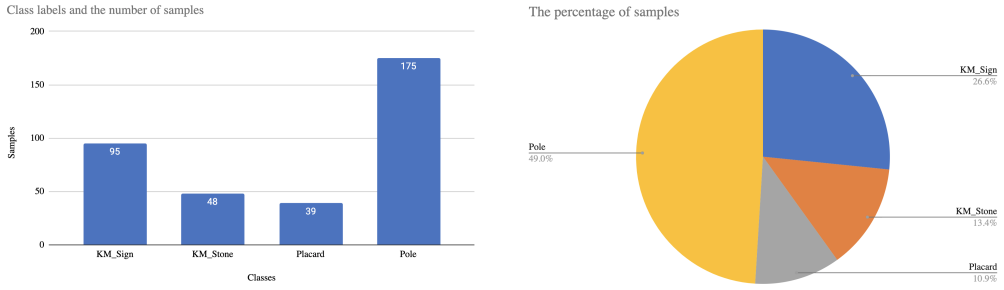


Figure 2. Our Ladybug5+ 360 degrees spherical camera.

As our deep learning environment setup, We use "TensorFlow Core v2.7.0 (TF)" [35] to create an end-to-end open-source platform for deep learning. The entire experiments were performed on servers with Intel® Xeon® Scalable 4210R (10 core, 2.4 GHz, 13.75 MB, 100 W), 256 GB of memory, and the NVIDIA RTX™ 1080Ti (11 GB) × 2 cards.



(**a**) (**b**)

Figure 3. The sample size of annotated class and its corresponding object class in the Road Asset corpus. Sample size for each class (**a**) and Percent of sample size for each class (**b**).

4. Proposed Method

4.1. Transformer Based YOLOX

Although currently, YOLOv5 [13] is already performing well, some recent work on object detection has triggered the development of this new YOLOX algorithm [14]. The most important focus points in object detection are anchor-free detectors, advanced label assignment strategies, and end-to-end detectors. These new focal points are still not integrated into the YOLO algorithm, and YOLOv4 and YOLOv5 are still anchor-based detectors and use hand-crafted assigning rules for training. It is a fundamental reason for the development of the YOLOX algorithm.

Our Transformer-based YOLOX follows a sequence-to-sequence vector with transformers from [36] and the corresponding output vector with input vector fabrication as in Natural Language Processing (NLP), the capacity of a machine application to understand mortal language. The most famous image classification network simply employs the Transformer Encoder to convert the multiple input tokens. However, the decoder component of the conventional transformer network is also employed for other purposes.

The regular ViT-YOLO model [15] and its conversion for computer vision tasks where the relations between a token (image patches) and each other tokens are calculated. The global figure leads to quadratic complexity concerning the number of image patches,

addressing it unsuitable for many image problems requiring an immense set of tokens for the softmax layer.

A pure transformer-based encoder learns feature representations furnished the $1D$ vector of embedding sequence E input. It means each ViT layer has a global receptive field, answering the insufficient receptive field problem of the existing encoder-decoder deep neural network already and for all. The ViT encoder consists of L_e layers of multilayer perceptron (MLP) and multi-head self-attention (MSA) modules.

This distinct behavior appears to be due to the inclusion of some inductive biases in CNNs, which these networks can use to comprehend the particularities of the analyzed image more rapidly, even if they end up restricting them and making it more difficult to grasp global relationships. Input visual distortions such as adversarial patches or permutations were also significantly more resistant to Vision Transformers. In reality, CNNs generate outstanding outcomes even when trained on data sets that are not as huge as Vision Transformers trade.

In case that a conventional encoder designed for image labeling would downsample a $2D$ image $x \in R^{HW3}$ into a grid of into a featuremap $x_f \in R^{\frac{H}{16} \times \frac{W}{16} \times C}$ we thus decide to set the transformer input sequence length L as $\frac{H}{16} \times \frac{W}{16} = \frac{W}{256}$. This means the output of the vector sequence of the ViT can be clearly reshaped to the point kernel map x_f.

