# Transformers in Small Object Detection: A Benchmark and Survey of State-of-the-Art

AREF MIRI REKAVANDI, The University of Western Australia, Perth, Australia
SHIMA RASHIDI, RMIT University, Melbourne, Australia
FARID BOUSSAID, The University of Western Australia, Perth, Australia
STEPHEN HOEFS, Defence Science and Technology Group, Perth, Australia
EMRE AKBAS, Middle East Technical University, Ankara, Turkey and Helmholtz Center Munich German Research Center for Environmental Health, Neuherberg, Germany
MOHAMMED BENNAMOUN, The University of Western Australia, Perth, Australia

Transformers have rapidly gained popularity in computer vision, especially in the field of object detection. Upon examining the outcomes of state-of-the-art object detection methods, we noticed that transformers consistently outperformed well-established CNN-based detectors in almost every video or image dataset. Small objects have been identified as one of the most challenging object types in detection frameworks due to their low visibility. This article aims to explore the performance benefits offered by such extensive networks and identify potential reasons for their Small Object Detection (SOD) superiority. We aim to investigate potential strategies that could further enhance transformers' performance in SOD. This survey presents a taxonomy of over 60 research studies on developed transformers for the task of SOD, spanning the years 2020 to 2023. These studies encompass a variety of detection applications, including SOD in generic images, aerial images, medical images, active millimeter images, underwater images, and videos. We also compile and present a list of 12 large-scale datasets suitable for SOD that were overlooked in previous studies and compare the performance of the reviewed studies using popular metrics such as mean Average Precision (mAP), Frames Per Second (FPS), and number of parameters. Researchers can keep track of newer studies on our web page, which is available at: https://github.com/arekavandi/Transformer-SOD.

CCS Concepts: • **Computing methodologies → Computer vision**; **Machine learning;**

Additional Key Words and Phrases: Object recognition, small object detection, vision transformers, object localization, deep learning, attention, MS COCO dataset

This research was partially supported by the Australian Research Council (ARC DP210101682, DP210102674) and Defence Science and Technology Group (DSTG) for the project "Low Observer Detection of Small Objects in Maritime Scenes". Dr. Akbas gratefully acknowledges the support of TUBITAK 2219 program of The Scientific and Technological Research Council of Türkiye.

Authors' Contact Information: Aref Miri Rekavandi, The University of Western Australia, Perth, Australia; e-mail: aref.mirirekavandi@uwa.edu.au; Shima Rashidi, RMIT University, Melbourne, Victoria, Australia; e-mail: shima.rashidi7@gmail.com; Farid Boussaid, The University of Western Australia, Perth, Western Australia, Australia; e-mail: farid.boussaid@uwa.edu.au; Stephen Hoefs, Defence Science and Technology Group, Perth, Australia; e-mail: stephen.hoefs@defence.gov.au; Emre Akbas, Middle East Technical University, Ankara, Ankara, Turkey and Helmholtz Center Munich German Research Center for Environmental Health, Neuherberg, Germany; e-mail: emre@ceng.metu.edu.tr; Mohammed Bennamoun, The University of Western Australia, Perth, Western Australia, Australia; e-mail: mohammed.bennamoun@uwa.edu.au.

ACM 0360-0300/2025/09-ART64
https://doi.org/10.1145/3758090

**ACM Reference Format:**
Aref Miri Rekavandi, Shima Rashidi, Farid Boussaid, Stephen Hoefs, Emre Akbas, and Mohammed Bennamoun. 2025. Transformers in Small Object Detection: A Benchmark and Survey of State-of-the-Art. *ACM Comput. Surv.* 58, 3, Article 64 (September 2025), 33 pages. https://doi.org/10.1145/3758090

## 1 Introduction

**Small Object Detection** (**SOD**) has been recognized as a significant challenge for **State-Of-The-Art** (**SOTA**) object detection methods [67]. The term "small object" refers to objects that occupy a small fraction of the input image. For example, the widely used MS COCO dataset [61] defines small objects as those with bounding boxes of 32 × 32 pixels or less in a typical 480 × 640 image (Figure 1). Other datasets have their own definitions, e.g. objects that occupy up to 10% of the image. Small objects are often missed or detected with incorrectly localized bounding boxes, and sometimes with incorrect labels. The main reason for the deficient localization in SOD stems from the limited information provided in the input image or video frame, compounded by the subsequent spatial degradation experienced as they pass through multiple layers in deep networks. Since small objects frequently appear in various application domains, such as pedestrian detection [110], medical image analysis [82], face recognition [20], traffic sign detection [68], traffic light detection [121], ship detection [101], **Synthetic Aperture Radar** (**SAR**)-based object detection [87], it is worth examining the performance of modern deep learning SOD techniques. Object detection and in particular, SOD, has long relied on **Convolutional Neural Network** (**CNN**)-based deep learning models. Several single-stage and two-stage detectors have emerged over time, including **You Only Look Once** (**YOLO**) variants [3, 46, 49, 84–86, 103], **Single Shot multi-box Detector** (**SSD**) [66], RetinaNet [60], **Spatial Pyramid Pooling Network** (**SPP-Net**) [40], Fast R-CNN [35], Faster R-CNN [89], **Region-Based Fully Convolutional Networks** (**R-FCN**) [23], Mask R-CNN [41], **Feature Pyramid Networks** (**FPN**) [59], cascade R-CNN [5], and Libra R-CNN [79]. Various strategies have been used in conjunction with these techniques to improve their detection performance for SOD, with multi-scale feature learning being the most commonly used approach. In our previous work, we surveyed numerous strategies employed in deep learning to enhance the performance of SOD in optical images and videos up to the year 2022 [88]. We showed that beyond the adaptation of newer deep learning structures such as transformers, prevalent approaches include data augmentation, super-resolution, multi-scale feature learning, context learning, attention-based learning, region proposal, loss function regularization, leveraging auxiliary tasks, and spatiotemporal feature aggregation. Additionally, we observed that transformers are among the leading methods in localizing small objects across most datasets. However, given that [88] predominantly evaluated over 160 papers focusing on CNN-based networks, an in-depth exploration of transformer-centric methods was not undertaken. Given the rapid advancements and ongoing exploration in the field, there is a timely opportunity to investigate the latest transformer models designed for SOD. In this article, we aim to provide a comprehensive analysis of the factors driving the strong performance of transformers in SOD and to highlight how these approaches differ from those used in generic object detection.

Since 2017, numerous review articles have been published in the field. A comprehensive discussion and listing of these reviews can be found in our previous survey [88]. Another recent survey [17] primarily focuses on CNN-based techniques. The narrative of this current survey stands distinct from its predecessors. Our focus in this article narrows down specifically to transformers – an aspect not explored previously – positioning them as the dominant network architecture for image and video SOD. This entails a unique taxonomy tailored to this innovative architecture, consciously sidelining CNN-based methods. Given that this topic is relatively new, our survey prioritizes

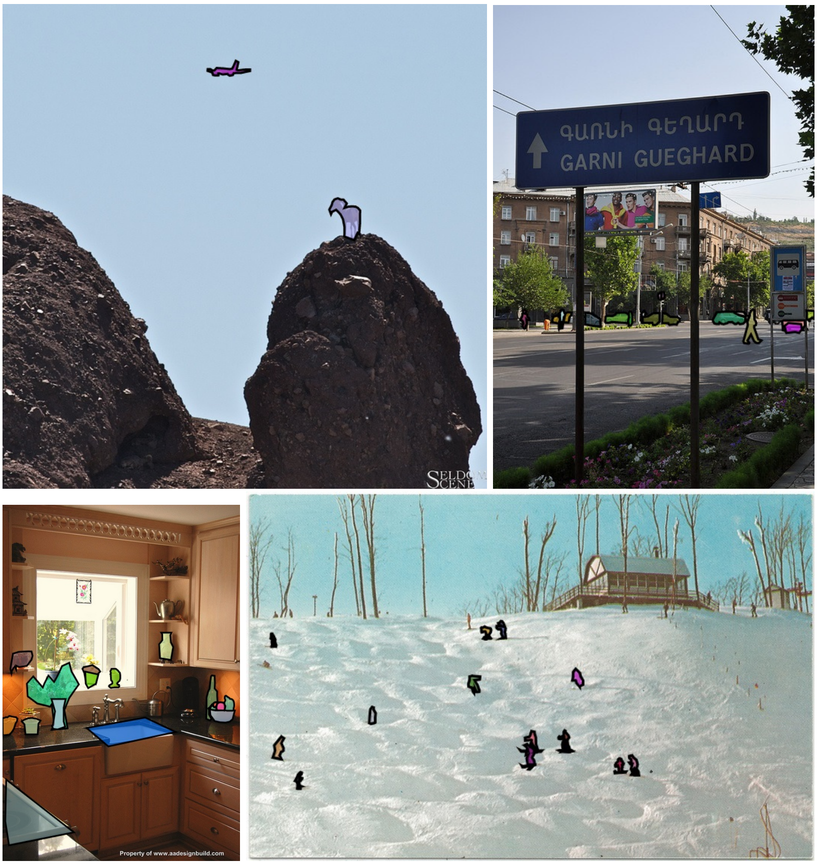

Fig. 1. Examples of small size objects from MS COCO dataset [61]. The objects are highlighted with color segments.

works published primarily post-2021, while including selected earlier studies from 2020 and 2021 for completeness. We also highlight newly introduced datasets relevant to SOD across various application domains. The studies included were identified through keyword searches (e.g., “small object detection”, “tiny object detection”, “transformers for SOD”) using platforms such as Google Scholar, IEEE Xplore, arXiv as well as through reference chaining from influential works. Priority was given to papers that directly addressed SOD or provided quantitative performance breakdowns for small objects. Studies that mentioned small objects but lacked SOD-specific methodological focus or yielded subpar results were excluded. In this survey, we assume the reader is already familiar with generic object detection techniques, their architectures, and relevant performance measures. If the reader requires foundational insight into these areas, we refer the reader to our previous work [88]. In summary, the key contributions of this survey are as follows:

- We provide a comprehensive review of recent transformer-based SOD methods (over 60 research studies) across both image and video domains. A novel taxonomy is proposed to categorize these approaches, with detailed descriptions and critical insights into their strengths and limitations.
- We present 12 newly published or previously underrepresented SOD datasets, which were not covered in earlier surveys, thereby expanding the dataset landscape available to researchers.

Table 1. A List of Terminologies Used in This Article with Their Meanings

| Full Term | Description |
|---|---|
| Encoder | Encoder in transformers consists of multiple layers of self-attention modules and feed-forward neural networks to extract local and global semantic information from the input data. |
| Decoder | Decoder module is responsible to generate the output (either sequence or independent) based on the concept of self and cross attention applied to the object queries and encoder's output. |
| Token | Token refers to the most basic unit of data input into the transformers. It can be image pixels, patches, or video clips. |
| Multi-Head Attention | Multi-Head Attention is a mechanism in transformers that enhances the learning capacity and representational power of self-attention. It divides the input into multiple subspaces and performs attention computations independently on each subspace, known as attention heads. |
| Spatial Attention | Spatial attention in transformers refers to a type of attention mechanism that attends to the spatial positions of tokens within a sequence. It allows the model to focus on the relative positions of tokens and capture spatial relationships. |
| Channel Attention | Channel attention in transformers refers to an attention mechanism that operates across different channels or feature dimensions of the input. It allows the model to dynamically adjust the importance of different channels, enhancing the representation and modeling of channel-specific information in tasks. |
| Object Query | It refers to a learned vector representation that is used to query and attend to specific objects or entities within a scene. |
| Positional Embedding | It refers to a learned representation that encodes the positional information of tokens in an input sequence, enabling the model to capture sequential dependencies. |

— We provide a comparative analysis of existing models across general-purpose and domain-specific applications, including aerial imagery, medical imaging, underwater imagery, active millimeter wave imaging, and video-based detection tasks.
— We discuss the current challenges and limitations in transformer-based SOD and outline promising future research directions to advance the field.

The structure of this article is as follows. Section 2 introduces foundational transformer architectures–detailing key components such as the encoder and decoder–and presents two pioneering transformer-based models, DETR and ViT-FRCNN. Section 3 shifts focus to transformer-based methods specifically engineered for SOD, incorporating specialized designs and techniques to enhance performance on small-scale targets. We then present a comprehensive taxonomy of these approaches and examine each category in depth. Section 4 showcases the different datasets used for SOD and evaluates them across a range of applications. In Section 5, we analyze and contrast these outcomes with earlier results derived from CNN networks. Finally, Section 6 presents our concluding remarks.

