Towards Efficient Use of Multi-Scale Features in Transformer-Based Object Detectors

Gongjie Zhang\textsuperscript{1,2} Zhipeng Luo\textsuperscript{1,3} Zichen Tian\textsuperscript{1} Jingyi Zhang\textsuperscript{1} Xiaqin Zhang\textsuperscript{4} Shijian Lu\textsuperscript{1}

\textsuperscript{1}S-Lab, Nanyang Technological University  \textsuperscript{2}Black Sesame Technologies  \textsuperscript{3}SenseTime Research  \textsuperscript{4}Wenzhou University

gjz@ieee.org  zhipeng001@e.ntu.edu.sg  shijian.lu@ntu.edu.sg

Abstract

Multi-scale features have been proven highly effective for object detection but often come with huge and even prohibitive extra computation costs, especially for the recent Transformer-based detectors. In this paper, we propose Iterative Multi-scale Feature Aggregation (IMFA) — a generic paradigm that enables efficient use of multi-scale features in Transformer-based object detectors. The core idea is to exploit sparse multi-scale features from just a few crucial locations, and it is achieved with two novel designs. First, IMFA rearranges the Transformer encoder-decoder pipeline so that the encoded features can be iteratively updated based on the detection predictions. Second, IMFA sparsely samples scale-adaptive features for refined detection from just a few keypoint locations under the guidance of prior detection predictions. As a result, the sampled multi-scale features are sparse yet still highly beneficial for object detection. Extensive experiments show that the proposed IMFA boosts the performance of multiple Transformer-based object detectors significantly yet with only slight computational overhead.

1. Introduction

Detecting objects of vastly different scales has always been a major challenge in object detection [28]. Fortunately, strong evidence [11, 22, 25, 48, 69, 72] shows that object detectors can significantly benefit from multi-scale features while dealing with large scale variation. For ConvNet-based object detectors like Faster R-CNN [42] and FCOS [49], Feature Pyramid Network (FPN) [25] and its variants [12, 18, 19, 30, 48, 69, 70] have become the go-to components for exploiting multi-scale features.

Other than ConvNet-based object detectors, the recently proposed DEtection TRansformer (DETR) [4] has established a fully end-to-end object detection paradigm with promising performance. However, naively incorporating multi-scale features using FPN in these Transformer-based detectors [4, 11, 20, 29, 35, 55, 66, 72] often brings enormous and even unfeasible computation costs, primarily due to the poor efficiency of the attention mechanism in processing high-resolution features. Concretely, to handle a feature map with a spatial size of $H \times W$, ConvNet requires a computational cost of $O(HW)$, while the complexity of the attention mechanism in Transformer-based object detectors is $O(H^2W^2)$. To mitigate this issue, Deformable DETR [72] and Sparse DETR [43] replace the original global dense attention with sparse attention. SMCA-DETR [11] restricts most Transformer encoder layers to be scale-specific, with only one encoder layer to integrate multi-scale features. However, as the number of tokens increases quadratically w.r.t. feature map size (typically 20x~80x compared to single-scale), these methods are still costly in computation and memory consumption, and rely on special operators like

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{The proposed Iterative Multi-scale Feature Aggregation (IMFA) is a generic approach for efficient use of multi-scale features in Transformer-based object detectors. It boosts detection accuracy on multiple object detectors at minimal costs of additional computational overhead. Results are obtained with ResNet-50. Best viewed in color.}
\end{figure}
deformable attention [72] that introduces extra complexity for deployment. To the best of our knowledge, there is yet no generic approach that can efficiently exploit multi-scale features for Transformer-based object detectors.

In this paper, we present Iterative Multi-scale Feature Aggregation (IMFA), a concise and effective technique that can serve as a generic paradigm for efficient use of multi-scale features in Transformer-based object detectors. The motivation comes from two key observations: (i) the computation of high-resolution features is highly redundant as the background usually occupies most of the image space, thus only a small portion of high-resolution features are useful to object detection; (ii) unlike ConvNet, the Transformer’s attention mechanism does not require grid-shaped feature maps, which offers the feasibility of aggregating multi-scale features only from some specific regions that are likely to contain objects of interest. The two observations motivate us to sparsely sample multi-scale features from just a few informative locations and then aggregate them with encoded image features in an iterative manner.