To recover the $\frac{HW}{256}$-long vector sequence of our input, we divide an image $x \in R^{H \times W \times 3}$ into a grid of as $\frac{H}{16} \times \frac{W}{16}$ patches uniformly, several ViT modules with modified self-attention calculation (SwinTF modules) are adapted on these image patch tokens. The ViT module maintain the number of patches $\frac{H}{4} \times \frac{W}{4}$ and then make a series out of this grid. Each vectorized patch p is mapped into a latent C-dimensional embedding space using a linear projection function. $f : p \to e \in R^C$, for a patch x, we obtain a 1D series of vector embeddings. We get a unique embedding p_i for each position i to encode the patch spatial information, which is then added to e_i to generate the final sequence input $E = \{e_1 + p_1, e_2 + p_2, ..., e_L + p_L\}$. In this process, spatial data is kept notwithstanding the order-less attention type of transformers.

A classical transformer-based encoder accepts feature representations when given the 1D embedding sequence E as input. It means that each ViT layer has a global receptive field, resolving the problem of the standard deep learning encoder's restricted sensory area once and for all. The encoder of SwinTF consists of L_e vector of MLP and MSA modules (Figure 4). At each layer l, the input to self-attention is in a triplet of $(query, key, value)$ calculated from the input $Z^{l-1} \in R^{L \times C}$ as:

$$query = Z^{l-1}W_Q, key = Z^{l-1}W_K, value = Z^{l-1}W_V \tag{1}$$

where $W_Q/W_K/W_V \in R^{C \times d}$ are the learnable weights of three linear projection vectors and d is the dimension of $(query, key, value)$. Self-attention (SA) is then expressed as:

$$SA(Z^{l-1}) = Z^{l-1} + softmax(\frac{Z^{l-1}W_Q(ZW_K)^T}{\sqrt{d}})(Z^{l-1}W_V) \tag{2}$$

MSA is a reckoning with m self-supporting SA actions and projects their concatenated outputs: $MSA(Z^{l-1}) = [SA_1(Z_l - 1); SA_2(Z_l - 1); ...; SA_m(Z_l - 1)]W_O$, where $W_O \in R^{md} \times C$. d is typically set to C/m. The output of MSA is then transformed by an MLP module with residual skip as the output layer as:

$$Z^l = MSA(Z_{l-1}) + MLP(MSA(Z^{l-1})) \in R^{L \times C}. \tag{3}$$

Lastly, a normalized layer is employed before MLP and MSA modules which are omitted for clearness. We express $Z^1, Z^2, Z^3, ..., Z^{L_e}$ as the weights of transformer vectors.