## 2 Background

Transformers represent a category of neural networks renowned for their prowess in natural language processing tasks [102]. They stand out for their capacity to grasp contextual relationships and understanding through the analysis of sequential data, such as text or time-series data. Through mechanisms like self-attention, transformers excel in capturing interdependencies and correlations among elements within input sequences. The transformer model comprises two primary modules: the encoder and the decoder. Visual representations of the processing blocks within each module are depicted in Figure 2. Table 1 serves as a reference for readers unfamiliar with the terminology commonly used in transformers for computer vision. In the context of Object Detection, the encoder module processes input tokens, which can be image patches or video clips, employing various feature embedding approaches. This includes leveraging pre-trained CNNs to extract suitable representations. The positional encoding block enriches the feature representations of each token with positional information, a technique that has shown significant performance enhancements across various applications. Subsequently, the encoded representations undergo processing in a Multi-Head Attention block, parameterized with three main matrices, namely $\mathbf{W}_q \in \mathbf{R}^{d_q \times d}$,

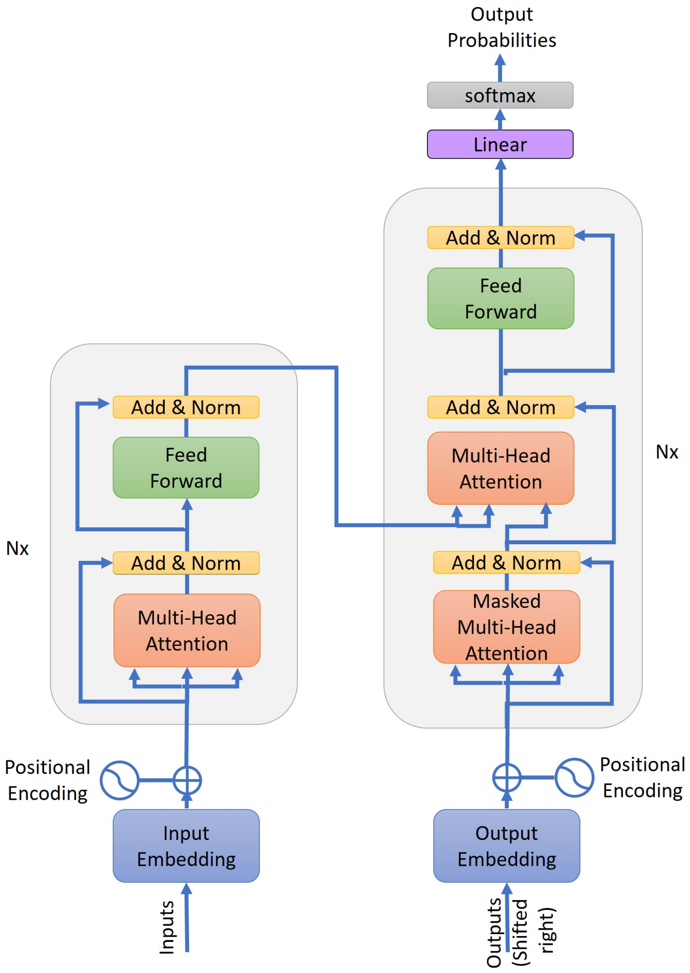

Fig. 2. Transformer architecture containing encoder (left module) and decoder (right module) used in sequence to sequence translation (figure from [102]).

$\mathbf{W}_k \in \mathbf{R}^{d_k \times d}$, and $\mathbf{W}_v \in \mathbf{R}^{d_v \times d}$ to obtain query, key, and value vectors, denoted as $\mathbf{q}$, $\mathbf{k}$, $\mathbf{v}$, respectively. In other words,

$$\mathbf{q}_i = \mathbf{W}_q \mathbf{x}_i, \quad \mathbf{k}_i = \mathbf{W}_k \mathbf{x}_i, \quad \mathbf{v}_i = \mathbf{W}_v \mathbf{x}_i, \quad i = 1, \ldots, T, \tag{1}$$

where $T$ is the total number of tokens with each token being represented by $\mathbf{x}$. The Multi-Head Attention block produces the output:

$$\text{MH Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O. \tag{2}$$

where $\mathbf{W}^O \in \mathbf{R}^{hd_v \times d}$, $d_k = d_q$, and

$$\text{head}_h = \text{Attention}(\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h) = \text{Softmax}\left(\frac{\mathbf{K}_h^\top \mathbf{Q}_h}{\sqrt{d_k}}\right)\mathbf{V}_h^\top. \tag{3}$$

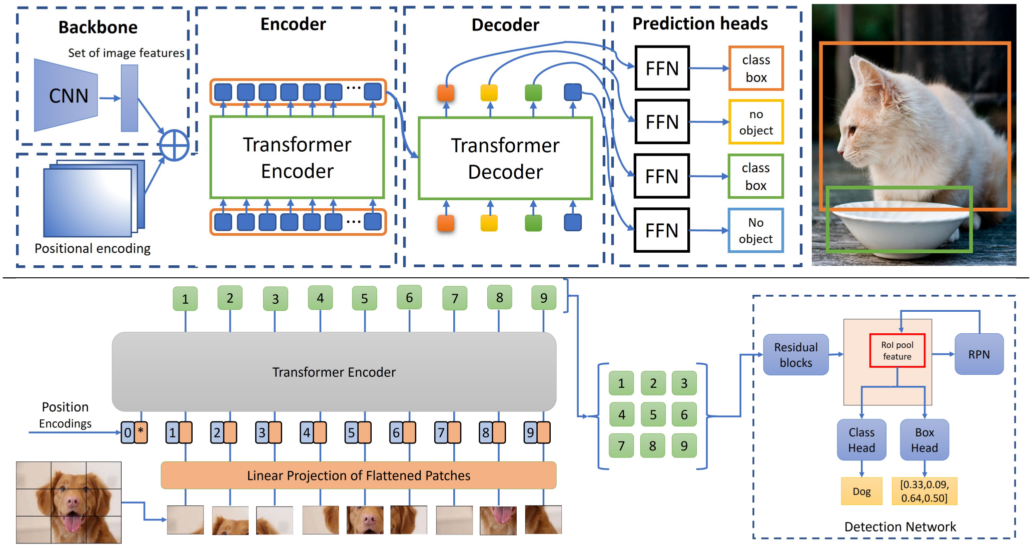

Fig. 3. Top: DETR (figure from [7]). Bottom: ViT-FRCNN (figure from [2]).

These outputs are then combined with a skip connection and a normalization block. Each vector then undergoes independent processing through a fully connected layer, where an activation function introduces non-linearity into the network. The parameters of this block are shared across all vectors. This process repeats $N$ times, with $N$ corresponding to the number of layers in the deep network. In the decoder module, a similar process occurs using the vectors generated in the encoder. Ultimately, the system computes the output probabilities for potential output classes, with attention achieved through the dot product operation between the key and query matrices. An alternative representation for the transformer is also given by

$$\text{MH Attention}^i = \sum_h \mathbf{W}_h^O \left[ \sum_{k=1}^{T} A_{hik} \mathbf{W}_v \mathbf{x}_k \right], i = 1, \ldots, T, \tag{4}$$

where $\mathbf{W}_h^O$ is a submatrix of $\mathbf{W}^O$ that corresponds to $h$th head, and $A_{hik}$ is the attention weight in $h$th head, which is the element in $i$th row (corresponds to $i$th query) and $k$th column (corresponds to $k$th key) of the matrix: Softmax$(\frac{\mathbf{K}_h^\top \mathbf{Q}_h}{\sqrt{d_k}})$.

Dosovitskiy *et al.* were the first to utilize the architecture of transformers in computer vision tasks, including image recognition [29]. The remarkable performance exhibited by transformers in various vision tasks has paved the way for their application in the domain of object detection research. Two pioneering works in this area are the **DEtection TRansformer** (**DETR**) [7] (Figure 3, Top) and ViT-FRCNN [2] (Figure 3, Bottom).

DETR aimed to reduce the reliance on CNN-based techniques during post-processing by employing a set-based global loss. This particular loss function aids in the collapse of near-duplicate predictions through bipartite matching, ensuring each prediction is uniquely paired with its matching ground truth bounding boxes. As an end-to-end model, DETR benefits from global computation and perfect memory, making it suitable for handling long sequences generated from videos/images. The bipartite matching loss utilized in DETR is defined as follows:

$$\hat{s} = \arg\min_{s \in \mathcal{S}} \sum_{i}^{N} \mathcal{L}_{match}(y_i, \hat{y}_{s(i)}), \tag{5}$$

where $\mathcal{L}_{match}(y_i, \hat{y}_{s(i)})$ measures the pair-wise matching cost between ground truth box $y_i$ with size $N$ and the prediction with index $s(i)$ where $s$ is a specific order of predicted bounding boxes. In this formulation, $N$ is the largest possible number of objects within an image. In the case of fewer objects in predictions and ground-truth, $y$ and $\hat{y}$ will be padded with $\emptyset$ (indicating no object). Consequently, this loss function considers all possible matching policies between predictions and ground truth, selecting the one that yields the minimum loss value. The optimal pairing can be efficiently computed using the Hungarian algorithm, as demonstrated in [95]. DETR used a CNN backbone to extract compact feature representations and an encoder-decoder transformer with a feed-forward network to produce the final predictions (see Figure 3, Top). In contrast, ViT-FRCNN uses the **Vision Transformer** (**ViT**) [29] for object detection and demonstrates that pre-trained ViT on large-scale datasets enhances the detection performance through rapid fine-tuning. While ViT-FRCNN, like DETR, incorporates CNN-based networks in its pipeline, specifically in the detection head, it diverges from DETR by using the Transformer (encoder only) to encode visual attributes. Additionally, a conventional **Region Proposal Network** (**RPN**) [89] is used for generating detections (illustrated in Figure 3, Bottom). Both DETR and ViT-FRCNN have shown subpar results in the detection and classification of small objects. ViT-FRCNN even exhibited worse results when increasing the token size of the input image. The best outcomes were achieved when the token size was set to 16 × 16, and all intermediate transformer states were concatenated with the final transformed layer. Additionally, both detectors rely on CNNs at different stages, in DETR as the backbone for feature extraction and in ViT-FRCNN for the detection head. To improve the results of SOD, it is crucial to retain the image patches as small as possible to preserve spatial resolution, which consequently increases the computational costs. To address these limitations and challenges, further research has been conducted, which will be discussed in detail in Section 3 and in particular Section 3.4.

## 3 Transformers for Small Object Detection

This section focuses on transformer-based methods specifically designed to address the challenges of SOD. Building upon the general transformer architectures introduced in Section 2, we now shift attention to approaches that incorporate specialized designs and techniques to improve performance in detecting small-scale targets. A taxonomy of small object detectors is shown in Figure 4. We show that existing detectors based on novel transformers can be analyzed through one or more of the following perspectives: object representation, fast attention for high-resolution or multi-scale feature maps, fully transformer-based detection, architecture and block modification, auxiliary techniques, improved feature representation, and spatio-temporal information. Each of these categories is examined in detail in the following subsections, with a focus on how they uniquely contribute to overcoming the limitations of detecting small objects.