Concretely, IMFA consists of two novel designs in the Transformer-based detection pipelines. First, IMFA rearranges the encoder-decoder pipeline so that each encoder layer is immediately connected to its corresponding decoder layer. This design enables iterative update of encoded image features along with refined detection predictions. Second, IMFA sparsely samples multi-scale features from the feature pyramid generated by the backbone, with the sampling process guided by previous detection predictions. Specifically, motivated by the spatial redundancy of high-resolution features, IMFA only focuses on a few promising regions with high likelihood of object occurrence based on prior predictions. Furthermore, inspired by the significance of objects’ keypoints for recognition and localization [39, 59, 66, 71], IMFA first searches several keypoints within each promising region, and then samples useful features around these keypoints at adaptively selected scales. The sampled features are finally fed to the subsequent encoder layer along with the image features encoded by the previous layer. With the two new designs, the proposed IMFA aggregates only the most crucial multi-scale features from those informative locations. Since the number of the aggregated features is small, IMFA introduces minimal computational overhead while consistently improving the detection performance of Transformer-based object detectors. It is noteworthy that IMFA is a generic paradigm for efficient use of multi-scale features: (i) as shown in Fig. 1, it can be easily integrated with multiple Transformer-based object detectors with consistent performance boosts; (ii) as discussed in Section 5.4, IMFA has the potential to boost DETR-like models on tasks beyond object detection.

To summarize, the contributions of this work are threefold.

• We propose a novel DETR-based detection pipeline, where encoded features can be iteratively updated along with refined detection predictions. This new pipeline allows to leverage intermediate predictions as guidance for robust and efficient multi-scale feature encoding.

• We propose a sparse sampling strategy for multi-scale features, which first identifies several promising regions under the guidance of prior detections, then searches several keypoints within each promising region, and finally samples their features at adaptively selected scales. We demonstrate that such sparse multi-scale features can significantly benefit object detection.

• Based on the two contributions above, we propose Iterative Multi-scale Feature Aggregation (IMFA) – a simple and generic paradigm that enables efficient use of multi-scale features in Transformer-based object detectors. IMFA consistently boosts detection performance on multiple object detectors, yet remains computationally efficient. This is the pioneering work that investigates a generic approach for exploiting multi-scale features efficiently in Transformer-based object detectors.

2. Related Work

Object Detection. Most modern object detectors, like Faster R-CNN [42], YOLO [40], and FCOS [49], are ConvNet-based. They have achieved promising results on various detection benchmarks [2, 7, 17, 24, 38, 44, 48, 54, 57, 61, 62]. However, these methods detect objects by defining surrogate regression and classification tasks, which rely on many hand-crafted components, such as anchors, rule-based training target assignment, and non-maximum suppression (NMS). Thus the detection pipelines of these ConvNet-based detectors are complex, hyper-parameter-intensive, and not fully end-to-end, leading to sub-optimal performance. Unlike ConvNet-based detectors, the recently proposed DETR [4] has revolutionized the paradigm for object detection using a Transformer [50] encoder-decoder architecture, eliminating the need for those hand-crafted components. Inspired by DETR [4], many Transformer-based object detectors [1, 3, 5, 8, 13, 16, 20, 23, 29, 34, 36, 45, 52, 53, 64–68, 72] are proposed and achieve state-of-the-art detection accuracy as well as fast convergence.

Multi-Scale Features for Object Detection. One major challenge in object detection is to effectively represent objects at distinct scales. This is especially crucial for detecting small objects in images. In modern ConvNet-based detectors [26, 42, 48, 49, 54, 56, 70], Feature Pyramid Network (FPN) [25] and its variants [12, 18, 30, 69, 70] have become the go-to solutions to exploit multi-scale features. However, as feature pyramids require computation on high-resolution feature maps, FPN and its variants also introduce substantial computational overhead.
Multi-scale features are also helpful for Transformer-based object detectors. However, due to the inefficiency of Transformer’s attention mechanism [50] to process high-resolution feature maps, it requires special modifications to reduce the computational complexity to a feasible level. Concretely, Deformable DETR [72] proposes deformable attention, which reduces the complexity via key sparsification in the attention module. SMCA-DETR [11] uses only one multi-scale attention encoder layer while restricting other layers to be scale-specific. CF-DETR [3] embeds the Transformer encoder into an FPN [25] to produce feature pyramids, and extracts multi-scale features with RoIAlign [14]. These methods enable the use of multi-scale features in Transformer-based detectors, but introduce huge computational overhead, require large-memory GPUs for training and inference, and rely on special operators like deformable attention or RoIAlign. To the best of our knowledge, there is yet no generic approach to efficiently leverage multi-scale features for Transformer-based detectors so far.

