applied sciences

MDPI

Article

Multi-Resolution and Semantic-Aware Bidirectional Adapter for Multi-Scale Object Detection

Zekun Li [1,†], Jin Pan [1], Peidong He [2,3,†], Ziqi Zhang [4,†], Chunlu Zhao [1,*] and Bing Li [4]

[1] National Computer Network Emergency Response Technical Team/Coordination Center of China (CNCERT/CC), Beijing 100029, China; lzk@cert.org.cn
[2] Aerospace Information Research Institute, Chinese Academy of Sciences, No. 9 Dengzhuang South Road, Haidian District, Beijing 100094, China
[3] Department of Key Laboratory of Computational Optical Imaging Technology, Chinese Academy of Sciences, No. 9 Dengzhuang South Road, Haidian District, Beijing 100094, China
[4] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100094, China
* Correspondence: chunluzhao@cert.org.cn
† These authors contributed equally to this work.

Abstract: Scale variation presents a significant challenge in object detection. To address this, multi-level feature fusion techniques have been proposed, exemplified by methods such as the feature pyramid network (FPN) and its extensions. Nonetheless, the input features provided to these methods and the interaction among features across different levels are limited and inflexible. In order to fully leverage the features of multi-scale objects and amplify feature interaction and representation, we introduce a novel and efficient framework known as a multi-resolution and semantic-aware bidirectional adapter (MSBA). Specifically, MSBA comprises three successive components: multi-resolution cascaded fusion (MCF), a semantic-aware refinement transformer (SRT), and bidirectional fine-grained interaction (BFI). MCF adaptively extracts multi-level features to enable cascaded fusion. Subsequently, SRT enriches the long-range semantic information within high-level features. Following this, BFI facilitates ample fine-grained interaction via bidirectional guidance. Benefiting from the coarse-to-fine process, we can acquire robust multi-scale representations for a variety of objects. Each component can be individually integrated into different backbone architectures. Experimental results substantiate the superiority of our approach and validate the efficacy of each proposed module.

Keywords: object detection; scale variation; transformer; multi-level fusion

Citation: Li, Z.; Pan, J.; He, P.; Zhang, Z.; Zhao, C.; Li, B. Multi-Resolution and Semantic-Aware Bidirectional Adapter for Multi-Scale Object Detection. *Appl. Sci.* **2023**, *13*, 12639. https://doi.org/10.3390/app132312639

Academic Editor: Andrea Prati

Received: 15 September 2023
Revised: 9 November 2023
Accepted: 20 November 2023
Published: 24 November 2023

1. Introduction

Object detection, a crucial task in computer vision, entails the classification and localization of pertinent objects within an image. As convolutional neural networks (CNN) and vision transformers have experienced significant advancements, object detection methods have made considerable progress, contributing to the enhancement of recognition performance across diverse visual tasks. Numerous methods [1–4] have been proposed to enhance performance from various perspectives, demonstrating remarkable results on popular benchmarks like MS-COCO [5].

The object detection task involves predicting objects in natural and real-world scenes, which encompasses objects of varying scales. Nevertheless, scale variation poses a challenging dilemma that hampers the performance of detection methods. Several studies [6,7] have confirmed the sensitivity of CNNs to object scale and image resolution. Moreover, following a series of pooling and convolution operations on the input image, the features lose a noticeable amount of information, particularly pertaining to the fine details of microscopic objects. Furthermore, there is an imbalance of information across different levels. High-level features encompass semantic information, albeit lacking in spatial details,

while low-level features preserve detailed information but grapple with capturing semantic context. The aforementioned issues have emerged as bottlenecks for contemporary detection algorithms.

The implementation of multi-level feature integration serves as an effective strategy to mitigate these issues. For instance, FPN [8] employed a top-down feature integration method to combine features at different scales. Nonetheless, the input features of FPN are directly extracted from the backbone network. They may have already lost original information during the inference process. Additionally, some studies [9] have validated the significance of semantic information in the high-level features. However, there exists insufficient exploration and utilization of the high-level features in FPN. Additionally, the merging approach incorporates a rigid fusion of two features, neglecting the variability in features across different levels. Relying solely on high-level features to direct low-level features leads to an absence of low-level spatial information in the high-level features. Moreover, the direct fusion process from top to bottom may dilute the semantic information within the high-level features. This suggests an insufficient interaction among multi-level features. As depicted in Figure 1, the baseline network (FPN) in the left column struggles to effectively address multi-scale object challenges, resulting in numerous false positive cases. This is primarily due to the underutilization of features and the absence of precise object representations.

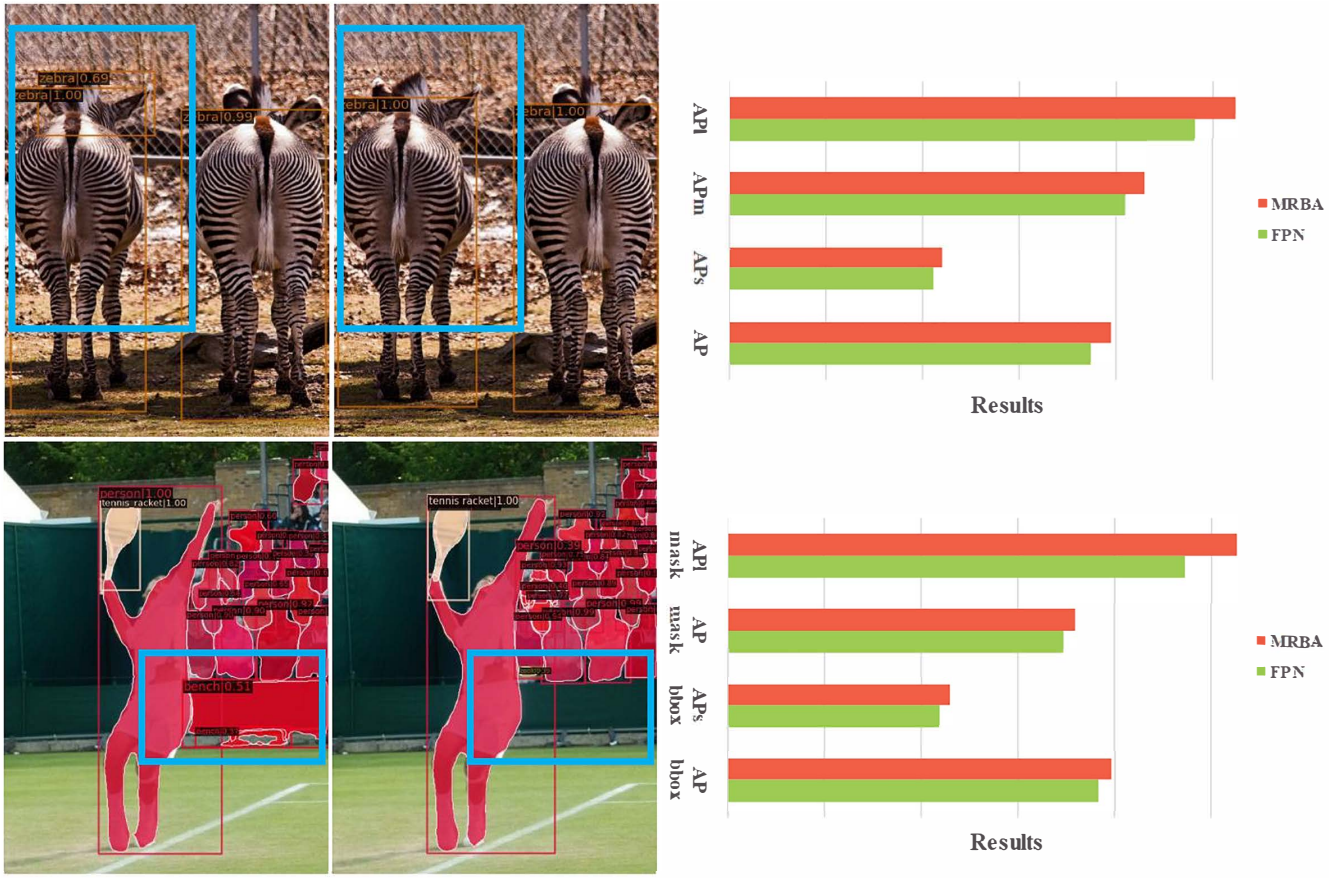

Figure 1. Visual Comparison of Results. The top row displays the detection outcomes of Faster R-CNN using FPN (*left*) and MSBA (*right*). The MSBA-based results exhibit a significant reduction in false positives and a qualitative performance enhancement. AP_S, AP_M, and AP_L denote the AP of small, medium, and large objects. In the bottom row, a similar trend is observed for Mask R-CNN, where our approach (*right*) consistently outperforms the baseline (*left*). *AP bbox* and AP_S *bbox* pertain to detection performance and bbox AP for small objects. *AP mask* and AP_L *mask* correspond to instance segmentation performance and mask AP for large objects.