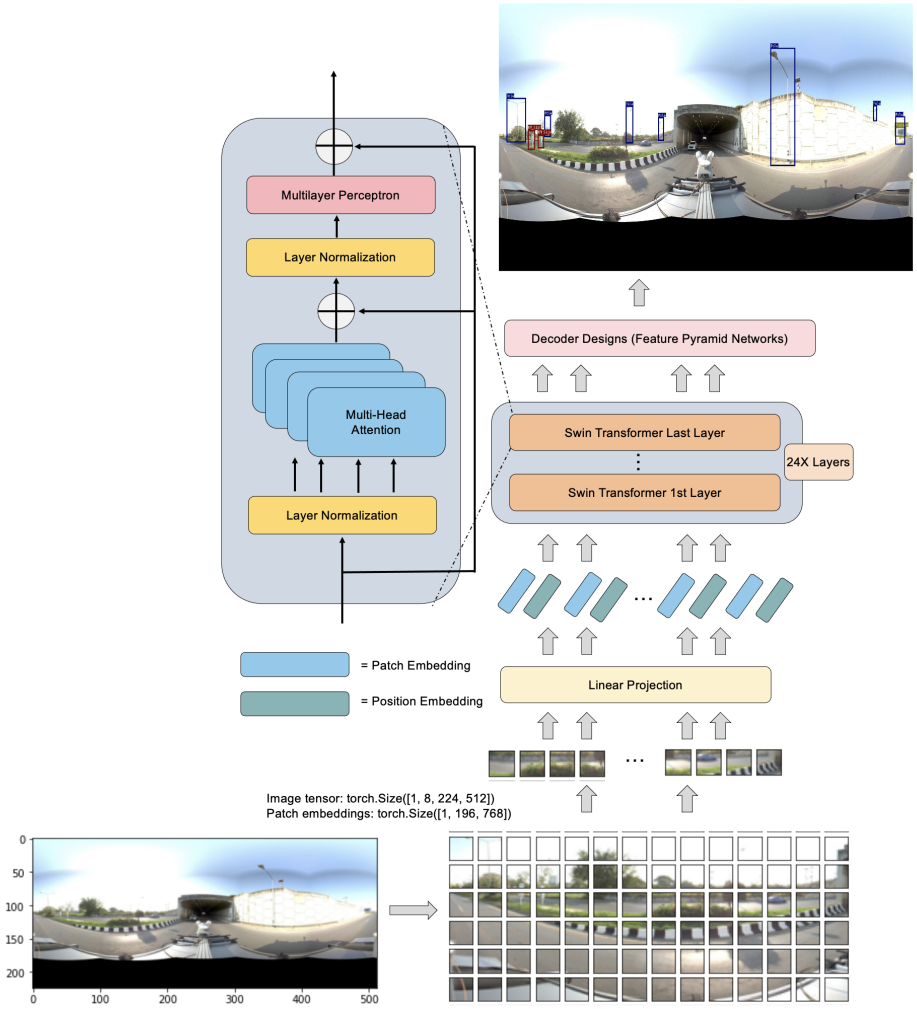


Figure 4. Overall architecture of our Transformer-Based YOLOX.

4.2. Feature Pyramid Network (FPN) Decoder Design

Objects from the road captured images vary a lot in sizes while the feature map from a single layer of the convolutional neural network has limited capacity of representation, so its crucial to effectively represent and process multi-scale features. FPN decoder designs as portrayed in Figure 5 are set up to achieve pixel-level labeling.

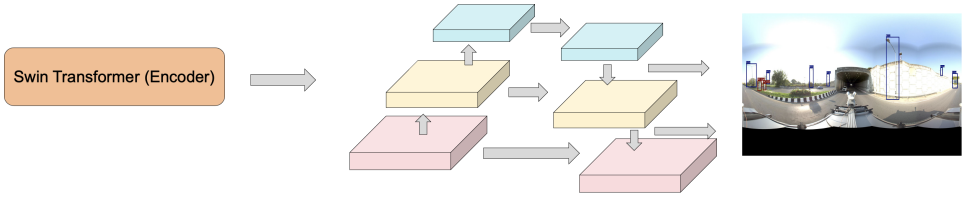


Figure 5. Our feature pyramid decoder design.

FPN [24] is a characteristic extractor created with accuracy and speed in mind for such a pyramid idea. It takes the place of detectors like Faster R-feature CNN's extractor [37]. Image recognition generates many feature map layers (multi-scale feature maps) with superior quality information than the traditional feature pyramid. It also utilizes specifically constructed transformers in a self-level, top-down, and bottom-up interactive pattern to change any feature pyramid into another feature pyramid of the same size but with richer contexts. It features a simple query, key, and value operation (Equation (1)) that is demonstrated to be important in choosing informative long-range interaction, which fits

our objective of non-local interaction at appropriate sizes. We depict the higher-level "idea" using the visual qualities of the lower-level "pixels" intuitively. Each level's altered feature maps (red, yellow, and blue) are resized to their matching map size and concatenated with the original map before being sent into the convolution layer, which resizes them to the accurate "thickness." Higher-resolution features are upsampled from higher-pyramid-level feature maps, which are spatially coarser but semantically more robust. The spatial resolution is upsampled by a factor of two, with the nearest neighbor being used for simplicity. Each lateral link combines feature maps from the bottom-up and top-down paths of the same spatial size. To minimize the channel dimensions, the feature maps from the bottom-up course are convolutional 11 times. In addition, element-wise addition is used to combine the feature maps from the bottom-up and top-down pathways. Finally, a 33 convolution is applied to each merged map to form the final feature map to reduce the aliasing impact of upsampling. This last collection of feature maps corresponds to the precise spatial dimensions. Because all layers of the pyramid, like a standard featured picture pyramid, employ joint classifiers/regressors, the feature dimension at output d is fixed at $d = 256$. As a result, the outputs of all further convolutional layers are 256-channel.