### 3.1 Object Representation

Various object representation techniques have been adopted in object detection techniques. The object of interest can be represented by rectangular boxes [35], points such as center points [129] and point sets [119], probabilistic objects [105], and keypoints [48]. Each object representation technique has its own strengths and weaknesses, with respect to the need for annotation formats and small object representation. The pursuit of finding the optimal representation technique, while keeping all the strengths of the existing representations, began with RelationNet++ [19]. This

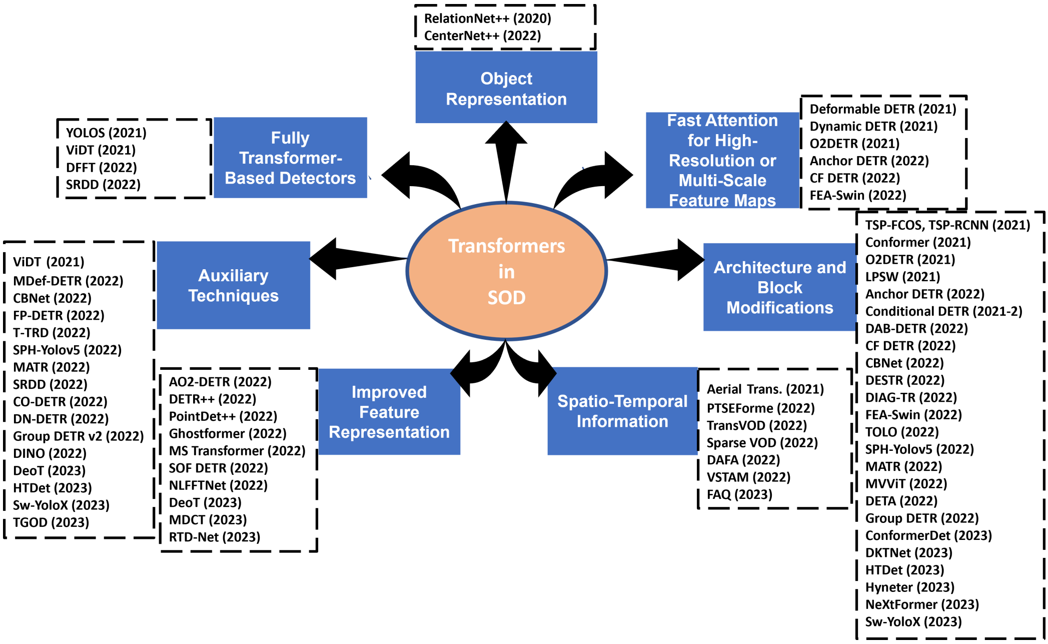

Fig. 4. Taxonomy of SOD using transformers and popular object detection methods in each category.

approach bridges various heterogeneous visual representations and combines their strengths via a module called **Bridging Visual Representations** (**BVR**). BVR operates efficiently without disrupting the overall inference process employed by the main representations, leveraging novel techniques of key sampling and shared location embedding. More importantly, BVR relies on an attention module that designates one representation form as the "master representation" (or query), while the other representations are designated as "auxiliary" representations (or keys). The BVR block is shown in Figure 5, where it enhances the feature representation of the anchor box by seamlessly integrating center and corner points (keys) into the anchor-based (query) object detection methodology. Different object representations are also shown in Figure 5. CenterNet++ [30] was proposed as a novel bottom-up approach. Instead of estimating all the object's parameters at once, CenterNet++ strategically identifies individual components of the object separately, i.e., top-left, bottom-left, and center keypoints. Then, post-processing methodologies are adopted to cluster points associated with the same objects. This technique has demonstrated a superior recall rate in SOD compared to top-down approaches that estimate entire objects as a whole.

### 3.2 Fast Attention for High-Resolution or Multi-Scale Feature Maps

Previous research has shown that maintaining a high resolution of feature maps is a necessary step for maintaining high performance in SOD. Transformers, inherently exhibit a notably higher complexity compared to CNNs due to their quadratic increase in complexity with respect to the number of tokens (e.g., pixel numbers). This complexity emerges from the requirement of pairwise correlation computation across all tokens. Consequently, both training and inference times exceed expectations, rendering the detector inapplicable for SOD in high-resolution images and videos.

In their work on Deformable DETR, Zhu *et al.* [130] addressed this issue that had been observed in DETR for the first time. They proposed attending to only a small set of key sampling points around a reference, significantly reducing the complexity. By adopting this strategy, they effectively preserved

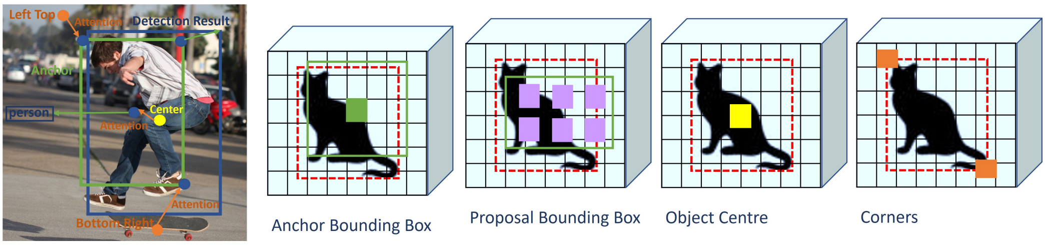

Fig. 5. BVR uses different representations. i.e., corner and center points to enhance features for anchor-based detection (left figure). Object representations are shown for another image (cat) where red dashes show the ground truth (figure from [19]).

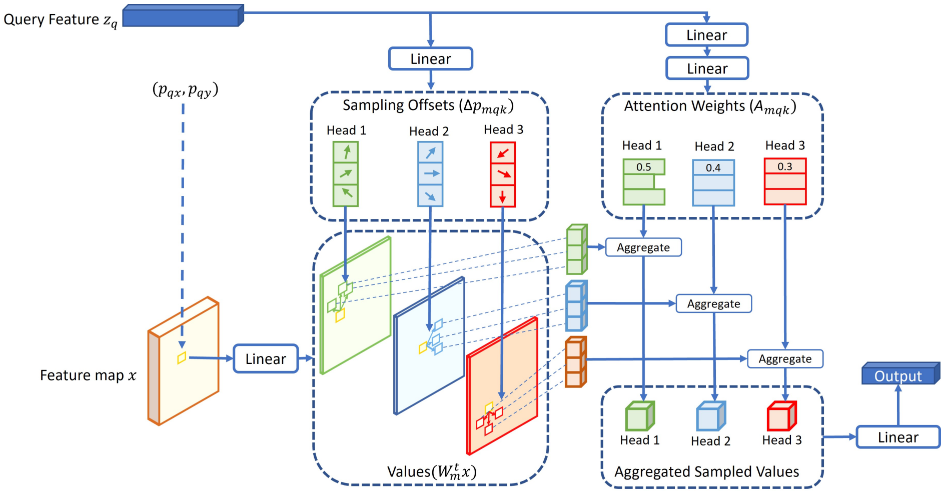

Fig. 6. The block diagram for the deformable attention module. $\mathbf{z}_q$ is the content feature of the query, $\mathbf{x}$ is the feature map and $\mathbf{p}_q$ is the reference point in 2-D grid. In short, the deformable attention module only attends to a small set of key sampling points around the reference point (different in each head). This significantly reduces the complexity and further improves the convergence (figure from [130]).

spatial resolution through the use of multi-scale deformable attention modules. Remarkably, this method eliminated the necessity for FPN, thereby greatly enhancing the detection and recognition of small objects. The $i$th output of a multi-head attention module in Deformable attention is given by

$$\text{MH Attention}^i = \sum_h \mathbf{W}_h^O \left[ \sum_{k=1}^{K} A_{hik} \mathbf{W}_v \mathbf{x}_k (\mathbf{p}_i + \Delta \mathbf{p}_{hik}) \right], \tag{6}$$

where $i = 1, \ldots, T$ and $\mathbf{p}_i$ is the reference point of the query and $\Delta \mathbf{p}_{hik}$ is the sampling offset (in 2D) in $h$th head with K samplings (K<<T=HW). Figure 6 illustrates the computation process within its multi-head attention module. Deformable DETR benefits from both its encoder and decoder modules, with the complexity order within the encoder being $\mathcal{O}(HWC^2)$ where $H$ and $W$ are the height and width of the input feature map and $C$ is the number of channels. In contrast for the DETR encoder, the order of complexity is $\mathcal{O}(H^2W^2C)$, displaying a quadratic increase as $H$ and $W$ increase in size. Deformable attention has played a prominent role in various other detectors, e.g., in T-TRD [53]. Subsequently, Dynamic DETR was proposed in [25], featuring a dynamic encoder and

a dynamic decoder that harness feature pyramids from low to high-resolution representations, resulting in efficient coarse-to-fine object detection and faster convergence. The dynamic encoder can be viewed as a sequentially decomposed approximation of full self-attention, dynamically adjusting attention mechanisms based on scale, spatial importance, and representation. Both Deformable DETR and Dynamic DETR make use of deformable convolution for feature extraction. In a distinct approach, O$^2$DETR [71] demonstrated that the global reasoning offered by a self-attention module is actually not essential for aerial images, where objects are usually densely packed in the same image area. Hence, replacing attention modules with local convolutions coupled with the integration of multi-scale feature maps was proven to improve the detection performance in the context of oriented object detection. The authors in [108] proposed the concept of **Row-Column Decoupled Attention** (**RCDA**), decomposing the 2D attention of key features into two simpler forms: 1D row-wise and column-wise attentions. In the case of CF-DETR [6], an alternative approach to FPN was proposed whereby C5 features were replaced with encoder features at level 5 (E5), resulting in improved object presentation. This innovation was named **Transformer Enhanced FPN** (**TEF**) module. In another study, Xu *et al.* [114] developed a weighted **Bidirectional Feature Pyramid Network** (**BiFPN**) through the integration of skip connection operations with the Swin transformer. This approach effectively preserved information pertinent to small objects.

### 3.3 Fully Transformer-Based Detectors

The advent of transformers and their outstanding performance in many complex tasks in computer vision has gradually motivated researchers to shift from CNN-based or mixed systems to fully transformer-based vision systems. This line of work started with the application of a transformer-only architecture to the image recognition task, known as ViT, proposed in [29]. In [94], ViDT extended the YOLOS model [32] (the first fully transformer-based detector) to develop the first efficient detector suitable for SOD. In ViDT, the ResNet used in DETR for feature extraction is replaced with various ViT variants, such as Swin Transformer [69], ViTDet [56], and DeiT [100], along with the **Reconfigured Attention Module** (**RAM**). The RAM is capable of handling [PATCH] × [PATCH], [DET] × [PATCH], and [DET] × [DET] attentions. These cross and self-attention modules are necessary because, similar to YOLOS, ViDT appends [DET] and [PATCH] tokens in the input. ViDT only utilizes a transformer decoder as its neck to exploit multi-scale features generated at each stage of its body step. Figure 7 illustrates the general structure of ViDT and highlights its differences from DETR and YOLOS.

Recognizing that the decoder module is the main source of inefficiency in transformer-based object detection, the **Decoder-Free Fully Transformer** (**DFFT**) [11] leverages two encoders: **Scale-Aggregated Encoder** (**SAE**) and **Task-Aligned Encoder** (**TAE**), to maintain high accuracy. SAE aggregates the multi-scale features (four scales) into a single feature map, while TAE aligns the single feature map for object type and position classification and regression. Multi-scale feature extraction with strong semantics is performed using a **Detection-Oriented Transformer** (**DOT**) backbone.

In **Sparse RoI-based deformable DETR** (**SRDD**) [131], the authors proposed a lightweight transformer with a scoring system to ultimately remove redundant tokens in the encoder. This is achieved using RoI-based detection in an end-to-end learning scheme.

### 3.4 Architecture and Block Modifications

DETR, the first end-to-end object detection method, struggles with extended converge times during training and performs poorly on small objects. Several research works have addressed these issues to improve SOD performance. One notable contribution comes from Sun et al. [97], who, drawing inspiration from FCOS [99] (a fully convolutional single-stage detector) and Faster RCNN, proposed

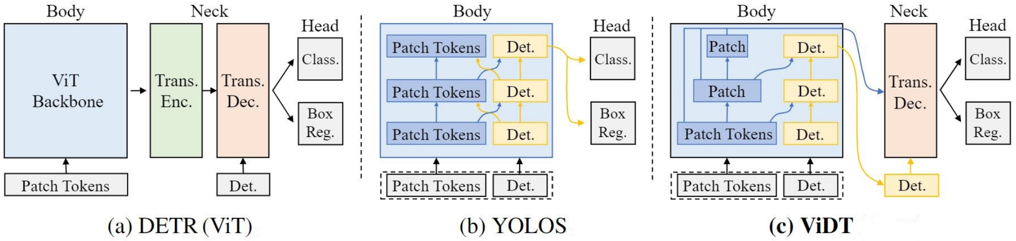

Fig. 7. ViDT (a) mixes DETR (with ViT backbone or other fully transformer-based backbones) (b) with YOLOS architecture (c) in a multi-scale feature learning pipeline to achieve SOTA results (figure from [94]).