To mitigate the constraints of current approaches, we introduce a novel and potent framework, termed a multi-resolution and semantic-aware bidirectional adapter for multi-scale object detection (MSBA). More precisely, this framework comprises three sequential components: multi-resolution cascaded fusion (MCF), a semantic-aware refinement transformer (SRT), and bidirectional fine-grained interaction (BFI). Respectively, these three components target the input, enhancement, and interaction aspects of the feature integration process through a coarse-to-fine strategy. The MCF component receives inputs in the form of multi-stage features and multi-resolution images from the backbone. It then adaptively extracts suitable multi-level features tailored to distinct object instances through a cascaded fusion strategy involving multiple receptive fields. Additionally, SRT is introduced to enhance the multi-scale semantic representation by refining both detailed and global semantic information while minimizing computational costs. SRT is designed with a semantic association strategy and employs multi-branch attention to effectively integrate semantic information across diverse scales. Moreover, to achieve a versatile and effective feature interaction, we introduce BFI, a mechanism for establishing a bidirectional flow of information. The bottom-up interaction is intended to furnish spatial guidance transitioning from low-level to high-level layers, fostering interaction across multiple levels. By leveraging intricate spatial information from low-level layers, high-level layers can effectively identify salient regions and provide enhanced semantic information with greater accuracy. Conversely, the top-down interaction is employed to establish semantic enhancement from high-level layers to low-level layers. Building upon the copious semantic information in the high-level layers, low-level layers can exhibit a comprehensive comprehension of object instances. In conclusion, the introduced coarse-to-fine process allows for the attainment of a more potent representation of objects across multiple scales.

Thorough experiments are carried out to validate the efficacy of the proposed approach. The introduced MSBA serves as a plug-and-play framework that seamlessly integrates with diverse backbones and detectors. On the MS COCO dataset, our method consistently outperforms state-of-the-art methods, achieving superior performance across different backbones and detectors, without any additional bells and whistles. As depicted in Figure 1, our detection results, presented in the second column, demonstrate superiority in accurately detecting multi-scale objects. In summary, this study offers the following key contributions:

- To mitigate the challenge of scale variation, we introduce a novel multi-resolution and semantic-aware bidirectional adapter for multi-scale object detection, referred to as MSBA. It alleviates the scale-variant issue by addressing the input, refinement, and interaction facets of feature integration.
- Our proposition, MSBA, is composed of multi-resolution cascaded fusion (MCF), a semantic-aware refinement transformer (SRT), and bidirectional fine-grained interaction (BFI). SRT is dedicated to refining the multi-scale semantic representation, while BFI is employed to foster ample interaction across various levels. Importantly, all these modules are pluggable.
- The proposed method is rigorously evaluated on the widely used MS-COCO dataset, demonstrating its superiority over state-of-the-art approaches. Thorough ablation experiments are conducted to confirm the efficacy of the proposed modules within the MSBA framework.

2. Related Work

Object detection is a fundamental task in computer vision that finds wide application in other visual fields, including remote sensing [10,11] and self-driving [12,13] technologies. It involves identifying and classifying objects of interest within an image. Object detection has made remarkable advancements in terms of accuracy and speed, thanks to convolution-based and transformer-based algorithms.

2.1. Object Detection

In the field of object detection, when it comes to the convolution-based network architectures, most detectors can be organized into two types: two-stage detectors [14–17] and one-stage detectors [18–21]. Two-stage detectors can achieve better performance with longer computation time, and the one-stage detectors show superiority in speed with inferior accuracy. In terms of the representation of the object, there can be divided into anchor-based and anchor-free detectors. Anchor-based [16,20] methods employ a multitude of anchor boxes to classify and locate objects, while anchor-free methods [22–24] utilize key points (e.g., center or corner points) for detection rather than relying on intricate manual design and hyperparameter settings. ATSS [25] has been proposed as a flexible label assignment method to narrow the discrepancy between anchor-free and anchor-based approaches. Recently, transformer-based methods [4,26–30] have made significant advancements. DETR [4] is the first end-to-end detector based on transformer blocks that achieves comparable performance at a high computation cost. Subsequently, deformable DETR [26] is proposed to enhance performance while mitigating computation costs through the use of deformable attention strategies. Additionally, Sparse R-CNN [27] employs sparse boxes to accomplish multi-stage refinement using a combination of self-attention modules and iterative structures. MCCL [31] is introduced to apply a novel training-time technique for reducing calibration errors. NEAL [32] is dedicated to training an attentive CNN model without the introduction of additional network structures. PROB [33] presents a novel probabilistic framework for objectness estimation within the context of open-world object detection.

2.2. Approaches for Scale Variation

Scale variation in object instances poses a significant challenge in object detection, hindering the improvement of detection accuracy. Singh et al. introduces SNIP [6] and SNIPER [34] as solutions to address this issue. The proposed method acknowledges the sensitivity of CNN to scales and advocates for detecting objects within a specified scale range. Consequently, a scale normalization training scheme is devised to facilitate the detection of objects at varying scales. These concepts have been widely adopted to acquire multi-scale information. However, SNIP exhibits high complexity, limiting its suitability for certain practical applications. FPN [8] introduces a novel feature pyramid architecture to solve the problem of scale variation by merging adjacent layers from top to bottom. It has achieved significant advancements and serves as a fundamental structure in many detectors. However, there is still room for performance improvement. PANet [35] is subsequently proposed to enhance FPN by introducing a new bottom-up structure that shortens information propagation. Moreover, FPG [36] stacks multi-pathway pyramids to enrich feature representations. DSIC [37] utilizes a gating mechanism to dynamically control the flow of data, enabling the automatic selection of different connection styles based on input samples. Furthermore, to address scale variation, PML [38] designs an enhanced loss function by modeling the likelihood function. HRViT [39] combines high-resolution multi-branch architectures with vision transformers (ViTs). MViTv2 [40] includes residual pooling connections and decomposed relative positional embeddings. In contrast to the aforementioned methods, our approach highlights the roles of different layers and maximizes information exchange between high-level and low-level layers to enhance feature representations. In contrast to the aforementioned methods, our approach incorporates both multi-stage features and multi-resolution images as suitable inputs, employing a cascaded fusion strategy. Furthermore, the proposed MSBA highlights the roles of different layers and maximizes information exchange between high-level and low-level layers to enhance feature representations.

2.3. Vision Transformer

The application of transformers in diverse visual tasks has made significant advancements. ViT [41] employs a standard transformer backbone for image classification, but this

approach incurs significant computational overhead. Subsequently, a series of studies are conducted to enhance ViT. For instance, T2T-ViT [42] divides the image into overlapping patches as tokens, enhancing token interactions. TNT [43] investigates both patch-level and pixel-level representations using nested transformers. Additionally, CPVT [44] introduces implicit conditional position encodings that depend on the local context of the input token. Notably, the Swin transformer [45] introduces a hierarchical approach that incorporates multi-level features and window-based attention. Moreover, the application of the transformer to other vision tasks has achieved remarkable progress, such as video captioning [46,47], vision-language navigation [48,49], and visual voice cloning [50,51]. These excellent works have witnessed the milestone success of the vision transformer. Furthermore, numerous endeavors [52–54] have been dedicated to leveraging the strengths of both the CNN and transformer, resulting in improved performance while reducing computational overhead. However, the majority of the aforementioned studies concentrate on enhancing the attention mechanism within individual feature states, disregarding the variations among features across different receptive fields. Conversely, our transformer-based approach can amalgamate global and local semantic information within high-level features, due to the proposed effective attention mechanism. Furthermore, our proposed method places greater emphasis on exploring interactions among diverse receptive fields and accentuating the reusability of features to enhance their representational capacity.

3. The Proposed Method

3.1. Foundation

The overview of the proposed MSBA is illustrated in Figure 2. As depicted in Figure 2a, MCF comprises two feature information streams. The $\{C_2', C_3', C_4', C_5'\}$ indicate the features derived from the multi-resolution input image, processed through multiple convolutions to capture sufficient coarse-grained information. $\{C_2, C_3, C_4, C_5\}$ represent features from distinct stages of the single-resolution image undergone by the backbone network. In Figure 2b, to ensure consistent notation within the same module, we employ $\{M_2', M_3', M_4', M_5'\}$ in BFI to denote features derived from MCF's output. SRT concentrates on enhancing the multi-scale semantic representation in the high-level feature, specifically targeting C_5. Besides, Additionally, BFI encompasses pixel-level filter interaction (PLI) and channel-wise prompt interaction (CWI). The output of PLI is denoted as $\{M_2, M_3, M_4, M_5\}$, where M_2 remains unchanged (M_2') without any further operations. Similarly, $\{P_2', P_3', P_4', P_5'\}$ mirrors $\{M_2, M_3, M_4, M_5\}$ and represents features resulting from PLI's output. Additionally, $\{P_2, P_3, P_4, P_5\}$ signify features enriched with meticulous semantic prompt information, primed for predictions.