Spatial Redundancy and Sparse Features. Not all features are equally important. In most cases, only a small portion of features are crucial for recognition. With this motivation, several works [9, 10, 41, 43, 51, 52, 72] perform sparse operations over feature maps to avoid computation at less informative locations. Specifically, in object detection, AutoFocus [37] first predicts and crops regions at coarse scales, and then makes final predictions on those regions at a higher resolution. PnP-DETR [52] and Sparse DETR [43] adaptively allocate encoding operations to informative feature tokens. One similar work to our proposed IMFA is QueryDet [58], which first coarsely predicts over low-resolution features, and then sparsely exploits multi-scale features based on the coarse predictions to generate the final detection results, thus improving inference speed. However, unlike our proposed IMFA, QueryDet is designed for single-stage ConvNet-based detectors with FPN [25], and it only accelerates the inference procedure.

Our proposed IMFA is also inspired by the spatial redundancy in high-resolution features. IMFA only exploits sparse features from only a few highly informative locations to get the best of both worlds for Transformer-based detectors – high detection accuracy and low computational cost.

3. A Revisit of Transformer-Based Detection

Since our proposed method is developed on top of the recently proposed Transformer-based object detectors, we first briefly review the detection pipeline of Transformer-based object detectors [4, 29, 35, 55], taking the pioneering work DETR [4] as an example.

DETR [4] formulates object detection as a direct set prediction problem and uses a Transformer [50] encoder-decoder architecture to solve it. Given an image $I \in \mathbb{R}^{H_0 \times W_0 \times 3}$, the backbone network generates its feature maps, which are further fed to the Transformer encoder to produce the encoded image features $F \in \mathbb{R}^{H \times W \times d}$, where $d$ denotes the feature dimension, and $H_0$, $W_0$ and $H$, $W$ are the spatial sizes of the input image and its feature maps, respectively. Then, the encoded features are fed to the Transformer decoder to interact with a set of object queries representing potential objects at different spatial locations. The object queries are finally used to produce final detection predictions with a feed-forward network (FFN). The entire detection pipeline is supervised by a set-based global loss with bipartite matching.

Specifically, both the Transformer encoder and decoder are composed of multiple layers. As shown in Fig. 2 (left),
existing methods \cite{4,11,29,35,55,72} usually process the input image features with a stack of encoder layers and obtain a fixed set of encoded features, which are further fed to the Transformer decoder layers to update the detection results iteratively. Differently, as illustrated in Fig. 2 (right), one major difference introduced by IMFA is that it rearranges the encoder-decoder pipeline into multiple stacked detection stages, so that encoded features can also be iteratively updated along with refined detection predictions. This design modification lays the foundation for efficient use of multi-scale features guided by prior detection results, which is to be detailed in the next section.

4. Iterative Multi-Scale Feature Aggregation

4.1. Overview

**Iterative Multi-scale Feature Aggregation (IMFA)** is a generic paradigm for efficient use of multi-scale features in Transformer-based object detectors, such as DETR \cite{4}. Fig. 3 illustrates the detection pipeline of the proposed IMFA. For computational efficiency, IMFA exploits multi-scale features with dual-sparsity: (i) it samples multi-scale features from just a few promising regions with high likelihood of object occurrence as guided by prior detection predictions; (ii) for each promising region, it only samples features from several keypoints with the most informative features at adaptively selected scales. The dual-sparsity is achieved with two novel designs, which are to be described in detail in the following subsections.

4.2. Iterative Update of Encoded Features

The iterative update of encoded image features is the basis for IMFA to exploit multi-scale features efficiently. As introduced in Section 3, most existing Transformer-based detectors use fixed encoded image features to make predictions. In order to guide the multi-scale sampling process with prior detections, IMFA rearranges the Transformer encoder-decoder pipeline, as shown in Fig. 2 (right).