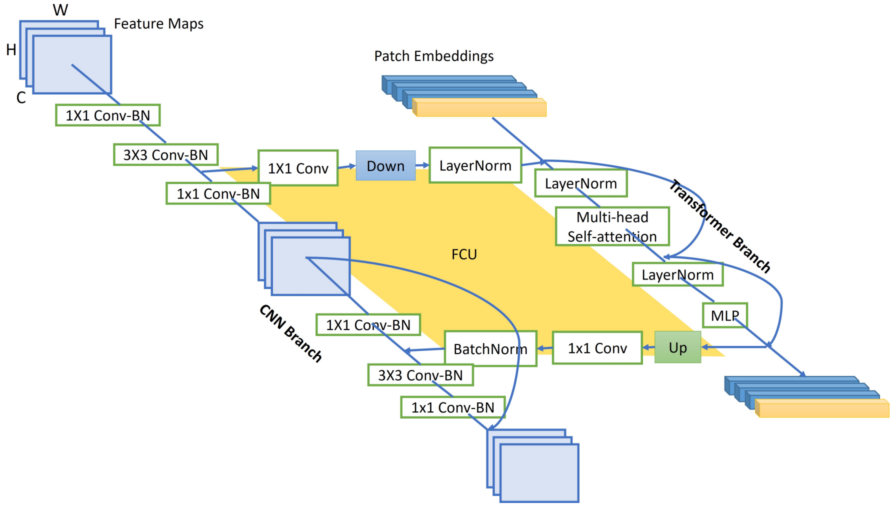

Fig. 8. Conformer architecture which leverages both local features provided by CNNs and global features provided by transformers in FCU (figure from [80]).

two encoder-only DETR variants with feature pyramids called TSP-FCOS and TSP-RCNN. This was accomplished by eliminating cross-attention modules from the decoder. Their findings demonstrated that cross-attention in the decoder and the instability of the Hungarian loss were the main reasons for the late convergence in DETR. This insight led them to discard the decoder and introduce a new bipartite matching technique in these new variants, i.e., TSP-FCOS and TSP-RCNN.

In a combined approach using CNNs and transformers, Peng et al. [80, 81] proposed a hybrid network structure called “Conformer”. This structure fuses the local feature representation provided by CNNs with the global feature representation provided by transformers at varying resolutions (see Figure 8). This was achieved through **Feature Coupling Units** (**FCUs**), with experimental results demonstrating its effectiveness compared to ResNet50, ResNet101, DeiT, and other models. A similar hybrid technique combining CNNs and transformers was proposed in [70]. Recognizing the importance of local perception and long-range correlations, Xu et al. [115] added a **Local Perception Block** (**LPB**) to the Swin Transformer block in the Swin Transformer. This new backbone, called the Local Perception Swin Transformer (LPSW), improved the detection of small-size objects in aerial images significantly. DIAG-TR [116] introduced a **Global-Local Feature Interweaving** (**GLFI**) module in the encoder to adaptively and hierarchically embed local features

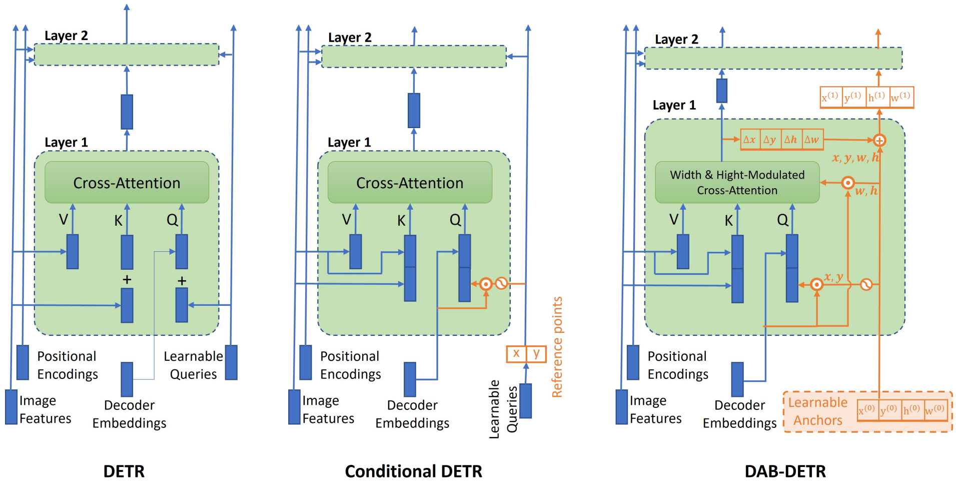

Fig. 9. DAB-DETR improves conditional DETR and utilizes dynamic anchor boxes to sequentially provide better reference query points and anchor sizes (figure from [64]).

into global representations. This technique counterbalances for the scale discrepancies of small objects. Furthermore, learnable anchor box coordinates were added to the content queries in the transformer decoder, providing an inductive bias. In a recent study, Chen et al. [8] proposed the Hybrid network Transformer (Hyneter), which extends the range of local information by embedding convolutions into the transformer blocks. This improvement led to enhanced detection results on the MS COCO dataset. Similar hybrid approaches have been adopted in [26]. In another study [118], the authors proposed a new backbone called NeXtFormer, which combines CNN and transformer to boost the local details and features of small objects, while also providing a global receptive field.

Among various methods, O$^2$DETR [71] substituted the attention mechanism in transformers with depthwise separable convolution. This change not only decreased memory usage and computational costs associated with multi-scale features but also potentially enhanced the detection accuracy in aerial photographs.

Questioning the object queries used in previous works, Wang et al. [108] proposed Anchor DETR, which used anchor points for object queries. These anchor points enhance the interpretability of the target query locations. The use of multiple patterns for each anchor point improves the detection of multiple objects in one region. In contrast, Conditional DETR [74] emphasizes on the conditional spatial queries derived from the decoder content leading to spatial attention predictions. A subsequent version, Conditional DETR v2 [14], enhanced the architecture by reformulating the object query into the form of a box query. This modification involves embedding a reference point and transforming boxes with respect to the reference point. In subsequent works, DAB-DETR [64] further improved on the idea of query design by using dynamically adjustable anchor boxes. These anchor boxes serve as both reference query points and anchor dimensions (see Figure 9).

In another work [6], the authors observed that while the **mean average precision** (**mAP**) of small objects in DETR is not competitive with SOTA techniques, its performance for small **intersection-over-union** (**IoU**) thresholds is surprisingly better than its competitors. This indicates that while DETR provides strong perception abilities, it requires fine-tuning to achieve better localization accuracy. As a solution, the **Coarse-to-Fine Detection Transformer** (**CF-DETR**) has been proposed to perform this refinement through **Adaptive Scale Fusion** (**ASF**) and **Local Cross-Attention**

(**LCA**) modules in the decoder layer. In [42] the authors contend that the suboptimal performance of transformer-based detectors can be attributed to factors such as using a singular cross-attention module for both categorization and regression, inadequate initialization for content queries, and the absence of leveraging prior knowledge in the self-attention module. To address these concerns, they proposed Detection Split Transformer (DESTR). This model splits cross-attention into two branches, one for classification and one for regression. Moreover, DESTR uses a mini-detector to ensure proper content query initialization in the decoder and enhances the self-attention module. Another research [114] introduced FEA-Swin, which leverages advanced foreground enhancement attention in the Swin Transformer framework to integrate context information into the original backbone. This was motivated by the fact that Swin Transformer does not adequately handle dense object detection due to missing connections between adjacent objects. Therefore, foreground enhancement highlights the objects for further correlation analysis. TOLO [112] is one of the recent works aiming to bring inductive bias (using CNN) to the transformer architecture through a simple neck module. This module combines features from different layers to incorporate high-resolution and high-semantic properties. Multiple light transformer heads were designed to detect objects at different scales. In a different approach, instead of modifying the modules in each architecture, CBNet, proposed by Liang et al. [57], groups multiple identical backbones that are connected through composite connections.

In the **Multi-Source Aggregation Transformer** (**MATR**) [96], the cross-attention module of the transformer is used to leverage other support images of the same object from different views. A similar approach is adopted in [45], where the **Multi-View Vision Transformer** (**MVViT**) framework combines information from multiple views, including the target view, to improve the detection performance when objects are not visible in a single view.

Other works prefer to adhere to the YOLO family architecture. For instance, SPH-Yolov5 [37] adds a new branch in the shallower layers of the Yolov5 network to fuse features for improved small object localization. It also incorporates for the first time the Swin Transformer prediction head in the Yolov5 pipeline.

In [77], the authors argue that the Hungarian loss's direct one-to-one bounding box matching approach might not always be advantageous. They demonstrate that employing a one-to-many assignment strategy and utilizing the **NMS** (**Non-Maximum Suppression**) module leads to better detection results. Echoing this perspective, Group DETR [12] implements K groups of object queries with one-to-one label assignment, leading to K positive object queries for each ground-truth object to enhance performance.

A **Dual-Key Transformer Network** (**DKTNet**) is proposed in [113], where two keys are used–one key along with the **Q** stream and another key along with the **V** stream. This enhances the coherence between **Q** and **V**, leading to improved learning. Additionally, channel attention is computed instead of spatial attention, and 1D convolution is used to accelerate the process.

### 3.5 Auxiliary Techniques

Experimental results have demonstrated that auxiliary techniques or tasks, when combined with the main task, can enhance performance. In the context of transformers, several techniques have been adopted, including: **(i)** Auxiliary Decoding/Encoding Loss: This refers to the approach where feed-forward networks designed for bounding box regression and object classification are connected to separate decoding layers. Hence individual losses at different scales are combined to train the models leading to better detection results. This technique or its variants have been used in ViDT [94], MDef-DETR [72], CBNet [57], SRDD [131]. **(ii)** Iterative Box Refinement: In this method, the bounding boxes within each decoding layer are refined based on the predictions from the previous layers. This feedback mechanism progressively improves detection accuracy. This technique has been used in ViDT [94]. **(iii)** Top-Down Supervision: This approach leverages human-understandable semantics to aid in the intricate task of detecting small or class-agnostic objects, e.g., aligned image-text pairs in MDef-DETR [72], or a text-guided object detector in TGOD [92]. **(iv)** Pre-training: This involves training on large-scale datasets followed by specific fine-tuning for the detection task. This technique has been used in CBNet V2-TTA [4], FP-DETR [107], T-TRD [53], SPH-Yolov5 [37], MATR [96], and extensively in Group DETR v2 [13]. **(v)** Data Augmentation: This technique enriches the detection dataset by applying various augmentation techniques, such as rotation, flipping, zooming in and out, cropping, translation, adding noise, and so on. Data augmentation is a commonly used approach to address various imbalance problems [76], e.g., imbalance in object size, within deep learning datasets. Data augmentation can be seen as an indirect approach to minimize the gap between train and test sets [83]. Several methods used augmentation in their detection task including T-TRD [53], SPH-Yolov5 [37], MATR [96], NLFFTNet [122], DeoT [28], HTDet [9], and Sw-YoloX [26]. (**vi**) One-to-Many Label Assignment: The one-to-one matching in DETR can result in poor discriminative features within the encoder. Hence, one-to-many assignments in other methods, e.g., Faster-RCNN, RetinaNet, and FCOS have been used as auxiliary heads in some studies such as CO-DETR [132]. (**vii**) Denoising Training: This technique aims to boost the convergence speed of the decoder in DETR, which often faces an unstable convergence due to bipartite matching. In denoising training, the decoder is fed with noisy ground-truth labels and boxes into the decoder. The model is then trained to reconstruct the original ground truth (guided by an auxiliary loss). Implementations like DINO [125] and DN-DETR [50] have demonstrated the effectiveness of this technique in enhancing the decoder's stability.