The matching gate functions as a controller, aiming to mitigate inconsistencies and redundancy arising from rigorous interaction between two features. It dynamically modulates the fusion process in response to the present input. In detail, when provided with input features $X, Y \in \mathbb{R}^{c\times h\times w}$ as input, the matching gate $\mathscr{G}(\cdot)$ can be described as:

$$\mathscr{G}(X, Y) = [F_{mul}(\alpha^{fine}, X) + F_{mul}(1 - \alpha^{fine}, Y)], \tag{1}$$

in which $\alpha^{fine} \in \mathbb{R}^{c\times 1\times 1}$ represents the control matrix of X and F_{mul} means the Hadamard product. α^{fine} can be obtained from the switch ($\mathcal{S}$) in the matching gate as:

$$\alpha^{fine} = \mathcal{S}(X), \tag{2}$$

$$\mathcal{S}(X) = \sigma[\mathscr{O}(\cdot), X], \tag{3}$$

where $\mathscr{O}(\cdot)$ represents the operations such as 3×3 convolution and pooling. $\sigma(\cdot)$ signifies a nonlinear activation function, executed as *Tanh* within our method. The matching gate adeptly fosters complementarity between the two features.

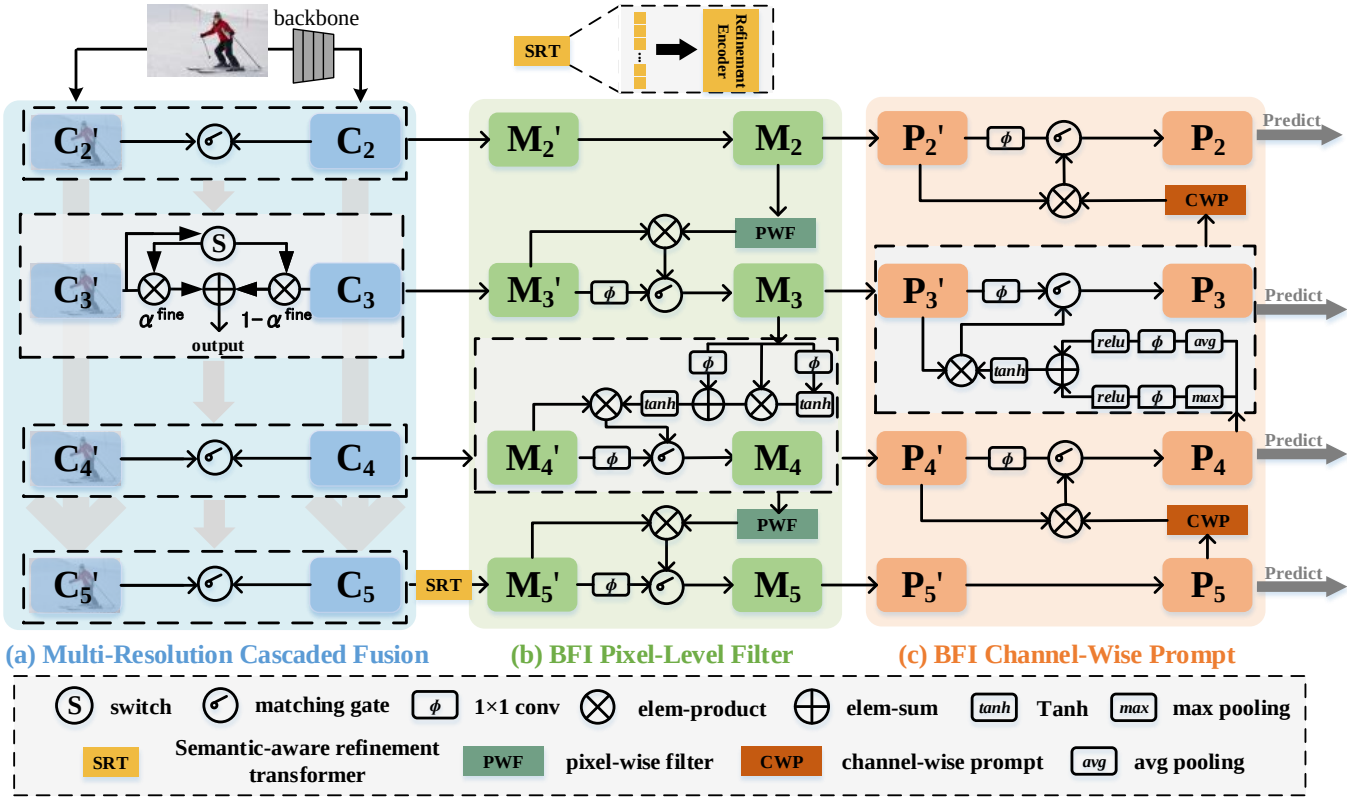

Figure 2. The overall architecture of MSBA. There are three components: multi-resolution cascaded fusion (MCF), semantic-aware refinement transformer (SRT) and bidirectional fine-grained interaction (BFI). MCF performs an adaptive fusion of multi-receptive-field and multi-resolution features, providing ample multi-scale information. Subsequently, SRT refines the features by amplifying long-range semantic information. Moreover, BFI ensures robust interaction by establishing two opposing directions of guidance for features containing fine-grained information. The pixel-level filter establishes a bottom-up pathway to convey spatial information from high-resolution levels. Concurrently, the channel-wise prompt guides low-level semantic information via the top-down structure.

3.2. Multi-Resolution Cascaded Fusion

FPN employs a single-resolution image as its input to create a feature pyramid. It can partially mitigate the challenge of scale variation. However, this approach is limited since a single-resolution image can only offer a restricted amount of object information within a specific scale. Using high-resolution images as input can be advantageous for detecting small objects, yet it might lead to relatively lower performance in detecting larger objects. Conversely, utilizing low-resolution images as input may lead to subpar performance in detecting small objects. Consequently, employing a single-resolution image as input might not suffice for effectively detecting objects across various scales.

Hence, the inclusion of a multi-scale image input is crucial for detectors to gather a broader spectrum of object information across different resolutions. This observation motivates our introduction of the multi-resolution cascaded fusion, which integrates multi-resolution data into the network architecture, as illustrated in Figure 2a. Initially, the input image undergoes both backbone processing and direct downsampling to align with the size of $C_i = \{C_2, C_3, C_4, C_5\}$ from the backbone as $Cds'_i = \{Cds'_2, Cds'_3, Cds'_4, Cds'_5\}$. Following this, the downsampled multi-resolution images undergo a sequence of convolution, batch normalization, and activation operations, culminating in the creation of corresponding features imbued with both coarse-grained spatial details and semantic insights. Furthermore, we employ a matching gate to adaptively manage the fusion process between the generated

multi-resolution features and the multi-stage features derived from the backbone. This procedure can be described as:

$$C'_i = \sum_{\Psi_i \in CBR} \Psi_i(Cds'_i). \tag{4}$$

Here, Cds'_i refers to the input image that has been downsampled to align with the suitable spatial dimensions of C_i, with i representing the feature level index from the backbone. $\Psi_i(\cdot)$ represents a sequence of operations, including a 3×3 Conv, BN, and ReLU to produce semantic features. Subsequently, we leverage C'_i to merge with the corresponding C_i using a matching gate, thereby generating a feature that is more effective. Additionally, we formulate a multi-receptive-field cascaded fusion strategy to extract multi-scale spatial information from the lower-level features. The entire procedure can be expressed as follows:

$$M'_i = \mathscr{G}(C'_i, C_i) + R_i(\mathscr{G}(C'_{i-1}, C_{i-1})) \quad i = (3, 4, 5), \tag{5}$$

where R'_i signifies the convolution operator applied with different dilation rates. M'_i corresponds to the input for the subsequent stage, enriched with ample coarse-grained and multi-scale spatial information. Notably, M'_2 is derived from the matching gate without the incorporation of dilated convolution.

Generally, our multi-resolution cascaded fusion supplies diverse resolution information. The proposed MCF is advantageous for object instances of varying scales. Additionally, we employ a matching gate as a controller to dynamically regulate the interaction process between multi-resolution images and the multi-stage features of the backbone. This adaptively controlled process aids in avoiding the inclusion of unnecessary information. Furthermore, the proposed multi-receptive-field cascaded fusion strategy contributes to the extraction of ample multi-scale spatial information for the high-level features. The resulting features consequently achieve a more comprehensive representation of different scales.

3.3. Semantic-Aware Refinement Transformer

Based on earlier investigations [9,55], it is evident that the semantic message contained in the high-level features significantly contributes to mitigating scale variations. However, in conventional approaches, there is a lack of distinction between different levels. Common methods merely employ high-level features to provide semantic information in their original states. Moreover, the transformer is designed to capture long-range semantic messages due to its self-attention mechanism. Nevertheless, directly applying the transformer to high-level features may disregard the variations in features across diverse representation situations. Thus, we propose the SRT transformer encoder to enhance the comprehensive semantic representation of high-level features across different feature states. This enhancement facilitates the acquisition of multi-scale semantic global information by high-level features.