5. Experimental Results

Our proposed YOLOX with Vision Transformer and FPN method reaches the highest performance on Average Precision (AP) rating at 61.15% in the testing set. At the same time, the YOLOv5L is the best baseline, including with its fixed backbone as modification of CSP-v5, rated lower in terms of average AP at 58.94%, which this baseline is less average AP than the proposed method at 2.21% as shown in Table 1. In the AP 50 precision, the YOLOX with Transformer with FPN outperforms the highest precision rating at 69.34%. In comparison, the YOLOv5L model is significantly less precise in terms of AP75 than the proposed method at 10.19%, as shown in Table 1. Moving to the most significant AP at 75, the YOLOX with Transformer and FPN reaches the highest measurement at 55.23%, while the best combination of YOLOv5L reaches the AP 75 at 53.66%, which this baseline is less than the proposed method by 1.57%. There is a trade-off between performance as precision (AP) and the complexity of the deep convolution neural networks architecture. The YOLOX method achieves more the Average AP than the YOLOv5L, around 2%. Still, the proposed method takes a long time to train when comparing YOLOv5L because the trainable parameters of YOLOX are significantly higher than YOLOv5L, about 20M, as shown in Table 1. Table 2 displays numbers of training, validation, and testing sets on our road asset data set.

Table 1. Comparison of the average precision of different object classifiers on Road Asset test-corpus. We select all the models trained on 100 epochs for a fair comparison.

	Backbone	**Model**	***AP*(%)**	***Param***	***F*1**	***FPS***
Baseline	Modified CSP v5	YOLOv5-S	49.2	7.8M	63.01	22
	Modified CSP v5	YOLOv5-M	52.33	21.8M	66.22	23
	Modified CSP v5	YOLOv5-L	58.94	47.8M	67.88	26
	ResNext	Faster R-CNN	56.32	43.12M	65.12	15
	ResNext	CentreNet	58.11	12.2M	62.33	31
Proposed	ViT + FPN	YOLOX	61.15	87.3M	71.11	11

The YOLOX, with the combination of the modern image classification front-end, namely Vision Transformer with FPN, achieves the highest AP in the large-scale object classes such as *KMSign* and *Placard*. The release of YOLOv5 includes five different models sizes: YOLOv5s (smallest), YOLOv5m, YOLOv5l, YOLOv5x (largest).

Furthermore, it reaches the highest average precision (AP) at 51.32% and 57.63%, as shown in Table 3, in those mentioned classes. Our proposed outperforms when comparing the performance of YOLO-v5-L on these large-scale object classes in terms of AP. The YOLOX contains higher AP than YOLO-v5-L in categories such as *KMSign* and *Placard* by

6.47% and 2.82%, as shown in Table 3, respectively. Turning to the smaller object classes such as *KMStone* and *Pole*, our proposed method is still the winner of AP's *KMStone* and *Pole* rating at 61.22% 60.88% as shown in Table 3, respectively. Compared to the YOLO-v5-L method, the proposed method outperforms AP on the smaller object classes, both *KMStone* and *Pole* by 3.79% and 2.45%, as shown in Table 3, respectively.

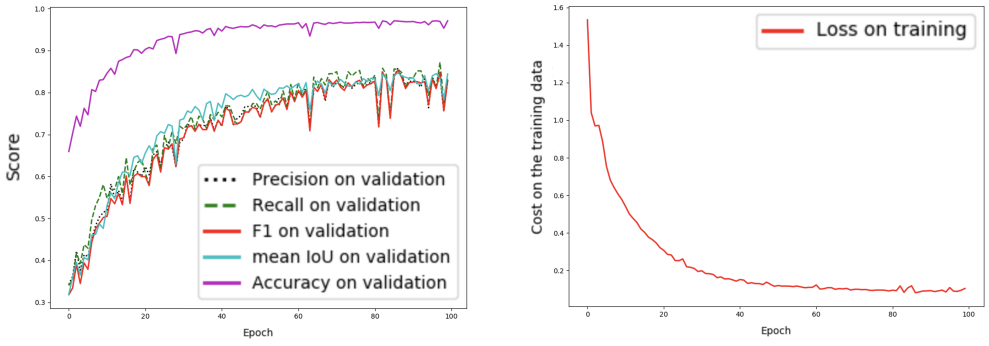
In the learning curve analysis, the cost function of YOLOX coupling with the Vision transformer with FPN, representing the line graph on the right-hand side, constantly learns into the local optima for all 100 epochs of the training set shown in Figure 6. In addition, the line graphs represent the performances of our proposed method in terms of Precision, Recall, F1, mean IoU, and Accuracy, which are evaluated on the validation set. On the left-hand side, the line graphs exhibit the upper tendency. Precision, Recall, F1, and mean IoU line graphs reach their highest performances, about 70% at the epoch of 100 on the validation set, while the line graph of Accuracy goes almost 100% at the end of 100 training epoch. Furthermore, these performances of line graphs drop at the epoch of 80, as shown in Figure 6.