### 3.6 Improved Feature Representation

Although current object detectors excel in a wide range of applications for regular-size or large objects, certain use-cases necessitate specialized feature representations for improved SOD. For instance, when it comes to detecting oriented objects in aerial imagery, any object rotation can drastically alter the feature representation due to increased background noise or clutter in the scene (region proposal). To address this, Dai *et al.* [24] ] proposed AO2-DETR, a method designed to be robust to arbitrary object rotations. This is achieved through three key components: **(i)** the generation of oriented proposals, **(ii)** a refinement module of the oriented proposal which extracts rotational-invariant features, and **(iii)** a rotation-aware set matching loss. These modules help to negate the effects of any rotations of the objects. In a related approach, DETR++[123] uses multiple Bi-Directional Feature Pyramid layers (BiFPN) that are applied in a bottom-up fashion to feature maps from C3, C4, and C5. Then, only one scale, which is representative of features at all scales, is selected to be fed into DETR framework for detection. For some specific applications, such as plant safety monitoring, where objects of interest are usually related to human workers, leveraging this contextual information can greatly improve feature representation. PointDet++ [98] capitalizes on this by incorporating human pose estimation techniques, integrating local and global features to enhance SOD performance. Another crucial element that impacts feature quality is the backbone network and its ability to extract both semantic and high-resolution features. GhostNet introduced in [54] offers a streamlined and more efficient network that delivers high-quality, multi-scale features to the transformer. Their Ghost module in this network partially generates the output feature map, with the remainder being recovered using simple linear operations. This is a key step to alleviate the complexity of the backbone networks. In the context of medical image analysis, MS Transformer [93] used a self-supervised learning approach to perform a random mask on the input image, which aids in reconstructing richer features, that are less sensitive to the noise. In conjunction with a hierarchical transformer, this approach outperforms DETR frameworks with various backbones. The **Small Object Favoring DETR** (**SOF-DETR**) [31] specifically favors the detection of small objects by merging convolutional features from layers 3 and 4 in a normalized inductive bias module prior to input into the DETR-Transformer. NLFFTNet [122] addresses the limitation of only considering local interactions in current fusion techniques by introducing a nonlocal feature-fused transformer convolutional network, capturing long-distance semantic relationships between different feature layers. DeoT [28] merges an encoder-only transformer with a novel feature pyramid fusion module. This fusion is enhanced by the use of channel and spatial attention in the **Channel Refinement Module** (**CRM**) and **Spatial Refinement Module** (**SRM**), enabling the extraction of richer features. The authors in HTDet [9] proposed a fine-grained FPN to cumulatively fuse low-level and high-level features for better object detection. Meanwhile, in MDCT [10] the author proposed a **Multi-kernel Dilated Convolution** (**MDC**) module to improve the performance of small object-related feature extraction using both the ontology and adjacent spatial features of small objects. The proposed module leverages depth-wise separable convolution to reduce the computational cost. Lastly, in [120], a feature fusion module paired with a lightweight backbone is engineered to enhance the visual features of small objects by broadening the receptive field. The hybrid attention module in RTD-Net [120] empowers the system to detect objects that are partially occluded, by incorporating contextual information surrounding small objects.

### 3.7 Spatio-Temporal Information

In this section, our focus is exclusively on video-based object detectors that aim to identify small objects. While many of these studies have been tested on the ImageNet VID dataset[1] [91], this dataset was not originally intended for SOD. Nonetheless, a few of the works also reported their results for small objects of ImageNet VID dataset. The topic of tracking and detecting small objects in videos has also been explored using transformer architectures. Although techniques for image-based SOD can be applied to video, they generally do not utilize the valuable temporal information, which can be particularly beneficial for identifying small objects in cluttered or occluded frames. The application of transformers to generic object detection/tracking started with TrackFormer [73] and TransT [15]. These models used frame-to-frame (setting the previous frame as the reference) set prediction and template-to-frame (setting a template frame as the reference) detection. Liu *et al.* in [63] were among the first to use transformers specifically for video-based SOD and tracking. Their core concept is to update template frames to capture any small changes induced by the presence of small objects and to provide a global attention-driven relationship between the template frame and the search frame.

Transformer-based object detection gained formal recognition with the introduction of TransVOD, an end-to-end object detector, as presented in [43] and [128]. This model applies both spatial and temporal transformers to a series of video frames, thereby identifying and linking objects across these frames. TransVOD has spawned several variants, each with unique features, including capabilities for real-time detection. PTSEFormer [104] adopts a progressive strategy, focusing on both temporal information and the objects' spatial transitions between frames. It employs multi-scale feature extraction to achieve this. Unlike other models, PTSEFormer directly regresses object queries from adjacent frames rather than the entire dataset, offering a more localized approach.

Sparse VOD [39] proposed an end-to-end trainable video object detector that incorporates temporal information to propose region proposals. In contrast, DAFA [90] highlights the significance of global features within a video as opposed to local temporal features. DEFA showed the inefficiency of the **First In First Out** (**FIFO**) memory structure and proposed a diversity-aware memory, which uses object-level memory instead of frame-level memory for the attention module. VSTAM [33] improves feature quality on an element-by-element basis and then performs sparse aggregation

[1] https://paperswithcode.com/sota/video-object-detection-on-imagenet-vid

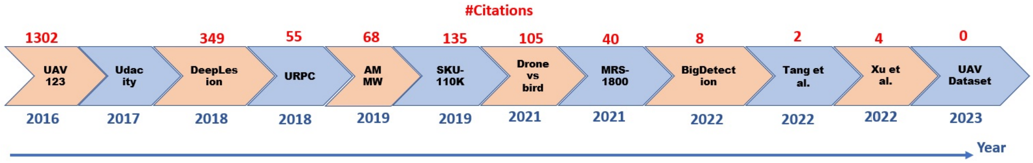

Fig. 10. Chronology of SOD datasets with number of citations (based on Google Scholar).

before these enhanced features are used for object candidate region detection. The model also incorporates external memory to take advantage of long-term contextual information.

In the FAQ work [22], a novel video object detector is proposed that uses query feature aggregation in the decoder module. This is different than the methods that focus on either feature aggregation in the encoder or the methods that perform post-processing for various frames. The research indicates that this technique improves the detection performance outperforming SOTA methods.

## 4 Results and Benchmarks

In this section, we present both quantitative and qualitative evaluations of previous works on SOD, identifying the most effective techniques tailored to specific applications. The quantitative analysis involved sourcing and examining data directly from original publications. Qualitative analysis, on the other hand, entailed applying models to selected images featuring numerous small objects, and subsequently comparing the results. Additionally, this section introduces new datasets specifically curated for SOD, encompassing both videos and images suitable for a wide array of applications.

### 4.1 Datasets

In this subsection, in addition to the widely used MS COCO dataset, we compile and present 12 new SOD datasets. These new datasets are primarily designed for specific applications, deliberately excluding datasets which were thoroughly covered in our previous survey [88], in order to minimize redundancy and avoid overlap. Figure 10 displays the chronological order of these datasets along with their citation count as of 15 June 2023, according to Google Scholar.

**UAV123** [75]: This dataset contains 123 videos acquired with UAVs and it is one of the largest object-tracking datasets with more than 110K frames.

**MRS-1800** [115]: This dataset consists of a combination of images from three other remote sensing datasets: DIOR [52], NWPU VHR-10 [18], and HRRSD [127]. MRD-1800 was created for the dual purpose of detection and instance segmentation, with 1800 manually annotated images which include three types of objects: airplanes, ships, and storage tanks.

**SKU-110K [36]**: This dataset serves as a rigorous testbed for commodity detection, featuring images captured from various supermarkets around the world. The dataset includes a range of scales, camera angles, lighting conditions, and so on.

**BigDetection**[4]: This is a large-scale dataset that is crafted by integrating existing datasets and meticulously eliminating duplicate boxes while labeling overlooked objects. It has a balanced number of objects across all sizes making it a pivotal resource for advancing the field of object detection. Using this dataset for pretraining and subsequently fine-tuning on MS COCO significantly enhances performance outcomes.

**Tang et al.** [98]: Originating from video footage of field activities within a chemical plant, this dataset covers various types of work such as hot work, aerial work, confined space operations, and so on. It includes category labels like people, helmets, fire extinguishers, gloves, work clothes and other relevant objects.

Table 2. Commonly Used Datasets for SOD

| Dataset | Application | Video | Image | Shooting Angle (Type) | Resolution (pixels) | #Object Classes | #Instances | #Images/Videos | Public? |
|---|---|---|---|---|---|---|---|---|---|
| **UAV123** [75] | UAV Tracking | ✓ | | Aerial Perspective(RGB) | – | – | – | 123 (>110K frames) | Yes: Click Here |
| **MRS-1800** [115] | Remote Sensing | | ✓ | Satellite based(RGB) | NF | 3 | 16,318 | 1800 | – |
| **SKU-110K**[36] | Commodity Detection | | ✓ | Normal | NF | 110,712 | 147.4 per image | 11,762 | Yes: Click Here |
| **BigDetection**[4] | Generic | | ✓ | Normal | NF | 600 | 36M | 3.4M Training<br>141K Test | Yes: Click Here |
| **Tang et al.** [98] | Cemical Plant Monitoring | | ✓ | Normal | – | 19 | – | 2400 | – |
| **Xu et al.** [114] | UAV-based Detection | | ✓ | Aerial (RGB) | 1920×1080 | 2 | 12.5K | 2K | Yes: Click Here |
| **DeepLesion** [117] | Lesion detection | | ✓ | (CT) | – | 8 | 32.7K | 32.1K | Yes: Click Here |
| **Udacity Self Driving Car** [1] | Self-Driving | ✓ | | Normal | 1920×1200 | 3 | 65K | 9,423 | Yes: Click Here |
| **AMMW Dataset** [65] | Security Inspection | | ✓ | Normal (AMMW) | 160×400 | >30 | – | >58K | – |
| **URPC 2018 Dataset**[2] | Underwater Detection | | ✓ | Normal | – | 4 | – | 2,901 Training<br>800 Test | – |
| **UAV dataset** [120] | UAV-based detection | | ✓ | Aerial Perspective (RGB) | – | 7 | 320,624 | 9,630 | – |
| **Drone-vs-bird** [21] | Drone Detection | ✓ | | Normal | NF | 2 | – | 77 Training 1,384 Frames | Yes: Click Here |

NF: Not fixed.

**Xu et al.** [114]: This publicly available dataset focuses on **UAV** (**Unmanned Aerial Vehicle**)-captured images and contains 2K images aimed at detecting both pedestrians and vehicles. The images were collected using a DJI drone and feature diverse conditions such as varying light levels and densely parked vehicles.

**DeepLesion** [117]: Comprising CT scans from 4,427 patients, this dataset ranks among the largest of its kind. It includes a variety of lesion types, such as pulmonary nodules, bone abnormalities, kidney lesions, and enlarged lymph nodes. The objects of interest in these images are typically small and accompanied by noise, making their identification challenging.

**Udacity Self Driving Car** [1]: Designed solely for educational use, this dataset features driving scenarios in Mountain View and nearby cities captured at a 2Hz image acquisition rate. The category labels within this dataset include cars, trucks, and pedestrians.

**AMMW Dataset** [65]: Created for security applications, this active millimetre-wave image dataset includes more than 30 different types of objects. These include two kinds of lighters (made of plastic and metal), a simulated firearm, a knife, a blade, a bullet shell, a phone, a soup, a key, a magnet, a liquid bottle, an absorbent material, a match, and so on.

**URPC 2018 Dataset**: This underwater image dataset includes four types of objects: holothurian, echinus, scallop, and starfish [62].

**UAV dataset** [120]: This image dataset includes more than 9K images captured via UAVs in different weather and lighting conditions and various complex backgrounds. The objects in this dataset are sedans, people, motors, bicycles, trucks, buses, and tricycles.

**Drone-vs-bird** [21]: This video dataset aims to address the security concerns of drones flying over sensitive areas. It offers labeled video sequences to differentiate between birds and drones under various illumination, lighting, weather, and background conditions.

A summary of these datasets, including their applications, type, resolutions, number of classes/instances/images/frame, and a link to their webpage, is provided in Table 2.

## 4.2 Benchmarks in Vision Applications

In this subsection, we introduce various vision-based applications where the detection performance of small objects is vital. For each application, we select one of the most popular datasets and report its performance metrics, along with details of the experimental setup.