Specifically, instead of using stacked encoder layers to produce a fixed set of feature tokens at one go, IMFA rearranges the detection pipeline into several stacked detection stages. Each detection stage consists of an encoder layer, a decoder layer, and an FFN. This design lays the foundation for incorporating sparse multi-scale features dynamically under the guidance of prior detection predictions, which is detailed in Section 4.3. It is noteworthy that, according to the experiments in Section 5.3, this design alone (shown in Fig. 2 (right), without incorporating multi-scale features) brings no performance gain over the baseline model.
Figure 4. Visualization of IMFA’s sampling locations and their adaptively selected feature scales. The searched sampling points mostly fall around the objects of interest, many of which are highly representative points with rich semantics, such as objects’ extremities. Besides, IMFA adaptively selects appropriate feature scales for each sampling point, generating sparse yet informative scale-adaptive features for refined detection predictions. Best viewed in color. More visualizations are provided in technical appendix.

4.3. Sparse Feature Sampling and Aggregation

Naively incorporating multi-scale features into the encoder leads to prohibitive computational complexity, as the number of feature tokens from all scales is too large to be processed by the attention mechanism. This motivates us to exploit only the most informative multi-scale features.

On the basis of Section 4.2, IMFA further performs sparse multi-scale feature sampling using prior detection predictions as guidance, as illustrated in Fig. 3. Specifically, IMFA first identifies a few promising regions with high likelihood of object occurrence. Then, it searches for several representative and informative keypoints within each promising region and samples their features at adaptively selected scales. Finally, the sampled features are fed to the subsequent encoder layers to aggregate with single-scale image features to produce refined detection predictions.

Identifying Promising Regions Based on Prior Predictions. In most cases, objects are sparsely distributed across images [27, 37, 58], which motivates us to exploit only the multi-scale features related to these objects. An intuitive solution is to guide the sampling process with the high-confidence detection predictions from the previous detection stage. Concretely, as shown in Fig. 3, for each detection stage except the first stage, we select $K$ predictions with the highest classification confidence scores from the previous detection stage as the promising regions. Here, $K = N \times r$, with $N$ denoting the number of object queries and $r$ denoting IMFA’s sampling ratio. Formally, we denote the selected box predictions and their corresponding object queries as $\{(B_1, Q_1), ..., (B_K, Q_K)\}$. The multi-scale features are then sampled within these promising regions, which is to be introduced in detail later. Since Transformer-based object detectors [4, 11, 29, 35, 55] already employ a sparse set (typically 100~300) of object queries to represent different objects, the promising regions sampled by IMFA remain sparse for efficient computation.

Sampling Scale-Adaptive Features from Representative Keypoints. IMFA directly samples multi-scale features from the feature pyramid that is generated from the backbone (C2-C5 from ResNet in our experiments). However, even the sparsely sampled promising regions still contain a substantial amount of feature tokens at high-resolution feature scales. To further sparsify the sampled multi-scale features, IMFA searches a small number of representative keypoints within each promising region and samples their corresponding features at adaptively selected scales.

As illustrated in Fig. 3, for each promising region, IMFA first uses its object query to predict $M$ keypoint locations within the region, which can be formulated as:

$$\{P_{ij}\}_{j=1}^{M} = \text{MLP}(Q_i) \quad \text{for } i = 1, 2, ..., K,$n(1)$$

where $i$ and $j$ index the queries and keypoints, respectively, and each keypoint $P_{ij} = (x_{ij}, y_{ij})$ lies within its corresponding box prediction $B_i$. Then, IMFA samples each keypoint’s features from the feature pyramid at all scales via bilinear interpolation, obtaining a set of features $\{F^s_{ij}\}_{s=1}^{S}$, where $S$ is the number of feature scales. Finally, to emphasize the distinct significance of different feature scales for each keypoint, we propose to perform adaptive scale selection by predicting scale-specific weights for each keypoint and obtaining scale-adaptive features through weighted summation:

$$F_{ij} = \sum_{s} \alpha^s_{ij} F^s_{ij} \quad \{\alpha^s_{ij}\}_{s=1}^{S} = \text{Softmax}(\gamma_{ij}(Q_i)),$n(2)$$

where the scale-selection weights $\alpha$ are generated by a linear projection $\gamma_{ij}$ followed by a Softmax function, so that $\sum_{s} \alpha^s_{ij} = 1$. In this way, IMFA only samples the most crucial and informative features, producing a set of sparse yet still highly informative multi-scale features for each promising region. Additionally, to further strengthen the representation capacity of the sampled multi-scale features, we feed the sampled features into a Dynamic Feed-
Forward Network (Dynamic FFN) to incorporate the semantics from their corresponding object queries via dynamic weighting [46], where FFN’s weights are dynamically generated by object queries. It can be formulated as:

\[ F'_{ij} = \text{MLP}_{\psi}(F_{ij}) \quad W_i = \psi(Q_i). \]  