Table 2. Numbers of training, validation, and testing sets.

Data Set	Total Images	Training Set	Validation Set	Testing Set
Road Asset Corpus	1300	800	300	200

Table 3. Comparison of the average precision between the results of four categories after subsequent operations on Road Asset test-dev corpus.

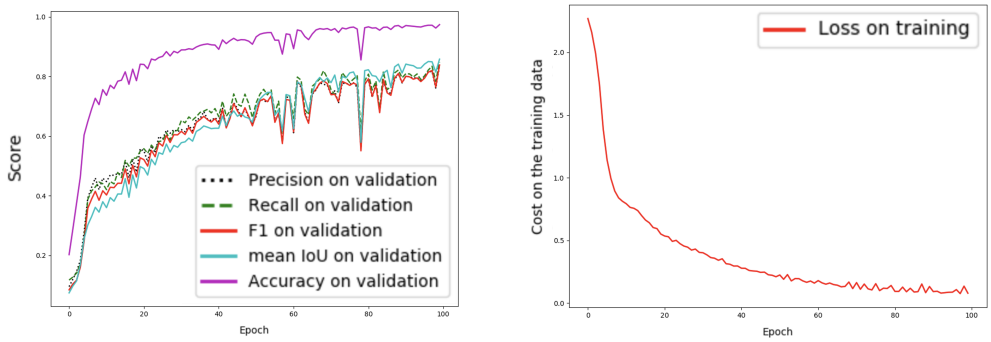
	Backbone	Model	*KMSign*	*KMStone*	*Pole*	*Placard*
Baseline	Modified CSP v5	YOLOv5-S	34.57	44.23	51.12	46.13
	Modified CSP v5	YOLOv5-M	40.12	53.12	54.73	47.23
	Modified CSP v5	YOLOv5-L	44.85	57.43	58.43	54.81
	ResNext	Faster R-CNN	46.75	54.22	57.34	56.22
	ResNext	CentreNet	37.625	54.35	56.66	53.21
Proposed	ViT + FPN	YOLOX	51.32	61.22	60.88	57.63



(**a**) (**b**)

Figure 6. Graph (learning curves) of the Road Asset corpus of the proposed approach, "YOLOX-based Vision Transformer with FPN"; x refers to epochs, and y refers to different measures (**a**) Plot of model loss (cross-entropy) on training and validation corpora; (**b**) performance plot on the validation corpus.

On the other hand, learning curves of both accuracy and loss graph of our best baseline (YOLOv5L) have shown via Figure 7. It shows their charts are smooth less than our proposed method.



(**a**) (**b**)

Figure 7. Graph (learning curves) of the Road Asset corpus of the baseline approach, "YOLOv5L"; x refers to epochs, and y refers to different measures (**a**) Plot of model loss (cross-entropy) on training and validation corpora; (**b**) performance plot on the validation corpus.

As shown in Figure 8, it provides qualitative object detection results of our YOLOX-Transformer with FPN model on an arbitrary image from Road Asset corpus. Lastly, we trained (finetuning) our YOLOX-Transformer with FPN model again via Pascal VOC data set [38] and prediction results shown in Figure 9.