*4.2.1 Generic Applications.* For generic applications, we evaluate the performance of all small object detectors on the challenging MS COCO benchmark [61]. The choice of this dataset is based

[2] http://en.cnurpc.org/

Table 3. Detection Performance (%) for Small-Scale Objects (area< $32^2$) on MS COCO Image Dataset [61]

| Model | Backbone | GFLOPS↓/FPS ↑ | #params↓ | $\text{mAP}^{@[0.5,0.95]}$ ↑ | Epochs↓ | URL |
|---|---|---|---|---|---|---|
| Faster RCNN-DC5 (NeurIPS2015)[89] | ResNet50 | 320/16 | 166M | 21.4 | 37 | Link |
| Faster RCNN-FPN (NeurIPS2015)[89] | ResNet50 | 180/26 | 42M | 24.2 | 37 | Link |
| Faster RCNN-FPN (NeurIPS2015)[89] | ResNet101 | 246/20 | 60M | 25.2 | – | Link |
| RepPoints v2-DCN-MS (NeurIPS2020)[16] | ResNeXt101 | –/– | – | 34.5* | 24 | Link |
| FCOS (ICCV2019)[99] | ResNet50 | 177/17 | – | 26.2 | 36 | Link |
| CBNet V2-DCN(ATSS[126]) (TIP2022)[57] | Res2Net101 | –/– | 107M | 35.7* | 20 | Link |
| CBNet V2-DCN(Cascade RCNN) (TIP2022)[57] | Res2Net101 | –/– | 146M | 37.4* | 32 | Link |
| DETR (ECCV2020)[7] | ResNet50 | 86/28 | 41M | 20.5 | 500 | Link |
| DETR-DC5 (ECCV2020)[7] | ResNet50 | 187/12 | 41M | 22.5 | 500 | Link |
| DETR (ECCV2020)[7] | ResNet101 | **52**/20 | 60M | 21.9 | – | Link |
| DETR-DC5 (ECCV2020)[7] | ResNet101 | 253/10 | 60M | 23.7 | – | Link |
| ViT-FRCNN (arXiv2020)[2] | – | –/– | – | 17.8 | – | – |
| RelationNet++ (NeurIPS2020)[19] | ResNeXt101 | –/– | – | 32.8* | – | Link |
| RelationNet++-MS (NeurIPS2020)[19] | ResNeXt101 | –/– | – | 35.8* | – | Link |
| Deformable DETR (ICLR2021)[130] | ResNet50 | 173/19 | 40M | 26.4 | 50 | Link |
| Deformable DETR-IBR (ICLR2021)[130] | ResNet50 | 173/19 | 40M | 26.8 | 50 | Link |
| Deformable DETR-TS (ICLR2021)[130] | ResNet50 | 173/19 | 40M | 28.8 | 50 | Link |
| Deformable DETR-TS-IBR-DCN (ICLR2021)[130] | ResNeXt101 | –/– | – | 34.4* | – | Link |
| Dynamic DETR (ICCV2021)[25] | ResNet50 | –/– | – | 28.6* | – | – |
| Dynamic DETR-DCN (ICCV2021)[25] | ResNeXt101 | –/– | – | 30.3* | – | – |
| TSP-FCOS (ICCV2021)[97] | ResNet101 | 255/12 | – | 27.7 | 36 | Link |
| TSP-RCNN (ICCV2021)[97] | ResNet101 | 254/9 | – | 29.9 | 96 | Link |
| Mask R-CNN (ICCV2021)[81] | Conformer-S/16 | 457.7/– | 56.9M | 28.7 | **12** | Link |
| Conditional DETR-DC5 (ICCV2021)[74] | ResNet101 | 262/– | 63M | 27.2 | 108 | Link |
| SOF-DETR (2022JVCIR) [31] | ResNet50 | –/– | – | 21.7 | – | Link |
| DETR++ (arXiv2022)[123] | ResNet50 | –/– | – | 22.1 | – | – |
| TOLO-MS (NCA2022) [112] | – | –/**57** | – | 24.1 | – | – |
| Anchor DETR-DC5 (AAAI2022) [108] | ResNet101 | –/– | – | 25.8 | 50 | Link |
| DESTR-DC5 (CVPR2022)[42] | ResNet101 | 299/– | 88M | 28.2 | 50 | – |
| Conditional DETR v2-DC5 (arXiv2022)[14] | ResNet101 | 228/– | 65M | 26.3 | 50 | – |
| Conditional DETR v2 (arXiv2022)[14] | Hourglass48 | 521/– | 90M | 32.1 | 50 | – |
| FP-DETR-IN (ICLR2022) [107] | – | –/– | **36M** | 26.5 | 50 | Link |
| DAB-DETR-DC5 (arXiv2022)[64] | ResNet101 | 296/– | 63M | 28.1 | 50 | Link |
| Ghostformer-MS (Sensors2022)[54] | GhostNet | –/– | – | 29.2 | 100 | – |
| CF-DETR-DCN-TTA (AAAI2022)[6] | ResNeXt101 | –/– | – | 35.1* | – | – |
| CBNet V2-TTA (CVPR2022)[4] | Swin Transformer-base | –/– | – | 41.7 | – | Link |
| CBNet V2-TTA-BD (CVPR2022)[4] | Swin Transformer-base | –/– | – | 42.2 | – | Link |
| DETA (arXiv2022)[77] | ResNet50 | –/13 | 48M | 34.3 | 24 | Link |
| DINO (arXiv2022)[125] | ResNet50 | 860/10 | 47M | 32.3 | **12** | Link |
| CO-DINO Deformable DETR-MS-IN (arXiv2022)[132] | Swin Transformer-large | –/– | – | 43.7 | 36 | Link |
| HYNETER (ICASSP2023)[8] | Hyneter-Max | –/– | 247M | 29.8* | – | – |
| DeoT (JRTIP2023) [28] | ResNet101 | 217/14 | 58M | 31.4 | 34 | – |
| ConformerDet-MS (TPAMI2023) [80] | Conformer-B | –/– | 147M | 35.3 | 36 | Link |
| YOLOS (NeurIPS2021)[32] | DeiT-base | –/3.9 | 100M | 19.5 | 150 | Link |
| DETR(ViT) (arXiv2021)[94] | Swin Transformer-base | –/9.7 | 100M | 18.3 | 50 | Link |
| Deformable DETR(ViT) (arXiv2021)[94] | Swin Transformer-base | –/4.8 | 100M | 34.5 | 50 | Link |
| ViDT (arXiv2022)[94] | Swin Transformer-base | –/9 | 100M | 30.6 | 50 | Link |
| DFFT (ECCV2022) [11] | DOT-medium | 67/– | – | 25.5 | 36 | Link |
| CenterNet++-MS (arXiv2022) [30] | Swin Transformer-large | –/– | – | 38.7* | – | Link |
| DETA-OB (arXiv2022)[77] | Swin Transformer-large | –/4.2 | – | 46.1* | 24 | Link |
| Group DETR v2-MS-IN-OB (arXiv2022) [13] | ViT-Huge | –/– | 629M | **48.4*** | – | – |
| Best Results | NA | DETR/TOLO | FP-DETR | Group DETR v2 | DINO | NA |

The top section shows results for CNN-based techniques, the middle section shows results for mixed architectures, and the bottom section presents from transformer-only networks. DC5: Dilated C5 stage, MS: Multi-scale network, IBR: Iterative bounding box refinement, TS: Two-stage detection, DCN: Deformable convnets, TTA: Test time augmentation, BD: Pre-trained on BigDetection dataset, IN: Pre-trained on ImageNet, OB: Pre-trained on Object-365 [34]. While * shows the results for COCO test-dev, the other values are reported for COCO val set.

on its wide acceptance in the object detection field and the accessibility of performance results. The MS COCO dataset consists of approximately 160K images across 80 categories. While the authors are advised to train their algorithms using the COCO 2017 training and validation sets, they are not restricted to these subsets.

In Table 3, we examine and evaluate the performance of all the techniques under review that have reported their results on MS COCO (due to inconsistent reporting practices across the literature, many studies did not present a complete set of performance metrics). The table provides information on the backbone architecture, GFLOPS/FPS (indicating the computational overhead and execution speed), number of parameters (indicating the scale of the model), mAP (mean average precision: a measure of object detection performance), and epochs (indicating the convergence time). Additionally, a link to each method's webpage is provided for further information. The methods are categorized into three groups: CNN-based, mixed, and transformer-only methods. The top-performing methods for each metric are shown in the table's last row. It should be noted that this comparison was only feasible for methods that have reported values for each specific metric. In instances where there is a tie, the method with the highest mAP was deemed the best. The default mAP values are for the ”COCO 2017 val” set, while those for the ”COCO test-dev” set are marked with an asterisk. Please be aware that the reported mAP is only for objects with area$< 32^2$.

Upon examining Table 3, it is obvious that most techniques benefit from using a mix of CNN and transformer architectures, essentially adopting hybrid strategies. Notably, Group DETR v2, which relies solely on a transformer-based architecture, attains a mAP of 48.4%. However, achieving such a performance requires the adoption of additional techniques such as pre-training on two large-scale datasets and multi-scale learning. In terms of convergence, DINO outperforms by reaching stable results after just 12 epochs, while also securing a commendable mAP of 32.3%. Conversely, the original DETR model has the lowest GFLOPS while TOLO has the fastest inference time. FP-DETR stands out for having the lightest network with only 36M parameters.

Drawing from these findings, we conclude that pre-training and multi-scale learning are currently the most effective strategies for achieving high performance in SOD. This conclusion can be justified by referring to all the methods with mAP over 40%, i.e., Group DETR v2, DETA-OB, CO-DINO Deformable DETR, CBNet V2 where pre-training and multiscale-learning are the common strategies among these detectors. This may be attributed to the imbalance problem of the dataset and the lack of informative features in small objects.

Figure 11, which spans two pages, along with its more detailed counterpart in Figure 12, illustrates the detection results of various transformers and CNN-based methods. These are compared to each other using selected images from the COCO dataset and implemented by us using their public models available on their GitHub pages. The analysis reveals that Faster RCNN and SSD fall short in accurately detecting small objects. Specifically, SSD either misses most objects or generates numerous bounding boxes with false labels and poorly located bounding boxes. While Faster RCNN performs better, it still produces low-confidence bounding boxes and occasionally assigns incorrect labels.

In contrast, DETR has the tendency to over-estimate the number of objects, leading to multiple bounding boxes for individual objects. It is commonly noted that DETR is prone to generating false positives. Finally, among the methods evaluated, CBNet V2 stands out for its superior performance. As observed, it produces high confidence scores for the objects it detects, even though it may occasionally misidentify some objects.

*4.2.2 Small Object Detection in Aerial Images .* Another interesting use of detecting small objects is in the area of remote sensing. This field is particularly appealing because many organizations and research bodies aim to routinely monitor the Earth's surface through aerial images to collect both national and international data for statistics. While these images can be acquired using various modalities, this survey focuses only on non-SAR images. This is because SAR images have been extensively researched and deserve their own separate study. Nonetheless, the learning techniques discussed in this survey could also be applicable to SAR images.

In aerial images, objects often appear small due to their significant distance from the camera. The bird's-eye view also adds complexity to the task of object detection, as objects can be situated anywhere within the image. To assess the performance of transformer-based detectors designed for such applications, we selected the DOTA image dataset [111], which has become a widely used benchmark in the field of object detection. Figure 13 displays some sample images from the DOTA dataset featuring small objects. The dataset includes a predefined Training set, Validation set, and Testing set. In comparison to generic applications, this particular application has received relatively less attention from transformer experts. However, as indicated in Table 4, ReDet distinguishes itself through its multi-scale learning strategy and pre-training on the ImageNet dataset, achieving the highest precision value (80.89%) and requiring only 12 training epochs. This mirrors the insights gained from the COCO dataset analysis, suggesting that optimal performance can be attained by addressing imbalances in downstream tasks and including informative features from small objects. **Note:** Some entries in the tables are incomplete due to inconsistencies in how performance metrics are reported in the original publications. In several cases, only partial results were available, and the absence of publicly released code or detailed implementation information–such as preprocessing steps, training configurations, hyperparameter settings, or evaluation scripts–prevented us from reproducing the missing values. To ensure accuracy, we report the results exactly as presented in the