Here, for each object query \( Q_i \), the dynamic weight \( W_i \) is obtained by a linear projection \( \psi \) of \( Q_i \). Then, \( W_i \) is applied to the scale-adaptive features \( F_{ij} \) to generate the final sampled features \( F'_{ij} \) with enhanced semantics. These sampled features, along with their positional embeddings obtained based on their keypoint locations, are further fed to the subsequent detection stage for aggregation.

**Iterative Aggregation of Multi-Scale Features.** To leverage the sampled multi-scale features for refined object detection, the sampled features and the encoded image features are fed into the subsequent encoder layer for aggregation using the attention mechanism. This is analogous to the top-down path created by FPN [25] for enhancing the semantics of low-level features. To avoid continuous growth of feature tokens and maintain efficiency, each detection stage does not inherit the multi-scale features that are generated from the previous stage, as shown in Fig. 3.

### 4.4. Visualization and Analysis

Fig. 4 visualizes IMFA’s sampling locations and their feature scales. It can be observed that the sampling locations mostly fall around the target objects, and typically at representative locations, such as object extremities. This proves the effectiveness of IMFA in searching sparse yet highly informative locations in the feature sampling process. Besides, it is noteworthy that IMFA tends to focus on higher-resolution features for small objects and lower-resolution features for large objects, which is intuitive as the detection of small objects relies more on finer details.

### 5. Experiments

#### 5.1. Experiment Setup

**Dataset and Evaluation Metrics.** We perform experiments on the COCO 2017 dataset [27]. We use ~117k images in train2017 for training and 5k images in val2017 for evaluation. We adopt COCO’s standard evaluation metrics for performance evaluation.

**Implementation Details.** As the proposed IMFA defines a generic paradigm, we mainly conduct experiments with DAB-DETR [29] – a state-of-the-art Transformer-based object detector with open-sourced implementation. We also integrate IMFA with DETR [4], Conditional DETR [35], and Anchor DETR [55], to demonstrate its generality.

A crucial implementation detail involves incorporating skip connections for encoded features between Transformer encoder layers, as motivated by [63] and [65, 66] to facilitate feature semantic alignment.

For IMFA-related hyper-parameters, we set the sampling ratio \( r \) at 20% and the keypoint number \( M \) at 8 by default. Other model-related setups align with their corresponding baselines [4, 29, 35, 55]. We use ImageNet-pretrained [6] ResNet [15] as backbone networks, and conduct training with AdamW optimizer [33]. The total batch size is set to 16 for training. The initial learning rate is \( 1 \times 10^{-5} \) for the backbone networks and \( 1 \times 10^{-4} \) for the Transformer architectures, along with a weight decay of \( 1 \times 10^{-4} \). Models are trained for 50 epochs, with the learning rate decayed at the 40th epoch by 0.1. The same data augmentation scheme used in [4, 29, 35, 55] is adopted.

#### 5.2. Experiment Results

**Compatibility with Transformer-Based Detectors.** We first evaluate the generality of IMFA by integrating it with multiple Transformer-based object detectors. As discussed in Section 1, these methods resort to higher-resolution backbones (denoted with ‘High-Res Feat’) as an alternative, as it
Comparison with State-of-the-Art Detectors. We integrate IMFA with DAB-DETR [29] to benchmark with other state-of-the-art single-stage Transformer-based detectors that utilize high-resolution or multi-scale features. We also include some popular two-stage detectors [42, 47, 60] for a comprehensive comparison. As shown in Table 2, our method can achieve comparable performance with the state-of-the-art methods, but with significantly lower computation. ‘MS’ denotes the use of multi-scale features. ‘SMS’ denotes the use of sparse multi-scale features with our proposed IMFA. ‘DC’ denotes the use of high-resolution features with R50-DC5.