As illustrated in Figure 3, we employ SRT on C'_5 to augment the semantic information. The entire process of SRT can be elucidated as follows:

$$\hat{M}'_5 = LN\{Attn_{\mathbb{SRT}}(PE(C'_5)) + PE(C'_5)\}, \tag{6}$$

$$M'_5 = LN\{(FFN(\hat{M}'_5) + \hat{M}'_5)\}, \tag{7}$$

where LN denotes the layer normailzation operation. PE introduces the position embedding for the feature and the FFN serves to enhance the non-linearity of these features. $Attn_{\mathbb{SRT}}$ signifies the novel SRT attention mechanism, enabling the query of the original feature to probe long-range semantic relationships across various feature states. Furthermore, the

sufficient semantic information can be integrated through the SRT attention mechanism effectively. The process can be delineated as:

$$Attn_{\mathbb{SRT}} = Concat[\{Attn_n(q_1, k_i, v_i)\}_{n=1}^{h}] \quad i = (1, 2, 3). \tag{8}$$

The term q_1 represents the query extracted from the original feature. The keys, namely k_2, k_3, along with the values v_2, v_3, signify the keys and values obtained through processing the corresponding features using average and max pooling operations. The processed features can achieve more expressive with tiny spatial size. The h denotes the number of attention heads. Following this, q_1 engages in interactions with the other keys to amplify the semantic representation of the high-level feature under various representation states. The mechanism $Attn$ is employed to calculate token-wise correlations among the features. Details can be formulated as follows:

$$Attention(q, k, v) = Softmax(\frac{qk^T}{\sqrt{d_k}})v, \tag{9}$$

where q, k, and v represent the query, key, and value, separately. d_k denotes the feature channels. Our proposed approach employs the initial query to compute correlations with other keys sourced from diverse sections of the feature. This process enables the sufficient extraction of semantic information from the high-level feature.

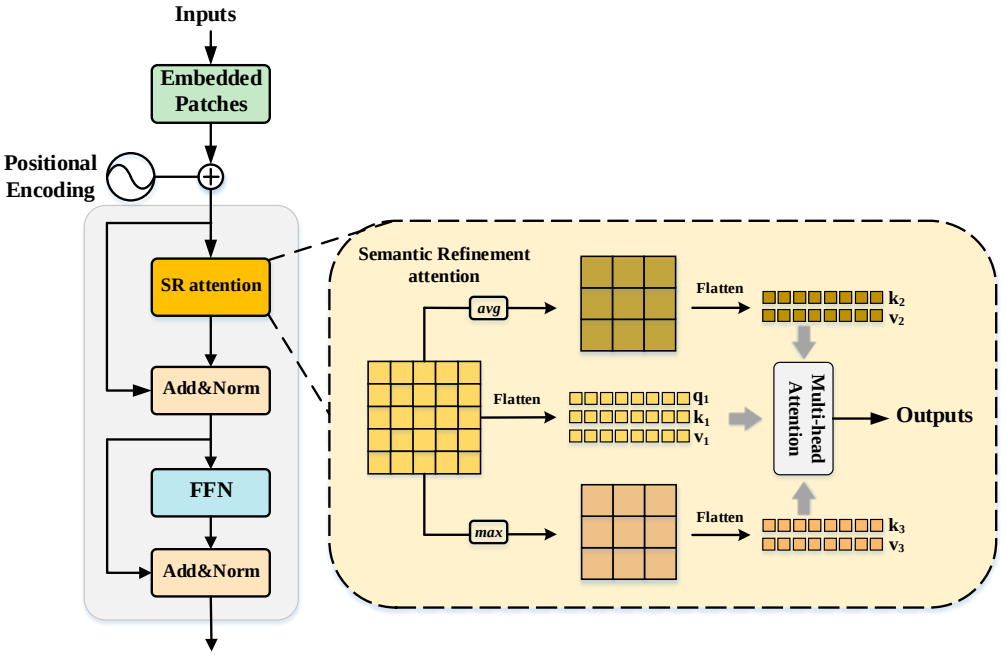

Figure 3. Illustration of semantic-aware refinement transformer encoder.

In summary, our proposed SRT comprehensively investigates the semantic information across different states of the high-level feature. This facilitates the refinement and enhancement of multi-scale semantic details through long-range relationship interactions. Moreover, the computational cost remains minimal due to the small spatial size of the high-level feature.

3.4. Bidirectional Fine-Grained Interaction

While acquiring the appropriate input for the merging process, a more effective interaction of features among various levels becomes essential. In a typical feature pyramid, a top-down pathway connects features from high to low levels in a progressive manner. Low-level features are enriched with semantic information from higher levels, which proves advantageous for classification tasks. Nevertheless, detection tasks demand sufficient

information pertinent to both classification and regression tasks, which poses a challenge due to the differing information needs of these tasks. The regression task mandates precise object contours and detailed information from high-resolution levels. Additionally, the classification task necessitates ample semantic information from low-resolution levels. However, the FPN scheme is not fully harnessed, resulting in the underutilization of high-resolution information from lower levels. The integration of numerous object contours and detailed information does not occur as effectively as anticipated. Furthermore, the semantic information gradually diminishes along the top-down path.

Building upon the aforementioned knowledge, we introduce bidirectional fine-grained interaction to address the challenge of underutilizing multi-scale features and to foster interplay across distinct levels. Initially, we recognize that a straightforward bottom-up path could potentially introduce additional noise in lower levels. Therefore, we devise a pixel-level filter (PLF), depicted in Figure 2b, which centers on salient locations and dynamically sieves out extraneous pixel-level information based on the current feature's characteristics. Moreover, high-level features often lack location-specific information. As a solution, we introduce a bottom-up scheme where low-level features employ the pixel-level filter to guide high-level features towards object-specific locations.

The pixel-level filter comprises two primary components: the identification of salient locations and the removal of superfluous pixel-level information, as well as the provision of fine-grained location guidance. The initial component, referred to as the pixel-level filter, can be outlined as follows:

$$W_i = Max[Tanh(\Phi(M_i) + Tanh(\Phi(M_i)) \times M_i), 0], \tag{10}$$

where $Tanh(\cdot)$ is tanh activation that transforms the operation into an encoded feature vector, ranging from $(-1, 1)$; $\Phi(\cdot)$ refers to a 1×1 conv operation; and Max ensures non-negativity. W_i is the output of PLF that denotes the filter result of M_i. The pixel-level filter effectively removes superfluous information by suppressing values below 0 and dynamically emphasizes the salient region. In the subsequent part, the adjacent layer M'_{i+1} is guided by the filter results W_i from preceding layers, facilitating focus on the desired region:

$$M_{i+1} = \mathscr{G}(\Phi(M'_{i+1}), F_{mul}(M'_{i+1}, W_i)) \tag{11}$$

$\Phi(\cdot)$ is a convolution operator applied to M_i with the intention of obtaining a focused region through a learning strategy. M_{i+1} signifies the output of interaction. It is obtained by matching the M'_{i+1} with the prominent information derived from preceding layers. M_2 remains unchanged, equivalent to M'_2.

Upon acquiring features enriched with accurate object contour and detailed information, we incorporate the concept of channel-wise prompt to facilitate the propagation of semantic information. As shown in Figure 2c, channel-wise prompt is devoted to extracting the semantic prompt map of the feature at the channel level, adaptively. Then, we utilize the semantic prompt map of higher levels to instruct the adjacent layer, which can heighten the semantic perception ability of objects. The detailed process can be articulated as:

$$R_i = Tanh\{Tanh[\Phi(avg(P_i))] + Tanh[\Phi(max(P_i))]\}, \tag{12}$$

where R_i denotes the semantic prompt map of high-level features, and avg and max represent the average pooling and max pooling operation block. Then, P'_{i-1} learns the semantic knowledge according to the prompt map. The process can be written as:

$$P_{i-1} = \mathscr{G}(\Phi(P'_{i-1}), F_{mul}(P'_{i-1}, R_i)). \tag{13}$$

The proposed bidirectional fine-grained interaction takes full advantage of multi-scale features. During the bidirectional interaction process, both semantic and spatial information can be effectively completed among different levels. The low-level layers, which possess

high-resolution information, effectively capture salient location information via pixel-level filtering at the pixel level. This information is then utilized to establish a bottom-up information flow. This aids in enhancing the essential location information of objects within high-level layers. Conversely, the high-level layers, abundant in semantic information, contribute significant semantic prompts when subjected to channel-wise prompting at the channel level. The prominent semantic prompt can be effectively transmitted to the low-level layers with minimal loss. BGI promotes adequate interaction among different levels with abundant multi-scale information.