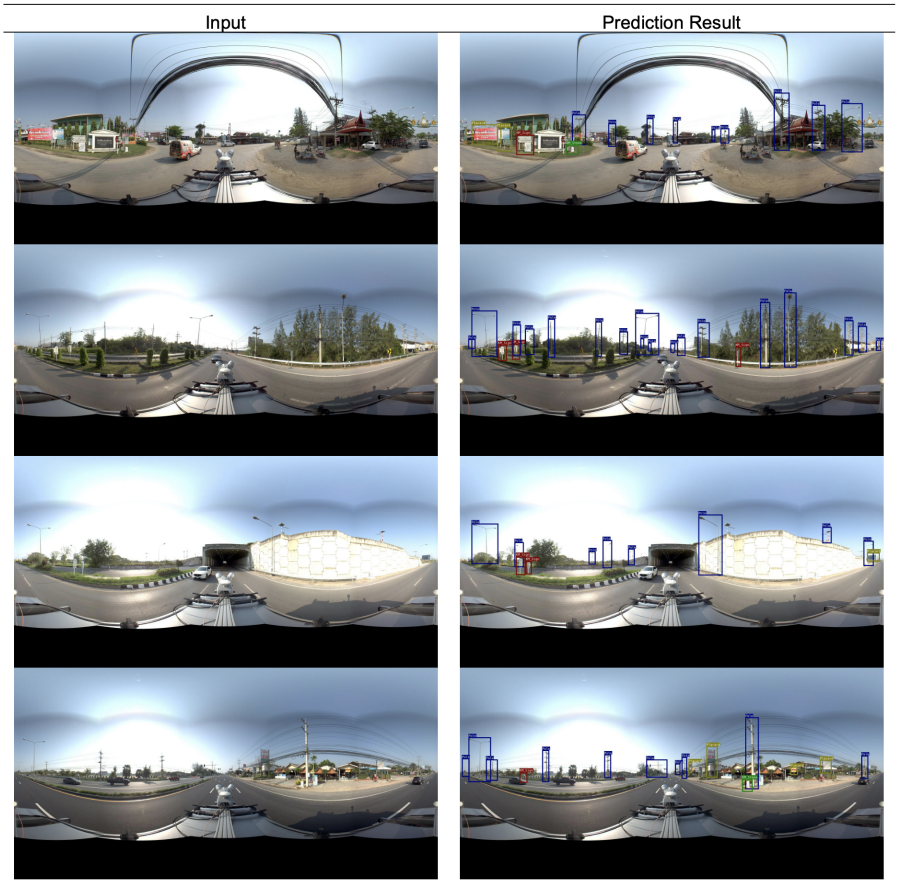


Figure 8. Visualized detection results of YOLOX-Transformer with FPN on an arbitrary image from Road Asset corpus.