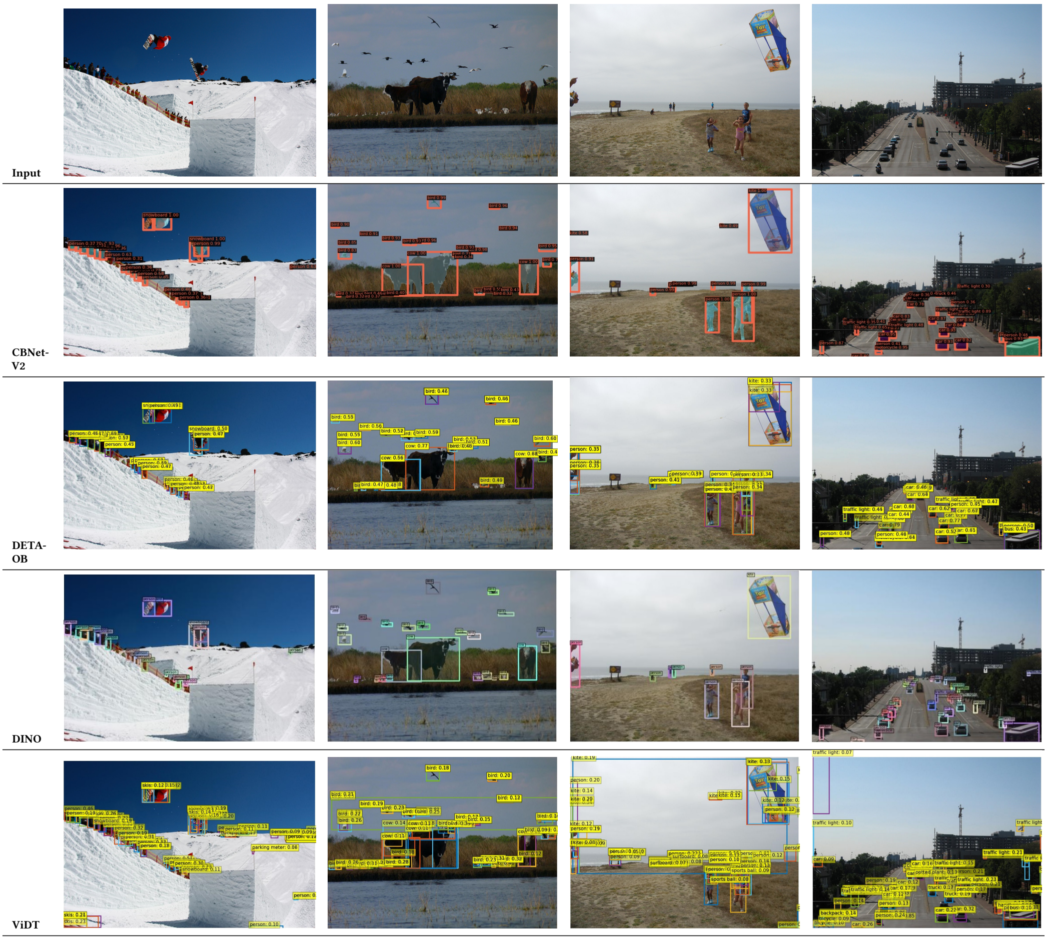

Fig. 11. Examples of detection results on COCO dataset [61] for transformer-based SOTA small object detectors compared with Convolutional networks.

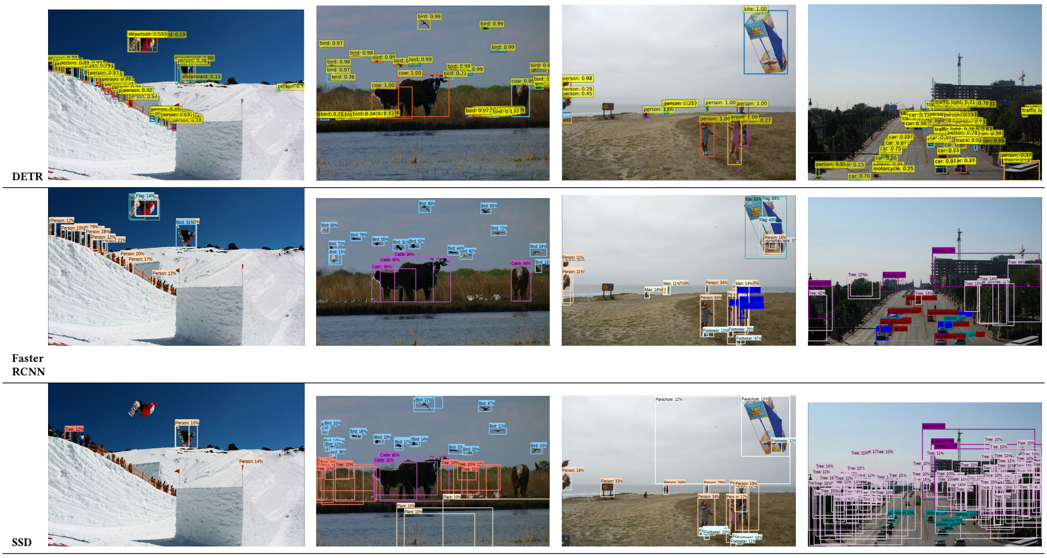

Fig. 11. (Continued)

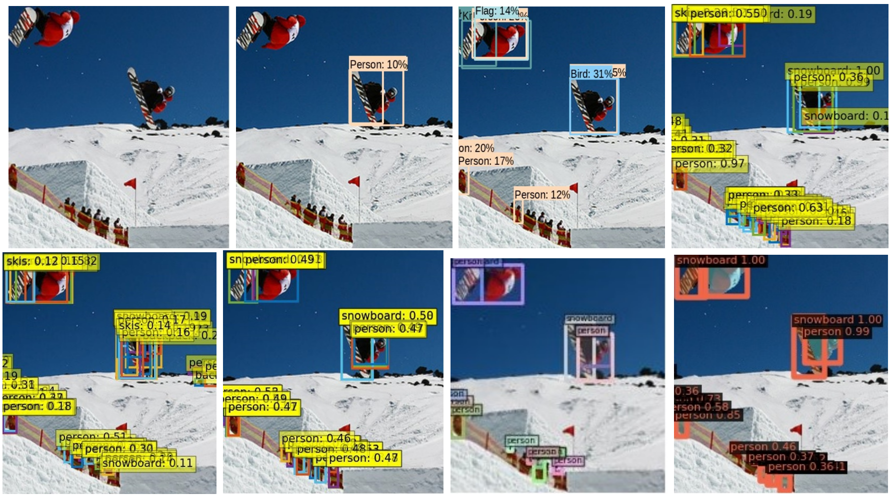

Fig. 12. Detection results on a sample image when zoomed in. First row from the left: Input image, SSD, Faster RCNN, DETR. Second row from the left: ViDT, DETA-OB, DINO, CBNet v2.

original sources. This limitation underscores broader challenges in the field related to reproducibility and the lack of standardized benchmarking practices.

*4.2.3 Small Object Detection in Medical Images.* In the field of medical imaging, specialists are often tasked with early detection and identification of anomalies. Missing even barely visible or small abnormal cells can lead to serious repercussions for patients, including cancer and life-threatening conditions. These small-sized objects can be found as abnormalities in the retina of diabetic patients, early tumors, vascular plaques, and so on. Despite the critical nature and the potential life-threatening impact of this research area, only a handful of studies have tackled the

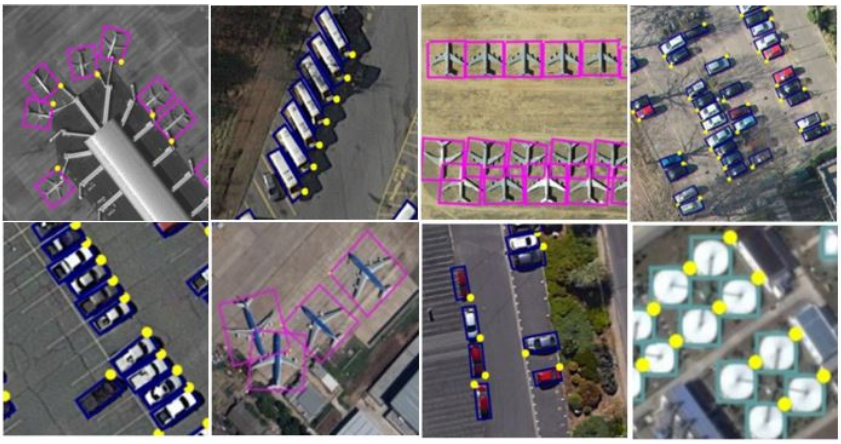

Fig. 13. Example of small objects in DOTA image dataset.

Table 4. Detection Performance (%) for Small-Scale Objects on DOTA Image Dataset [111]

| Model | Backbone | FPS ↑ | #params↓ | mAP ↑ | Epochs↓ | URL |
|---|---|---|---|---|---|---|
| Rotated Faster RCNN-MS (NeurIPS2015)[89] | ResNet101 | – | 64M | 67.71 | 50 | Link |
| SSD (ECCV2016) [66] | – | – | – | 56.1 | – | Link |
| RetinaNet-MS (ICCV2017)[60] | ResNet101 | – | **59M** | 66.53 | 50 | Link |
| ROI-Transformer-MS-IN (CVPR2019) [27, 106] | ResNet50 | – | – | 80.06 | **12** | Link |
| Yolov5 (2020)[46] | – | **95** | – | 64.5 | – | Link |
| ReDet-MS-FPN (CVPR2021)[38] | ResNet50 | – | – | 80.1 | – | Link |
| O$^2$DETR-MS (arXiv2021)[71] | ResNet101 | – | 63M | 70.02 | 50 | – |
| O$^2$DETR-MS-FT (arXiv2021)[71] | ResNet101 | – | – | 76.23 | 62 | – |
| O$^2$DETR-MS-FPN-FT (arXiv2021)[71] | ResNet50 | – | – | 79.66 | – | – |
| SPH-Yolov5 (RS2022) [37] | Swin Transformer-base | 51 | – | 71.6 | 150 | – |
| AO2-DETR-MS (TCSVT2022)[24] | ResNet50 | – | – | 79.22 | – | Link |
| MDCT (RS2023)[10] | – | – | – | 75.7 | – | – |
| ReDet-MS-IN (arXiv2023)[106] | ViTDet, ViT-B | – | – | **80.89** | **12** | Link |
| Best Results | NA | Yolov5 | RetinaNet | ReDet-MS-IN | ReDet-MS-IN | NA |

The top section shows results for CNN-based techniques, the middle section shows results for mixed architectures. MS: Multi-scale network, FT: Fine-tuned, FPN: Feature pyramid network, IN: Pre-trained on ImageNet.

challenges associated with detecting small objects in this crucial application. For those interested in this topic, the DeepLesion CT image dataset [117] has been selected as the benchmark due to the availability of the results for this particular dataset [51]. Sample images from this dataset are shown in Figure 14 (Left). This dataset is divided into three sets: training (70%), validation (15%), and test (15%) sets [93]. Table 5 compares the accuracy and mAP of three transformer-based studies against both two-stage and one-stage detectors. The MS Transformer emerges as the best technique with this dataset, albeit with limited competition. Its primary innovation lies in self-supervised learning and the incorporation of a masking mechanism within a hierarchical transformer model. Overall, with an accuracy of 90.3% and an mAP of 89.6%, this dataset appears to be less challenging compared to other medical imaging tasks, especially considering that some tumor detection tasks are virtually invisible to the human eyes.

*4.2.4 Small Object Detection in Underwater Images.* With the growth of underwater activities, the demand to monitor hazy and low-light environments has increased for purposes like ecological surveillance, equipment maintenance, and monitoring of wreck fishing. Factors like scattering and light absorption of the water, make the SOD task even more challenging. Example images of such challenging environments are displayed in Figure 14 (Right). Transformer-based detection methods

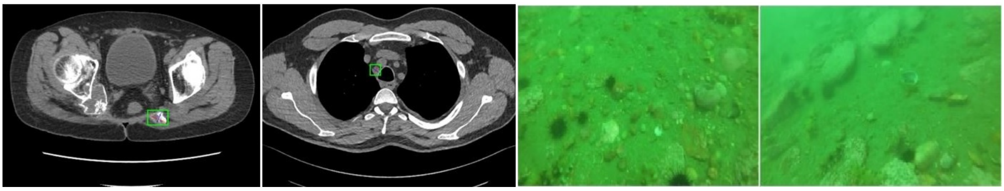

Fig. 14. Example of small abnormalities in DeepLesion image dataset (two images from the left) [117] and examples of low quality images in URPC2018 image dataset (two images from the right).

Table 5. Detection Performance (%) for DeepLesion CT Image Dataset [117]

| Model | Accuracy ↑ | $\text{mAP}^{0.5}$ ↑ |
|---|---|---|
| Faster RCNN (NeurIPS2015)[89] | 83.3 | 83.3 |
| Yolov5 [46] | 85.2 | 88.2 |
| DETR (ECCV2020)[7] | 86.7 | 87.8 |
| Swin Transformer | 82.9 | 81.2 |
| MS Transformer (CIN2022)[93] | **90.3** | **89.6** |
| Best Results | MS Transformer | MS Transformer |

The top section shows results for CNN-based techniques, the middle section shows results for mixed architectures.