4. Experiments

4.1. Settings

Dataset and Evaluation Metrics. Our experiments utilize the MS COCO dataset, a publicly available and reputable dataset comprising 80 distinct object categories. It consists of 115 k images for training (*train*2017) and 5k images for validation (*val*2017). Training is conducted on the *train*2017, while ablation experiments and comparable results are generated using the *val*2017. The performance assessment utilizes standard COCO-style average precision (AP) metrics, incorporating varying intersection over union (IoU) thresholds ranging from 0.5 to 0.95. AP_s, AP_m, and AP_l represent the AP of small, medium, and large objects. Moreover, AP^b and AP^m denote the AP of the bounding box and mask in the instance segmentation task.

Implementaion Details. To maintain experimental comparison fairness, all experiments are conducted utilizing PyTorch [56] and mmdetection [57]. In our configuration, input images are resized to ensure their shorter side measures 800 pixels. We train detectors with 8 Nvidia V100 GPUs (2 images per GPU) for 12 epochs. The initial learning rate is 0.02. And it is reduced by a factor of 0.1 after the 8th and 11th epochs, respectively. The backbones utilized in our experiments are publicly available and have been pretrained on ImageNet [58]. The training process incorporates linear warming up during the initial stage. All remaining hyperparameters remain consistent with the configurations outlined by mmdetection. Unless stated otherwise, all baseline methods incorporate FPN, and the ablation studies utilize Faster R-CNN based on ResNet50.

4.2. Ablation Studies

4.2.1. Ablation Studies on Three Components

To assess the significance of the components within MSBA, we progressively integrate three modules into the model. For all our ablation studies, the baseline method employed is Faster R-CNN with FPN, based on ResNet-50. As indicated in Table 1, MCF enhances the baseline method by 1.2 AP, owing to the utilization of diverse-resolution images and a cascaded dilated convolution fusion strategy. Multi-resolution images encompass ample spatial object information, while the cascaded method provides diverse receptive field messages. MCF effectively furnishes adequate information for objects of varying scales—small, medium, and large. SRT contributes a 1.3 AP enhancement to the baseline method by refining long-range relationships within high-level features. The most substantial contribution to the superior performance stems from the enhancements in AP_L (+2.9 AP), facilitated by ample semantic information. The findings suggest a deficiency in semantic information within the high-level features of the baseline method. SRT rectifies this shortfall by refining semantic information and enhancing feature representation in the high-level layer. BFI boosts detection performance by 1.4 AP, with a noteworthy improvement in AP_S. Evidently, robust interaction across various levels is conducive to mitigating scale variations. Furthermore, the fine-grained messages proficiently enhance detail and contour information across multi-scale features.

Combining any two of these components results in significantly improved performance compared to the baseline method, underscoring the efficacy of their synergistic interaction. For instance, the simultaneous integration of MCF and SRT yields an AP improvement of 39.0, surpassing the enhancement achieved by either module individually.

Furthermore, the incorporation of all three components with the baseline method results in an AP of 39.5. These ablation results substantiate the efficacy of the three individual components and their combined configurations, affirming their mutual complementarity.

Table 1. Effect of each component. Results are evaluated on COCO *val*2017. MCF: multi-resolution cascaded fusion, SRT: semantic-aware refinement transformer, BFI: bidirectional fine-grained interaction.

MCF	SRT	BFI	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
			37.4	58.1	40.4	21.2	41.0	48.1
√			38.6	59.4	41.9	22.2	42.1	49.9
	√		38.7	59.3	42.1	21.7	41.9	51.0
		√	38.8	59.7	42.4	22.6	42.4	50.7
√	√		39.0	59.9	42.4	22.0	42.4	50.7
√		√	39.1	60.4	42.6	22.4	42.9	50.6
	√	√	39.2	60.7	42.5	23.2	42.9	50.2
√	√	√	**39.5**	**60.4**	**42.8**	**22.1**	**42.9**	**52.3**

4.2.2. Ablation Studies of Various Dilation Rates

Table 2 presents the experimental results from various implementations of MCF. To validate the efficacy of MCF, we employed distinct dilation rates. Employing narrower dilation rates such as 1, 2, 3 and 2, 3, 4 yields constrained enhancements owing to the insufficiency of spatial information. Conversely, when employing dilation rates of 3, 6, 12, the performance fails to improve as anticipated. This suggests that the substantial disparity among the three dilation rates might result in incongruous receptive information. The more favorable outcome underscores the dominance of the appropriate configuration 1, 3, 6, which effectively provides ample pragmatic information for multi-level features.

Table 2. Comparsion of different dilation rates in MCF on COCO *val*2017.

Rates	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
(1, 2, 3)	38.0	58.7	41.4	21.8	41.6	48.8
(2, 3, 4)	37.9	58.6	41.1	21.6	41.3	48.8
(3, 6, 12)	38.2	59.1	41.5	22.0	41.5	49.6
(1, 3, 6)	**38.6**	**59.4**	**41.9**	**22.2**	**42.1**	**49.9**

4.2.3. Ablation Studies of Different Fusion Styles

Subsequently, we delve into the fusion techniques employed for combining two features within the MCF. The experiments are performed using distinct fusion styles within the matching gate. Initially, we employ the product operation on the two features to derive the fused feature. Subsequently, we sum the two features in another experiment for comparison purposes. As shown in Table 3, the summation operation applied to feature fusion yields superior performance, effectively preserving ample spatial and semantic information from both features.

Table 3. Comparsion of fusion style in the matching gate in MCF on COCO *val*2017.

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
baseline	37.4	58.1	40.4	21.2	41.0	48.1
product	38.3	58.9	41.8	21.6	41.9	50.0
sum	**38.6**	**59.4**	**41.9**	**22.2**	**42.1**	**49.9**

4.2.4. Ablation Studies of the Effect of Individual Component in BFI

In this section, we undertake comparative experiments to ascertain the efficacy of individual components within BFI. We employ two distinct directional structures to facilitate interaction independently. As shown in Table 4, both components enhance the performance of the baseline method. Furthermore, the outcomes reveal the superiority of combining both methods. The PLF and CWP are complementary and partially overlapping, leading to enhanced performance when combined.

Table 4. Comparsion of the effect of each component in BFI on COCO *val*2017. PLF: pixel-level filter, CWP: channel-wise prompt.

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
baseline	37.4	58.1	40.4	21.2	41.0	48.1
PLF	38.3	59.4	41.4	21.8	41.5	48.6
CWP	38.4	59.0	41.4	22.0	41.9	48.4
PLF with CWP	**38.8**	**59.7**	**42.4**	**22.6**	**42.4**	**50.7**

4.2.5. Ablation Studies of the Interaction Order

We subsequently undertake relevant experiments to validate the significance of interaction order between the two structures within BFI. The experiment is conducted by interchanging the positions of CWP and PLF. As shown in Table 5, the sequence of CWP followed by PLF surpasses other alternatives. However, following CWF, the PLF may introduce more noise and background information to high-level features. In contrast, when PLF precedes CWP, it effectively mitigates the aforementioned issues owing to the influence of semantic guidance.

Table 5. Comparsion of interaction orders in BFI on COCO *val*2017.

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	APL
baseline	37.4	58.1	40.4	21.2	41.0	48.1
CWP $\bigoplus$ PLF	38.6	59.1	42.0	21.5	42.1	50.2
PLF $\bigoplus$ CWP	**38.8**	**59.7**	**42.4**	**22.6**	**42.4**	**50.7**

4.3. Performance Comparison

To ascertain the efficacy and superiority, we perform comprehensive experiments encompassing both object detection and instance segmentation tasks. Furthermore, we re-implement the baseline methods using mmdetection to ensure equitable comparisons. Generally, the resulting performances surpass those reported in public articles. Additionally, we apply our proposed approach across multiple backbones and detectors, employing extended training schedules and techniques to demonstrate its generalizability.

4.3.1. Object Detection

As shown in Table 6, detectors incorporating MSBA consistently achieve substantial enhancements in comparison to conventional methods, encompassing both single-stage and multi-stage detectors. Our proposed MSBA demonstrates improvements of 1.5 and 2.1 points when integrated with RetinaNet and Faster R-CNN utilizing ResNet 50, respectively. Leveraging the ample coarse-grained information at lower levels, multi-stage detectors exhibit a more pronounced accuracy enhancement. Moreover, when combined with diverse backbones in conjunction with more sophisticated detectors, our approach attains superior outcomes, attributable to the reinforced multi-scale representation. Additionally, as depicted in Figure 4, MSBA effectively captures substantial spatial information through ample interaction, while mitigating the impact of erroneous and overlooked detections.