<table>
<thead>
<tr>
<th>Method</th>
<th>#Params</th>
<th>FLOPs</th>
<th>FPS</th>
<th>AP</th>
<th>AP50</th>
<th>AP75</th>
<th>APs</th>
<th>APM</th>
<th>APL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster-RCNN-FPN-R50 [25, 42]</td>
<td>✓</td>
<td>108</td>
<td>42M</td>
<td>180G</td>
<td>42.0</td>
<td>62.1</td>
<td>45.5</td>
<td>26.6</td>
<td>45.5</td>
</tr>
<tr>
<td>TSP-FCOS-FPN-R50 [47]</td>
<td>✓</td>
<td>36</td>
<td>52M</td>
<td>189G</td>
<td>43.1</td>
<td>62.3</td>
<td>47.0</td>
<td>26.6</td>
<td>46.8</td>
</tr>
<tr>
<td>TSP-RCNN-FPN-R50 [47]</td>
<td>✓</td>
<td>36</td>
<td>64M</td>
<td>188G</td>
<td>43.8</td>
<td>63.3</td>
<td>48.3</td>
<td>28.6</td>
<td>46.9</td>
</tr>
<tr>
<td>Sparse-RCNN-FPN-R50 [46]</td>
<td>✓</td>
<td>36</td>
<td>106M</td>
<td>166G</td>
<td>45.0</td>
<td>64.1</td>
<td>48.9</td>
<td>28.0</td>
<td>47.6</td>
</tr>
<tr>
<td>DETR-R50 [4]</td>
<td>✓ ✓</td>
<td>500</td>
<td>41M</td>
<td>187G</td>
<td>43.3</td>
<td>63.1</td>
<td>45.9</td>
<td>22.5</td>
<td>47.3</td>
</tr>
<tr>
<td>Deformable-DETR-R50 [72]</td>
<td>✓ ✓</td>
<td>50</td>
<td>40M</td>
<td>173G</td>
<td>43.8</td>
<td>62.6</td>
<td>47.7</td>
<td>26.4</td>
<td>47.1</td>
</tr>
<tr>
<td>Deformable-DETR-R50 [72] + Iter</td>
<td>✓ ✓</td>
<td>50</td>
<td>41M</td>
<td>173G</td>
<td>45.4</td>
<td>64.7</td>
<td>49.0</td>
<td>26.8</td>
<td>48.3</td>
</tr>
<tr>
<td>Efficient-DETR-R50 [60]</td>
<td>✓ ✓</td>
<td>36</td>
<td>32M</td>
<td>159G</td>
<td>44.2</td>
<td>62.2</td>
<td>48.0</td>
<td>28.4</td>
<td>47.5</td>
</tr>
<tr>
<td>Conditional-DETR-R50 [35]</td>
<td>✓ ✓</td>
<td>50</td>
<td>44M</td>
<td>195G</td>
<td>43.8</td>
<td>64.4</td>
<td>46.7</td>
<td>24.0</td>
<td>47.6</td>
</tr>
<tr>
<td>SMCA-DETR-R50 [11]</td>
<td>✓ ✓</td>
<td>50</td>
<td>40M</td>
<td>152G</td>
<td>43.7</td>
<td>63.6</td>
<td>47.2</td>
<td>24.2</td>
<td>47.0</td>
</tr>
<tr>
<td>YOLOS-Det-S [8]</td>
<td>✓ ✓</td>
<td>150</td>
<td>28M</td>
<td>172G</td>
<td>37.6</td>
<td>57.6</td>
<td>39.2</td>
<td>15.9</td>
<td>40.2</td>
</tr>
<tr>
<td>Anchor-DETR-R50 [55]</td>
<td>✓ ✓</td>
<td>50</td>
<td>37M</td>
<td>172G</td>
<td>44.2</td>
<td>64.7</td>
<td>47.5</td>
<td>24.7</td>
<td>48.2</td>
</tr>
<tr>
<td>DAB-DETR-R50 [29]</td>
<td>✓ ✓</td>
<td>50</td>
<td>44M</td>
<td>202G</td>
<td>44.5</td>
<td>65.1</td>
<td>47.7</td>
<td>25.3</td>
<td>48.2</td>
</tr>
<tr>
<td>SAM-DETR-R50 [66]</td>
<td>✓ ✓</td>
<td>50</td>
<td>58M</td>
<td>210G</td>
<td>43.3</td>
<td>64.4</td>
<td>46.2</td>
<td>25.1</td>
<td>46.9</td>
</tr>
<tr>
<td>SAM-DETR-R50 [66] w/ SMCA [11]</td>
<td>✓ ✓</td>
<td>50</td>
<td>58M</td>
<td>210G</td>
<td>45.0</td>
<td>65.4</td>
<td>47.9</td>
<td>26.2</td>
<td>49.0</td>
</tr>
<tr>
<td>DAB-DETR-R50 [29] + IMFA (Ours)</td>
<td>✓ ✓</td>
<td>50</td>
<td>53M</td>
<td>108G</td>
<td>45.5</td>
<td>65.0</td>
<td>49.3</td>
<td>27.3</td>
<td>48.3</td>
</tr>
</tbody>
</table>