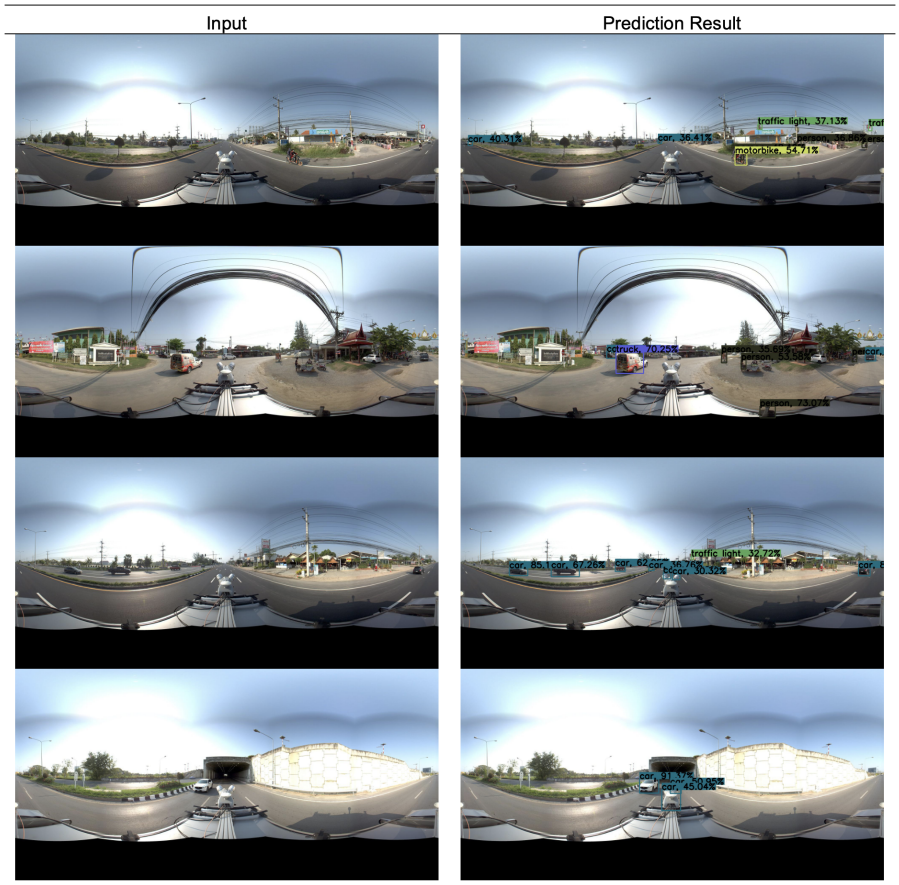


Figure 9. Visualized detection results of YOLOX-Transformer with FPN on Pascal VOC data set.

6. Conclusions

This paper proposes a novel Transformer-Based YOLOX with FPN, high-performance anchor-free YOLO for object detection. Our model can globally focus on dependencies between image feature patches and retain sufficient spatial information for object detection via multi-head self-attention. Furthermore, other effective techniques are adopted to achieve better accuracy and robustness. Furthermore, we apply FPN as learnable weights of decoder design to learn the importance of different input features instead of simply summing up or concatenating, which may cause feature mismatch and performance degradation. In particular, our Transformer-Based YOLOX with FPN achieves a new record of 61.15% box AP, 69.34% box AP50, and 55.23% box AP75 on Thailand highway test-dev, outperforming the prior SOTA model, including the following: YOLOv5S, YOLOv5M, and YOLOv5L.

Author Contributions: Conceptualization, T.P.; Formal analysis, T.P.; Investigation, T.P.; Methodology, T.P.; Project administration, T.P.; Resources, T.P.; Software, T.P.; Supervision, T.P.; Validation, T.P.; Visualization, T.P.; Writing–original draft, T.P.; Writing–review and editing, T.P.; Stand up and cheer, C.C., P.S., S.T., W.W., K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Center of Excellence in Infrastructure Management, Chulalongkorn University and the Ratchadapisek Somphot Fund for Postdoctoral Fellowship, Chulalongkorn University.

Acknowledgments: Teerapong Panboonyuen, also known as Kao Panboonyuen appreciates (thanks) and acknowledges the scholarship from Ratchadapisek Somphot Fund for Postdoctoral Fellowship, Chulalongkorn University, Thailand.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

FPN	Feature Pyramid Network
Param	Parameters
SwinTF	Swin Transformer
ViT	Vision Transformer