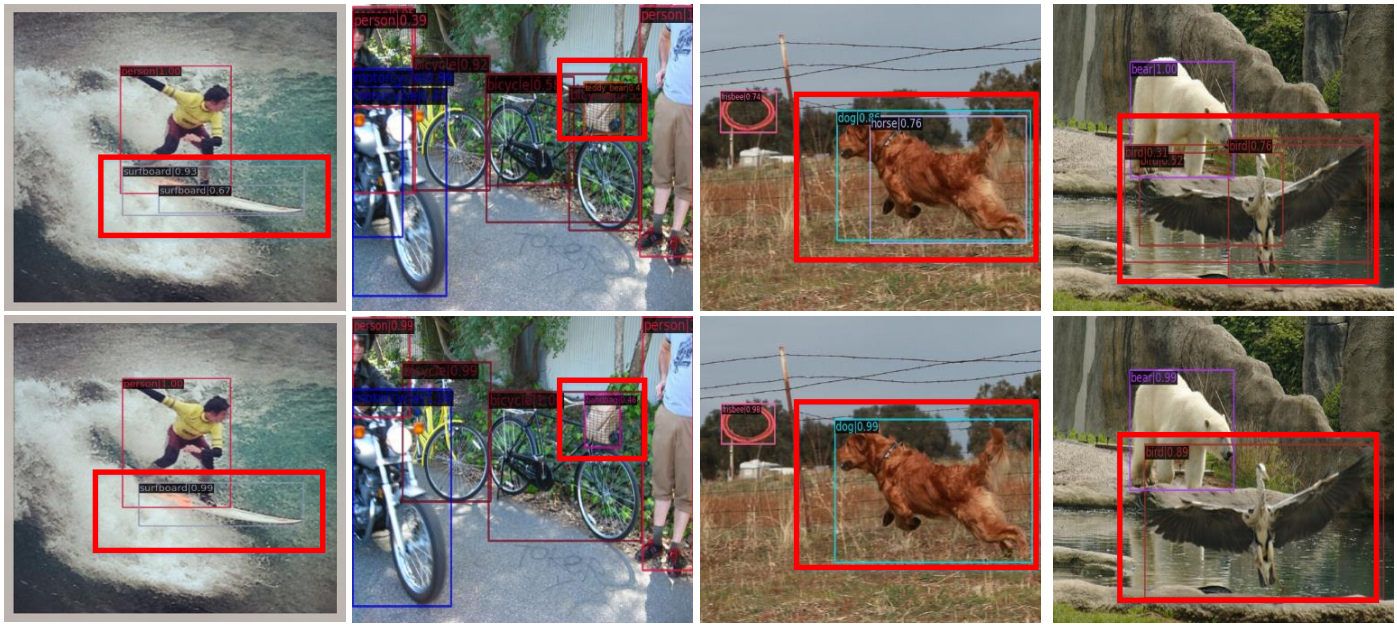

Figure 4. Example pairs of object detection results. (**Top row**) The outcomes are obtained using Faster R-CNN with FPN. (**Bottom row**) In contrast to Faster R-CNN with FPN, our MSBA method markedly enhances the localization capability of multi-scale objects through substantial interaction across diverse levels, as illustrated qualitatively.

Table 6. Object Detection: Performance comparisons with typical detectors based on FPN. "MSBA" represents our proposed adapter. "√" denotes the methods equipped with MSBA.

Method	Backbone	MSBA	AP^b	AP^b_S	AP^b_M	AP^b_L
RetinaNet	R50		36.5	20.4	40.3	48.1
		√	38.0	22.3	41.6	48.8
			(+1.5)	**(+1.9)**	**(+1.3)**	**(+0.7)**
	R101		38.5	21.7	42.8	50.4
		√	39.7	22.9	43.5	51.2
			(+1.2)	**(+1.2)**	**(+0.7)**	**(+0.8)**
Faster R-CNN	R50		37.4	21.2	41.0	48.1
		√	39.5	22.6	42.9	52.3
			(+2.1)	**(+1.4)**	**(+1.9)**	**(+4.2)**
	R101		39.4	22.4	43.7	51.1
		√	40.7	23.4	45.0	53.4
			(+1.3)	**(+1.0)**	**(+1.3)**	**(+2.3)**
Cascade R-CNN	R50		40.3	22.5	43.8	52.9
		√	41.9	23.9	45.5	55.4
			(+1.6)	**(+1.4)**	**(+1.7)**	**(+2.5)**
	R101		42.0	23.4	45.8	55.7
		√	42.6	23.8	46.8	57.0
			(+0.6)	**(+0.4)**	**(+1.0)**	**(+1.3)**

4.3.2. Instance Segmentation

We also conduct comprehensive experiments to confirm the superiority and generalizability of MSBA in the context of instance segmentation tasks. As shown in Table 7, our approach significantly enhances performance in both detection and instance segmentation tasks, exhibiting substantial advancements when contrasted with various robust models. Mask R-CNN achieves 41.7 AP on detection and 37.3 AP when equipped with MSBA based on ResNet-101. Despite the complexity of potent methods like HTC, MSBA exhibits a notable enhancement of 1.6 points in detection AP and 1.4 points in instance segmentation AP, both based on ResNet-50. Furthermore, MSBA achieves superior performance on large ob-

jects in both tasks, owing to substantial interaction and rich semantic information at higher levels. In addition, as shown in Figure 5, MSBA captures global semantic information, enabling accurate classification predictions and maintaining segmentation completeness.

Table 7. Instance Segmentation: Performance comparisons with powerful instance segmentation methodologies. All baseline approaches incorporate FPN. The † denotes the models trained with longer training schedules.

Method	Backbone	MSBA	AP^b	AP^b_S	AP^m	AP^m_L
Mask R-CNN	R50		38.2	21.9	34.7	47.2
		√	39.6	22.9	35.8	52.5
			(+1.4)	**(+1.0)**	**(+1.1)**	**(+5.3)**
	R101		40.0	22.6	36.1	49.5
		√	41.7	24.2	37.3	54.5
			(+1.7)	**(+1.6)**	**(+1.2)**	**(+5.0)**
Cascade Mask R-CNN	R50		41.2	23.9	35.9	49.3
		√	43.0	25.1	37.3	54.5
			(+1.8)	**(+1.2)**	**(+1.4)**	**(+5.2)**
	R101		42.9	24.4	37.3	51.5
		√	44.0	25.2	38.3	56.0
			(+1.1)	**(+0.8)**	**(+1.0)**	**(+4.5)**
HTC	R50		42.3	23.7	37.4	51.7
		√	43.9	25.6	38.8	56.7
			(+1.6)	**(+1.9)**	**(+1.4)**	**(+5.0)**
	R101 †		44.8	25.7	39.6	55.0
		√	45.7	27.0	40.2	59.2
			(+0.9)	**(+1.3)**	**(+0.6)**	**(+4.2)**

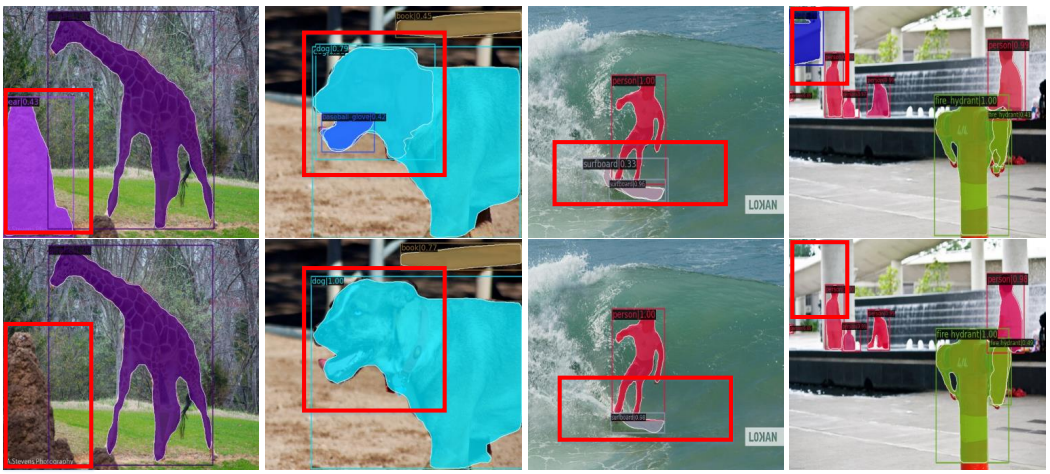

Figure 5. Example pairs of instance segmentation results. (**Top row**) The results are from Mask R-CNN with FPN. (**Bottom row**) our MSBA method significantly enhances the instance classification performance and effectively mitigates duplicate bounding boxes within densely populated regions, as demonstrated qualitatively.

4.3.3. Comparison on Transformer-Based Method

We further substantiate the generalizability of MSBA across transformer-based methods. As indicated in Table 8, we undertake relevant experiments encompassing both single-stage and two-stage detectors for both tasks. Our MSBA approach yields improvements of 1.2 and 0.9 points in the detection task when applied to pvt-tiny and swin-tiny methods, respectively. Moreover, even employing the same techniques, such as extended training schedules and multi-scale training, MSBA continues to demonstrate effectiveness and superiority when utilized with the more potent Swin-Small backbone, resulting in a 0.5-point enhancement over the baseline method. Due to the extensive multi-scale represen-

tation facilitated by MSBA, the performance improvement for small objects in the detection task is particularly notable.