Table 2. Comparison with state-of-the-art object detectors on COCO val 2017. Our proposed method achieves comparable performance with the state-of-the-art methods, but with significantly lower computation. ‘MS’ denotes the use of multi-scale features. ‘SMS’ denotes the use of sparse multi-scale features with our proposed IMFA. ‘DC’ denotes the use of high-resolution features with R50-DC5.

Comparison with State-of-the-Art Detectors. We integrate IMFA with DAB-DETR [29] to benchmark with other state-of-the-art single-stage Transformer-based detectors that utilize high-resolution or multi-scale features. We also include some popular two-stage detectors [42, 47, 60] for a comprehensive comparison. As shown in Table 2, our method can achieve comparable performance with the state-of-the-art methods, but with significantly less computational cost.

Results with Stronger Backbones. As shown in Table 3, when using stronger backbones [31, 32], IMFA still consistently improves detection performance at marginal costs.

5.3. Ablation Study

We conduct ablation studies with the strong baseline DAB-DETR-R50 [15, 29] to validate the effectiveness of our designs. Results are obtained on COCO val 2017.

Effect of IMFA's Design Choices. IMFA introduces two novel designs: i) iterative encoding described in Section 4.2 and Fig. 2 (right), and ii) sparse multi-scale feature sampling and aggregation described in Section 4.3 and Fig. 3. As shown in Table 4, the iterative encoding alone even slightly degrades the baseline’s performance. However, with IMFA’s sparsely sampled multi-scale features, our method significantly improves the detection performance of objects at all scales, especially at smaller scales. This proves that the multi-scale features sampled by IMFA are sparse yet highly effective for object detection.

We also study the three crucial components within the sparse feature sampling and aggregation process in Table 5. Without identifying representative keypoints (random spa-
Table 6. Ablation study on the sampling ratio $r$ of prior detection predictions. Results are obtained on COCO val 2017.

<table>
<thead>
<tr>
<th>$r$</th>
<th>#Params</th>
<th>FLOPs</th>
<th>$AP$</th>
<th>$AP_{50}$</th>
<th>$AP_{75}$</th>
<th>$AP_{S}$</th>
<th>$AP_{M}$</th>
<th>$AP_{L}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>53M</td>
<td>103G</td>
<td>44.2</td>
<td>64.0</td>
<td>47.5</td>
<td>25.9</td>
<td>47.3</td>
<td>60.6</td>
</tr>
<tr>
<td>15%</td>
<td>53M</td>
<td>105G</td>
<td>44.8</td>
<td>64.2</td>
<td>48.2</td>
<td>26.5</td>
<td>47.7</td>
<td>60.1</td>
</tr>
<tr>
<td>20%</td>
<td>53M</td>
<td>108G</td>
<td>45.5</td>
<td>65.0</td>
<td>49.3</td>
<td>27.3</td>
<td>48.3</td>
<td>61.6</td>
</tr>
<tr>
<td>25%</td>
<td>53M</td>
<td>111G</td>
<td>45.3</td>
<td>65.1</td>
<td>49.0</td>
<td>27.9</td>
<td>47.9</td>
<td>61.1</td>
</tr>
<tr>
<td>30%</td>
<td>53M</td>
<td>114G</td>
<td>45.1</td>
<td>64.5</td>
<td>48.9</td>
<td>28.4</td>
<td>48.2</td>
<td>60.2</td>
</tr>
</tbody>
</table>

Table 7. Ablation study on the keypoint number $M$ within each promising region. Results are obtained on COCO val 2017.