References

1. Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Yuan, L.; Zhang, L. Dynamic Head: Unifying Object Detection Heads with Attentions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 7373–7382.
2. Liu, S.; Zhang, L.; Lu, H.; He, Y. Center-Boundary Dual Attention for Oriented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5603914. [CrossRef]
3. Gu, Y.; Wang, L.; Wang, Z.; Liu, Y.; Cheng, M.M.; Lu, S.P. Pyramid constrained self-attention network for fast video salient object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10869–10876.
4. Haris, M.; Glowacz, A. Road Object Detection: A Comparative Study of Deep Learning-Based Algorithms. *Electronics* **2021**, *10*, 1932. [CrossRef]
5. Chen, G.; Chen, K.; Zhang, L.; Zhang, L.; Knoll, A. VCANet: Vanishing-Point-Guided Context-Aware Network for Small Road Object Detection. *Automot. Innov.* **2021**, *4*, 400–412. [CrossRef]
6. Wang, K.; Liu, M.; Ye, Z. An advanced YOLOv3 method for small-scale road object detection. *Appl. Soft Comput.* **2021**, *112*, 107846. [CrossRef]
7. Li, G.; Xie, H.; Yan, W.; Chang, Y.; Qu, X. Detection of road objects with small appearance in images for autonomous driving in various traffic situations using a deep learning based approach. *IEEE Access* **2020**, *8*, 211164–211172. [CrossRef]
8. Yang, J.; Li, C.; Zhang, P.; Dai, X.; Xiao, B.; Yuan, L.; Gao, J. Focal self-attention for local-global interactions in vision transformers. *arXiv* **2021**, arXiv:2107.00641.
9. Adarsh, P.; Rathi, P.; Kumar, M. YOLO v3-Tiny: Object Detection and Recognition using one stage improved model. In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 6–7 March 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 687–694.
10. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
11. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 325–341.
12. Wu, D.; Lv, S.; Jiang, M.; Song, H. Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. *Comput. Electron. Agric.* **2020**, *178*, 105742. [CrossRef]
13. Wu, W.; Liu, H.; Li, L.; Long, Y.; Wang, X.; Wang, Z.; Li, J.; Chang, Y. Application of local fully Convolutional Neural Network combined with YOLO v5 algorithm in small target detection of remote sensing image. *PLoS ONE* **2021**, *16*, e0259283. [CrossRef] [PubMed]
14. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
15. Zhang, Z.; Lu, X.; Cao, G.; Yang, Y.; Jiao, L.; Liu, F. ViT-YOLO: Transformer-Based YOLO for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 19–25 June 2021; pp. 2799–2808.
16. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2017; pp. 5998–6008.
18. He, X.; Chen, Y.; Lin, Z. Spatial-Spectral Transformer for Hyperspectral Image Classification. *Remote Sens.* **2021**, *13*, 498. [CrossRef]
19. Qing, Y.; Liu, W.; Feng, L.; Gao, W. Improved Transformer Net for Hyperspectral Image Classification. *Remote Sens.* **2021**, *13*, 2216. [CrossRef]
20. Sun, Z.; Cao, S.; Yang, Y.; Kitani, K.M. Rethinking transformer-based set prediction for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 19–25 June 2021; pp. 3611–3620.
21. Yang, F.; Zhai, Q.; Li, X.; Huang, R.; Luo, A.; Cheng, H.; Fan, D.P. Uncertainty-Guided Transformer Reasoning for Camouflaged Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 19–25 June 2021; pp. 4146–4155.
22. Wang, L.; Li, R.; Wang, D.; Duan, C.; Wang, T.; Meng, X. Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images. *Remote Sens.* **2021**, *13*, 3065. [CrossRef]
23. Jin, Y.; Han, D.; Ko, H. TrSeg: Transformer for semantic segmentation. *Pattern Recognit. Lett.* **2021**, *148*, 29–35. [CrossRef]
24. Kim, S.W.; Kook, H.K.; Sun, J.Y.; Kang, M.C.; Ko, S.J. Parallel feature pyramid network for object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 234–250.

25. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 19–25 June 2021; pp. 2778–2788.
26. Thuan, D. Evolution of Yolo Algorithm and Yolov5: The State-of-the-Art Object Detection Algorithm. 2021. Available online: https://www.theseus.fi/handle/10024/452552 (accessed on 12 November 2021).
27. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; Gao, W. Pre-trained image processing transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 12299–12310.
28. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision transformers for dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 19–25 June 2021; pp. 12179–12188.
29. Srinivas, A.; Lin, T.Y.; Parmar, N.; Shlens, J.; Abbeel, P.; Vaswani, A. Bottleneck transformers for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 16519–16529.
30. Kim, K.; Wu, B.; Dai, X.; Zhang, P.; Yan, Z.; Vajda, P.; Kim, S.J. Rethinking the Self-Attention in Vision Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 3071–3075.
31. Salvador, A.; Gundogdu, E.; Bazzani, L.; Donoser, M. Revamping Cross-Modal Recipe Retrieval with Hierarchical Transformers and Self-supervised Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 15475–15484.
32. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*; PMLR: London, UK, 2021; pp. 10347–10357.
33. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv* **2021**, arXiv:2103.14030.
34. Lin, A.; Chen, B.; Xu, J.; Zhang, Z.; Lu, G. DS-TransUNet: Dual Swin Transformer U-Net for Medical Image Segmentation. *arXiv* **2021**, arXiv:2106.06716.
35. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. Tensorflow: A system for large-scale machine learning. In Proceedings of the OSDI, Savannah GA USA, 2–4 November 2016; Volume 16, pp. 265–283.
36. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; 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, Nashville, TN, USA, 19–25 June 2021; pp. 6881–6890.
37. Cheng, B.; Wei, Y.; Shi, H.; Feris, R.; Xiong, J.; Huang, T. Revisiting rcnn: On awakening the classification power of faster rcnn. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 453–468.
38. Vicente, S.; Carreira, J.; Agapito, L.; Batista, J. Reconstructing pascal voc. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 41–48.