Table 8. Comparison with transformer-based backbone on object detection: Performance comparisons paired with Mask R-CNN. The baseline methods are integrated with FPN. † represents the models trained with extra tricks such as multi-scale crop and longer training schedule.

Method	Backbone	MSBA	AP^b	AP^b_S	AP^m	AP^m_L
RetinaNet	PVT-Tiny		36.6	21.9	-	-
		√	37.8	23.0	-	-
			(+1.2)	**(+1.1)**	-	-
	PVT-Small		40.4	24.8	-	-
		√	40.9	25.3	-	-
			(+0.5)	**(+0.5)**	-	-
Mask R-CNN	Swin-Tiny		42.7	26.5	39.3	57.8
		√	43.6	27.8	39.9	58.4
			(+0.9)	**(+1.3)**	**(+0.6)**	**(+0.6)**
	Swin-Tiny †		46.0	31.3	41.7	59.7
		√	47.1	31.9	42.4	60.5
			(+1.1)	**(+0.6)**	**(+0.7)**	**(+0.8)**
	Swin-Small †		48.2	32.1	43.2	62.1
		√	48.7	32.8	43.4	62.8
			(+0.5)	**(+0.7)**	**(+0.2)**	**(+0.7)**

4.3.4. Comparison with State-of-the-Art Methods

We evaluate MSBA based on more expressive methods with the longer training schedule and various tricks, compared with other state-of-the-art object detection approaches. To ensure equitable comparisons, we re-implement the corresponding baseline models, incorporating FPN within mmdetection. As shown in Table 9, MSBA consistently attains notable improvements, even when employed with more potent backbones, encompassing both CNN-based and Transformer-based configurations. MSBA achieves 42.1 AP and 43.0 AP when employing ResNeXt101-32×4d and ResNeXt101-64×4d as the feature extractors of Faster R-CNN, respectively. This marks an enhancement of 0.9 points compared to the FPN counterparts. When applied to transformer-based detectors employing identical training schedules and strategies, the consistently superior performance underscores the applicability of MSBA across various detector architectures. Additionally, we assess our approach on more potent models like HTC with a 20-epoch training schedule and Mask R-CNN with a 36-epoch training schedule. This leads to enhancements of 0.8 and 0.5 points in detection AP for ResNeXt101-32×4d and Swin-Small, respectively. Consequently, our approach yields substantial enhancements across diverse public backbones and distinct tasks. The enhanced performance serves as evidence of MSBA's capacity for generalization and robustness.

4.4. Error Analyses

Subsequently, we conduct error analyses to further substantiate the effectiveness of our approach. As illustrated in Figure 6, we randomly select four categories for error analysis, encompassing objects of diverse scales. Our approach outperforms the baseline method across various thresholds. When disregarding localization errors, MSBA surpasses the baseline, attributed to our approach's ability to offer more accurate classification information. Furthermore, when excluding errors associated with similar classes from the same supercategory and different classes, our method exhibits noteworthy enhancements compared to the baseline. This underscores MSBA's superior location accuracy.

Table 9. Comparisons with the states of the art: The symbol "*" signifies our re-implemented results on mmdetection. "Schedule" refers to the learning schedules of the respective methods. The † symbol indicates models trained with additional tricks, such as multi-scale training.

Method	Backbone	Schedule	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster R-CNN *	ResNet50-DCN	12	41.3	62.4	45.0	24.6	44.9	54.4
Faster R-CNN *	ResNet101-DCN	12	42.7	63.8	46.4	24.9	46.7	56.8
Faster R-CNN *	ResNeXt101-32×4d	12	41.2	62.1	45.1	24.0	45.5	53.5
Faster R-CNN *	ResNeXt101-64×4d	12	42.1	63.0	46.3	24.8	46.2	55.3
Mask R-CNN *	ResNet50-DCN	12	41.8	62.7	46.2	24.5	45.3	55.4
Mask R-CNN *	ResNet101-DCN	12	43.5	64.3	47.9	25.7	47.7	57.5
Mask R-CNN *	ResNeXt101-32×4d	12	41.9	62.5	45.9	24.4	46.3	54.0
Cascade R-CNN *	ResNet50-DCN	12	43.8	62.6	47.9	26.3	47.2	58.5
Cascade R-CNN *	ResNeXt101-32×4d	12	43.7	62.3	47.7	25.1	47.6	57.3
DETR[4]	ResNet50	500	42.0	62.4	44.2	20.5	45.8	61.1
DETR[4]	ResNet101	500	43.5	63.8	46.4	21.9	48.0	61.8
Deformable DETR[26]	ResNet50	50	43.8	62.6	47.7	26.4	47.1	58.0
Sparse R-CNN[27]	ResNet101	36	44.1	62.1	47.2	26.1	46.3	59.7
HTC *	ResNet101	20	44.8	63.3	48.8	25.7	48.5	60.2
Mask R-CNN *†	Swin-Tiny	36	46.0	68.2	50.3	30.5	49.2	59.5
HTC *	ResNeXt101-32×4d	20	46.1	65.3	50.1	27.1	49.6	60.9
Mask R-CNN *†	Swin-Small	36	48.2	69.8	52.8	32.1	51.8	62.7
MSBA Faster R-CNN	ResNet50-DCN	12	42.2	63.3	46.2	25.3	46.0	55.7
MSBA Faster R-CNN	ResNet101-DCN	12	43.4	64.4	47.5	25.7	47.4	57.8
MSBA Faster R-CNN	ResNeXt101-32×4d	12	42.1	63.3	45.7	24.7	46.6	54.8
MSBA Faster R-CNN	ResNeXt101-64×4d	12	43.0	64.3	47.1	25.3	46.9	56.9
MSBA Mask R-CNN	ResNet50-DCN	12	43.1	63.9	47.4	25.8	47.1	57.0
MSBA Mask R-CNN	ResNet101-DCN	12	44.2	64.9	48.4	25.9	48.3	58.5
MSBA Mask R-CNN	ResNeXt101-32×4d	12	43.1	64.0	46.9	26.2	47.1	56.2
MSBA Cascade R-CNN	ResNet50-DCN	12	44.6	63.6	48.8	27.0	48.2	59.3
MSBA Cascade R-CNN	ResNeXt101-32×4d	12	44.2	63.0	47.8	25.4	48.4	58.3
MSBA HTC	ResNet101	20	45.7	64.7	49.6	27.0	49.5	60.6
MSBA Mask R-CNN †	Swin-Tiny	36	47.1	68.8	51.5	31.9	50.2	60.6
MSBA HTC	ResNeXt101-32×4d	20	46.9	66.4	51.2	28.6	50.6	61.7
MSBA Mask R-CNN †	Swin-Small	36	**48.7**	**70.6**	**53.5**	**32.8**	**52.5**	**63.1**

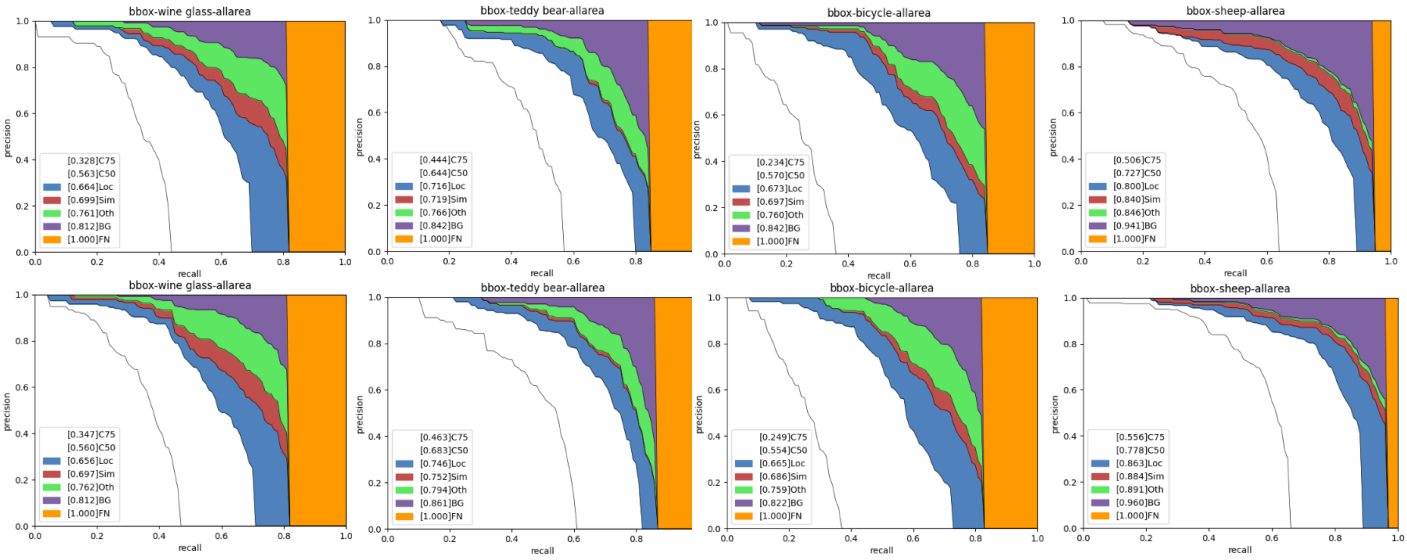

Figure 6. The error analyses of four categories: The results in the first row correspond to the baseline, while those in the second row correspond to MSBA.