<table>
<thead>
<tr>
<th>$M$</th>
<th>#Params</th>
<th>FLOPs</th>
<th>$AP$</th>
<th>$AP_{50}$</th>
<th>$AP_{75}$</th>
<th>$AP_{S}$</th>
<th>$AP_{M}$</th>
<th>$AP_{L}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>53M</td>
<td>101G</td>
<td>43.9</td>
<td>64.3</td>
<td>47.5</td>
<td>25.1</td>
<td>46.9</td>
<td>60.8</td>
</tr>
<tr>
<td>2</td>
<td>53M</td>
<td>102G</td>
<td>45.0</td>
<td>64.7</td>
<td>48.9</td>
<td>26.0</td>
<td>48.3</td>
<td>60.4</td>
</tr>
<tr>
<td>4</td>
<td>53M</td>
<td>104G</td>
<td>45.3</td>
<td>65.0</td>
<td>48.7</td>
<td>27.3</td>
<td>48.1</td>
<td>60.9</td>
</tr>
<tr>
<td>8</td>
<td>53M</td>
<td>108G</td>
<td>45.5</td>
<td>65.0</td>
<td>49.3</td>
<td>27.3</td>
<td>48.3</td>
<td>61.6</td>
</tr>
<tr>
<td>16</td>
<td>53M</td>
<td>117G</td>
<td>45.3</td>
<td>64.7</td>
<td>49.0</td>
<td>26.6</td>
<td>48.5</td>
<td>61.5</td>
</tr>
</tbody>
</table>

Effect of IMFA’s Hyper-Parameters. IMFA introduces two hyper-parameters: the sampling ratio of prior detection predictions and object queries ($r$) as well as the keypoint number in each promising region ($M$). We conduct sensitivity analysis on each of them.

Table 6 shows the effect of different $r$ values when $M$ is fixed at 8. As $r$ increases from 10% to 30%, the average precision ($AP$) first increases then decreases, while the computational cost keeps growing. An interesting trend is that the detection performance of small objects ($AP_S$) goes up with increasing $r$ consistently. We conjecture that small objects rely more on the fine details in high-resolution features, so that they can benefit from increased number of promising regions used for multi-scale feature sampling. However, the overall performance drops when $r$ is too large, which we conjecture is due to the increased difficulty in searching relevant features with overwhelming feature tokens involved. Based on the experimental results, we set the default value for $r$ as 20% in our system.

To study the effect of the number of keypoints $M$, we conduct experiments by fixing $r$ at 20% and report the results in Table 7. We can see a similar trend that the performance improves as $M$ increases but then drops when $M$ becomes too large. Therefore, we set $M$ as 8 by default.

5.4. Extension to Human Pose Estimation

We further apply the proposed IMFA to human pose estimation to verify its generality across different tasks. Concretely, we evaluate the performance on the COCO 2017 human pose estimation benchmark [27]. We adopt PRTR (two-stage variant) [21], a DETR-like human pose estimation method with open-sourced implementation, as our baseline. Please refer to the technical appendix for its full implementation details.

As shown in Table 8, on the task of human pose estimation, IMFA still clearly outperforms its baseline methods at the same input size with only slight extra computation. IMFA even surpasses its higher-resolution baselines at significantly reduced computational costs. The results indicate IMFA’s potential of boosting Transformer-based models on various vision tasks beyond object detection itself.

6. Conclusion

Multi-scale features are beneficial to object detection, but often come with large computational costs. This paper presents Iterative Multi-scale Feature Aggregation (IMFA) as the pioneering generic paradigm for efficient use of multi-scale features in Transformer-based object detectors. It gets the best of both worlds – high accuracy and low computational cost. IMFA identifies and extracts multi-scale features from the most promising and informative locations only and greatly improves detection accuracy on multiple object detectors at marginal additional costs. We expect IMFA will inspire more comprehensive research and applications on Transformer-based object detection.

Limitations. Although IMFA is compatible with many Transformer-based object detectors, it cannot be directly applied to Deformable DETR [72] and its extensions [43, 60]. This is due to undefined deformable operations on non-grid feature maps, which require extensive engineering efforts.

Acknowledgement:

This study is supported under the RIE 2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).
detector based on multi-level feature pyramid network. In 
AAAI, 2019. 1, 2

[71] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krähenbühl. 
Bottom-up object detection by grouping extreme and center 
points. In CVPR, 2019. 2

[72] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, 
and Jifeng Dai. Deformable DETR: Deformable transfor-
ners for end-to-end object detection. In ICLR, 2021. 1, 2, 3, 
4, 7, 8