5. Conclusions

In this paper, we introduce a novel and efficacious multi-resolution and semantic-aware bidirectional adapter, denoted as MSBA, for enhancing multi-scale object detection through adaptive feature integration. MSBA dissects the complete integration process into three segments, each dedicated to managing appropriate input, refined enhancement, and comprehensive interaction. The three corresponding constituents of MSBA, namely multi-resolution cascaded fusion (MCF), the semantic-aware refinement transformer (SRT), and bidirectional fine-grained interaction (BFI), are devised to address these three segments. Facilitated by these three simple yet potent components, MSBA demonstrates its adaptability across both two-stage and single-stage detectors, yielding substantial enhancements when contrasted with the baseline approach across the demanding MS COCO dataset.

Author Contributions: Conceptualization, Z.L. and J.P.; Methodology, Z.L. and B.L.; Validation, Z.L.; Formal analysis, P.H.; Writing—original draft, Z.L.; Writing—review & editing, Z.L. and Z.Z.; Visualization, Z.L.; Supervision, C.Z.; Funding acquisition, B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No. 62192785, No. 62372451, No. 62202469), the National Key Research and Development Program of China (No. 2022YFC3321000), and the Beijing Natural Science Foundation (No. M22005, 4224091).

Data Availability Statement: The MSCOCO dataset that supports this study is openly available online at https://arxiv.org/abs/1405.0312. It is cited as the reference [5] in our manuscript.

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

References

1. Yang, Z.; Liu, S.; Hu, H.; Wang, L.; Lin, S. RepPoints: Point Set Representation for Object Detection. *arXiv* **2019**, arXiv:1904.11490.
2. Wang, X.; Zhang, S.; Yu, Z.; Feng, L.; Zhang, W. Scale-Equalizing Pyramid Convolution for Object Detection. In Proceedings of the CVPR 2020: Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13359–13368.
3. Guo, C.; Fan, B.; Zhang, Q.; Xiang, S.; Pan, C. AugFPN: Improving Multi-Scale Feature Learning for Object Detection. In Proceedings of the CVPR 2020: Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12595–12604.
4. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
5. Lin, T.Y.; Maire, M.; Belongie, S.J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 8–11 September 2014; pp. 740–755.
6. Singh, B.; Davis, L.S. An analysis of scale invariance in object detection snip. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3578–3587.
7. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 379–387.
8. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
9. Li, Z.; Liu, Y.; Li, B.; Feng, B.; Wu, K.; Peng, C.; Hu, W. Sdtp: Semantic-aware decoupled transformer pyramid for dense image prediction. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6160–6173. [CrossRef]
10. Sun, X.; Wang, P.; Wang, C.; Liu, Y.; Fu, K. PBNet: Part-based convolutional neural network for complex composite object detection in remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 50–65. [CrossRef]
11. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [CrossRef]
12. Yi, H.; Shi, S.; Ding, M.; Sun, J.; Xu, K.; Zhou, H.; Wang, Z.; Li, S.; Wang, G. Segvoxelnet: Exploring semantic context and depth-aware features for 3d vehicle detection from point cloud. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–4 June 2020; pp. 2274–2280.
13. Anand, B.; Barsaiyan, V.; Senapati, M.; Rajalakshmi, P. Region of interest and car detection using lidar data for advanced traffic management system. In Proceedings of the 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), Online, 15 June 2020; pp. 1–5.
14. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.

15. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
17. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Montreal, QC, Canada, 7–12 December 2018; pp. 6154–6162.
18. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
19. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
20. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
21. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-shot refinement neural network for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4203–4212.
22. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9626–9635.
23. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
24. Zhou, X.; Zhuo, J.; Krahenbuhl, P. Bottom-up object detection by grouping extreme and center points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 850–859.
25. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection. In Proceedings of the CVPR 2020: Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9759–9768.
26. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2020**. arXiv:2010.04159.
27. Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. Sparse r-cnn: End-to-end object detection with learnable proposals. *arXiv* **2020**, arXiv:2011.12450.
28. Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; Zhang, L. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv* **2022**, arXiv:2201.12329.
29. Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L.M.; Zhang, L. Dn-detr: Accelerate detr training by introducing query denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 21–24 June 2022; pp. 13619–13627.
30. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.Y. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv* **2022**, arXiv:2203.03605.
31. Pathiraja, B.; Gunawardhana, M.; Khan, M.H. Multiclass Confidence and Localization Calibration for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19734–19743.
32. Ge, C.; Song, Y.; Ma, C.; Qi, Y.; Luo, P. Rethinking Attentive Object Detection via Neural Attention Learning. *IEEE Trans. Image Process.* **2023**. [CrossRef]
33. Zohar, O.; Wang, K.C.; Yeung, S. Prob: Probabilistic objectness for open world object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 11444–11453.
34. Singh, B.; Najibi, M.; Davis, L.S. SNIPER: Efficient multi-scale training. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 9310–9320.
35. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
36. Chen, K.; Cao, Y.; Loy, C.C.; Lin, D.; Feichtenhofer, C. Feature Pyramid Grids. *arXiv* **2020**, arXiv:2004.03580.
37. Li, Z.; Liu, Y.; Li, B.; Hu, W.; Zhang, H. DSIC: Dynamic Sample-Individualized Connector for Multi-Scale Object Detection. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021.
38. Yan, Z.; Qi, Y.; Li, G.; Liu, X.; Zhang, W.; Yang, M.H.; Huang, Q. Progressive Multi-resolution Loss for Crowd Counting. *IEEE Trans. Circuits Syst. Video Technol.* **2023**. [CrossRef]
39. Gu, J.; Kwon, H.; Wang, D.; Ye, W.; Li, M.; Chen, Y.H.; Lai, L.; Chandra, V.; Pan, D.Z. Multi-scale high-resolution vision transformer for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–21 June 2022; pp. 12094–12103.
40. Li, Y.; Wu, C.Y.; Fan, H.; Mangalam, K.; Xiong, B.; Malik, J.; Feichtenhofer, C. Mvitv2: Improved multiscale vision transformers for classification and detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–21 June 2022; pp. 4804–4814.
41. 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.

42. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv* **2021**, arXiv:2101.11986.
43. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in transformer. *arXiv* **2021**, arXiv:2103.00112.
44. Chu, X.; Zhang, B.; Tian, Z.; Wei, X.; Xia, H. Do we really need explicit position encodings for vision transformers? *arXiv* **2021**, arXiv:2102.10882.
45. 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.
46. Ye, H.; Li, G.; Qi, Y.; Wang, S.; Huang, Q.; Yang, M.H. Hierarchical modular network for video captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–21 June 2022; pp. 17939–17948.
47. Gu, X.; Chen, G.; Wang, Y.; Zhang, L.; Luo, T.; Wen, L. Text with Knowledge Graph Augmented Transformer for Video Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18941–18951.
48. An, D.; Qi, Y.; Li, Y.; Huang, Y.; Wang, L.; Tan, T.; Shao, J. BEVBert: Topo-Metric Map Pre-training for Language-guided Navigation. *arXiv* **2022**, arXiv:2212.04385.
49. Majumdar, A.; Shrivastava, A.; Lee, S.; Anderson, P.; Parikh, D.; Batra, D. Improving vision-and-language navigation with image-text pairs from the web. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 259–274.
50. Chen, Q.; Tan, M.; Qi, Y.; Zhou, J.; Li, Y.; Wu, Q. V2C: Visual voice cloning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–21 June 2022; pp. 21242–21251.
51. Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. Fastspeech: Fast, robust and controllable text to speech. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 3171–3180.
52. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv* **2021**, arXiv:2102.12122.
53. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. *arXiv* **2021**, arXiv:2103.15808.
54. Patel, K.; Bur, A.M.; Li, F.; Wang, G. Aggregating Global Features into Local Vision Transformer. *arXiv* **2022**, arXiv:2201.12903.
55. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You only look one-level feature. *arXiv* **2021**, arXiv:2103.09460.
56. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic Differentiation in Pytorch. In Proceedings of the NIPS 2017 Workshop Autodiff, Long Beach, CA, USA, 9 December 2017.
57. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv* **2019**, arXiv:1906.07155.
58. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.