Table 6. Detection Performance (%) for URPC2018 Dataset [62]

| Model | #Params↓ | $\text{mAP}^{@[0.5,0.95]}$ ↑ | $\text{mAP}^{0.5}$ ↑ |
|---|---|---|---|
| Faster RCNN (NeurIPS2015)[89] | 33.6M | 16.4 | – |
| Cascade RCNN (CVPR2018)[5] | 68.9M | 16 | – |
| Dynamic RCNN (ECCV2020) [124] | 41.5M | 13.3 | – |
| Yolov3 [46] | 61.5M | 19.4 | – |
| RoIMix (ICASSP2020) [62] | – | – | **74.92** |
| HTDet (RS2023) [9] | **7.7M** | **22.8** | – |
| Best Results | HTDet | HTDet | RoIMix |

The top section shows results for CNN-based techniques, the middle section shows results for mixed architectures.

should not only be adept at identifying small objects but also need to be robust against the poor image quality found in deep waters, as well as variations in color channels due to differing rates of light attenuation for each channel.

Table 6 shows the performance metrics reported in existing studies for this dataset. HTDet is the sole transformer-based technique identified for this specific application. It significantly outperforms the SOTA CNN-based method by a huge margin (3.4% in mAP). However, the relatively low mAP scores confirm that object detection in underwater images remains a difficult task. It is worth noting that the training set of the URPC 2018 contains 2901 labeled images, and the testing set contains 800 unlabeled images [9].

*4.2.5 Small Object Detection in Active Milli-Meter Wave Images.* Small objects can easily be concealed or hidden from normal RGB cameras, for example, within a person's clothing at an airport. Therefore, active imaging techniques are essential for security purposes. In these scenarios, multiple images are often captured from different angles to enhance the likelihood of detecting

Table 7. Detection Performance (%) for AMWW Image Dataset [65]

| Model | Backbone | $\text{mAP}^{0.5}\uparrow$ | $\text{mAP}^{@[0.5,0.95]}\uparrow$ |
|---|---|---|---|
| Faster RCNN (NeurIPS2015)[89] | ResNet50 | 70.7 | 26.83 |
| Cascade RCNN (CVPR2018)[5] | ResNet50 | 74.7 | 27.8 |
| TridentNet (ICCV2019) [55] | ResNet50 | 77.3 | 29.2 |
| Dynamic RCNN (ECCV2020) [124] | ResNet50 | 76.3 | 27.6 |
| Yolov5 [46] | ResNet50 | 76.67 | 28.48 |
| MATR (TCSVT2022) [96] | ResNet50 | **82.16** | **33.42** |
| Best Results | NA | MATR | MATR |

The top section shows results for CNN-based techniques, the middle section shows results for mixed architectures.

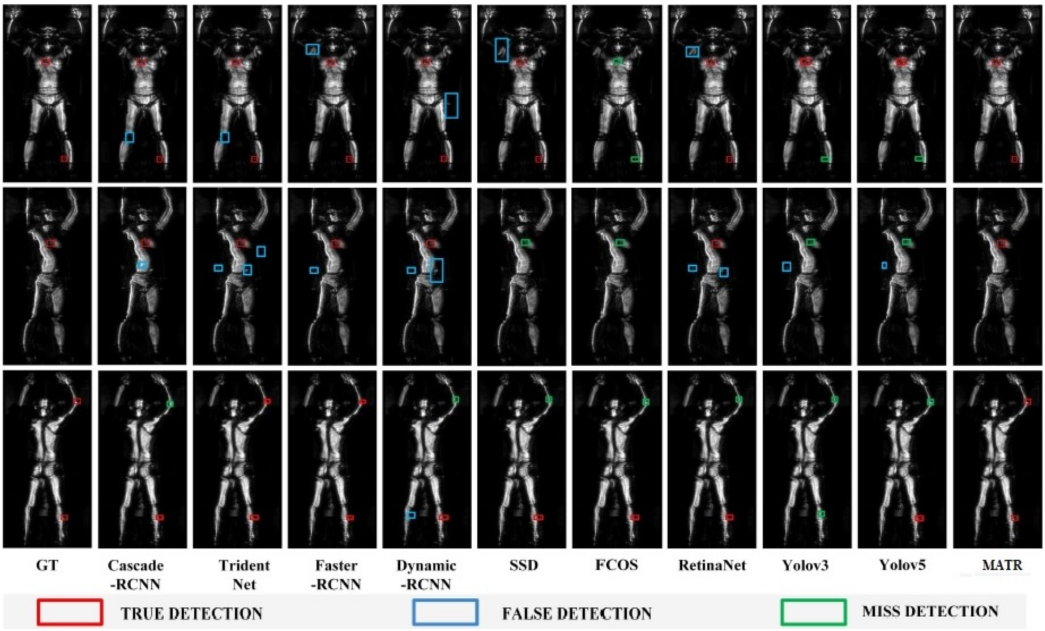

Fig. 15. Examples of detection results on AMMW image dataset [65] for SOTA small object detectors (figure from [96]).

even minuscule objects. Interestingly, much like in the field of medical imaging, transformers are rarely used for this particular application.

In our study, we focused on the detection performance of existing techniques using the AMMW Dataset [65] as shown in Table 7 (results are compiled from their papers). We have identified that MATR emerged as the sole technique that combines transformer and CNNs for this dataset. Despite being the only transformer-based technique, it could significantly improve the SOD performance (5.49% ↑ in $\text{mAP}^{0.5}$ with respect to Yolov5 and 4.22 % ↑ in $\text{mAP}^{@[0.5,0.95]}$ with respect to TridentNet) with the same backbone (ResNet50). Figure 15 visually compares MATR with other SOTA CNN-based techniques. Combining images from different angles largely helps to identify even small objects within this imaging approach. For training and testing, 35426 and 4019 images were used, respectively [96].

Table 8. Detection Performance (%) for ImageNet VID Dataset [91] for Small Objects

| Model | Backbone | $mAP^{@[0.5,0.95]}\uparrow$ |
|---|---|---|
| Faster RCNN (NeurIPS2015)[89]+SELSA[109] | ResNet50 | 8.5 |
| Deformable-DETR-PT [130] | ResNet50 | 10.5 |
| Deformable-DETR[130]+TransVOD-PT[128] | ResNet50 | 11 |
| DAB-DETR[64]+FAQ-PT[22] | ResNet50 | 12 |
| Deformable-DETR[130]+FAQ-PT[22] | ResNet50 | **13.2** |
| Best Results | NA | Deformable-DETR+FAQ |

The top section shows results for CNN-based techniques, the middle section shows results for mixed architectures. PT: Pre-trained on MS COCO.

*4.2.6 Small Object Detection in Videos.* The field of object detection in videos gained considerable attention recently, as the temporal information in videos can improve the detection performance. To benchmark the SOTA techniques, the ImageNet VID dataset has been used with results specifically focused on the dataset's small objects. This dataset includes 3862 training videos and 555 validation videos with 30 classes of objects. Table 8 reports the mAP of several recently developed transformer-based techniques. Despite the growing application of transformers in video object detection, their potential in SOD remains largely unexplored. Among the methods that have reported SOD performance on the ImageNet VID dataset, Deformable DETR with FAQ stands out for achieving the highest performance- although it is notably low at 13.2 % for $mAP^{@[0.5,0.95]}$). This highlights a significant research gap in the area of video-based SOD.

## 5 Discussion

In this survey article, we explored how transformer-based approaches can address the challenges of SOD. Our taxonomy divides transformer-based small object detectors into seven main categories: object representation, fast attention (useful for high-resolution and multi-scale feature maps), architecture and block modification, spatio-temporal information, improved feature representation, auxiliary techniques, and fully transformer-based detectors. When juxtaposing this taxonomy with the one for CNN-based techniques [88], we observe that some of these categories overlap, while others are unique to transformer-based techniques. Certain strategies are implicitly embedded into transformers, such as attention and context learning, which are performed via the self and cross-attention modules in the encoder and decoder. On the other hand, multi-scale learning, auxiliary tasks, architecture modification, and data augmentation are commonly used in both paradigms. However, it is important to note that while CNNs handle spatio-temporal analysis through 3D-CNN, RNN, or feature aggregation over time, transformers achieve this by using successive spatial and temporal transformers or by updating object queries for successive frames in the decoder. On the other hand, the tiling technique has shown promise in enhancing the detection of small objects by increasing their relative size in localized image regions [78]. While this strategy has been explored in CNN-based approaches, its integration with transformer-based detectors remains largely unexplored. Future research could investigate how spatial tiling combined with transformer architectures can improve SOD performance while maintaining scalability and efficiency. In summary, our findings underscore the effectiveness of strategies such as pre-training and multi-scale learning in achieving SOTA SOD performance across diverse datasets. Leveraging data fusion methods for SOD also presents a promising avenue, as it grants detectors access to rich, stacked information from the same object, potentially improving the performance compared to single image-based detection approaches. Despite the widespread adoption of transformers in SOD, which has notably improved performance across various applications, our analysis identifies several challenges, signaling key areas for future research. **Efficient Architecture/Learning:** While transformers have significantly improved the localization

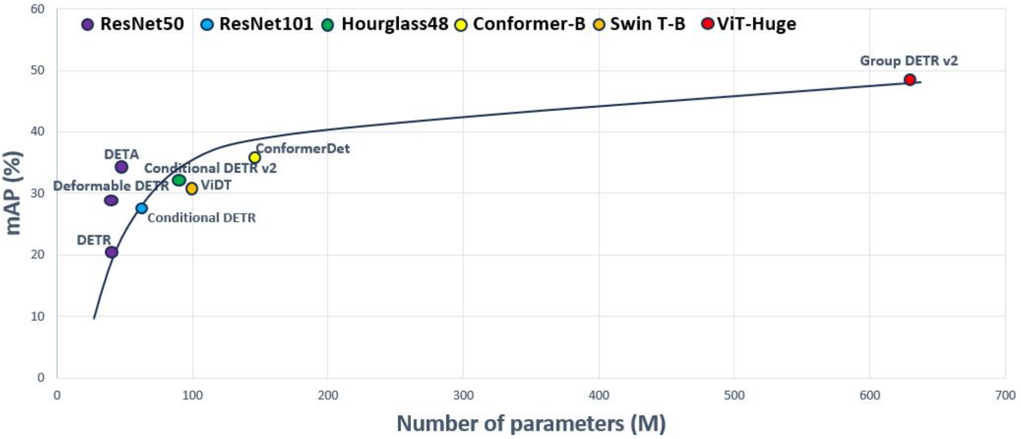

Fig. 16. mAP vs the number of model parameters in SOTA transformer-based small object detectors.

and classification of small objects, it is crucial to recognize the associated tradeoffs. These include a substantial parameter count (often in the billions), lengthy training periods (spanning several hundred epochs), and reliance on pretraining with extremely large datasets, a process often impractical without robust computational resources. These factors collectively limit accessibility to users who lack the necessary resources for training and testing these techniques for their specific tasks. Hence, there's a pressing need for lightweight networks equipped with domain-specific accelerators, novel optimization methods, efficient learning approaches, and tailored architectures [47]. Additionally, in some instances, it might be possible to substitute the global attention mechanism with lighter global interaction methods [58]. Despite the burgeoning parameter counts, now comparable to those of the human brain [44], performance in SOD still falls significantly short of human capabilities. Figure 16 illustrates this limited improvement in SOD performance, despite substantial growth in the scale of object detectors, both in terms of parameter count and backbone architecture.

**Missing Objects and Redundant Detections:** Our qualitative examination (Figures 11 and 12) indicates that classification errors are negligible in SOD, with the primary error source being inaccurate object localization. Analysis of the same figures reveals two primary challenges in SOD: (i) the occurrence of false negatives, or missing objects, and (ii) the presence of redundant detections. The challenge of missing objects often stems from the limited information encoded by the tokens. One potential solution is to employ higher-resolution images or enhance the feature pyramid architecture. However, these strategies may prolong processing times, which could be mitigated by adopting more streamlined and efficient network designs, as suggested in the first future direction. Addressing redundant detections, traditionally managed in CNN-based detectors through post-processing methods like NMS, presents a different challenge within the transformer framework. Here, a more effective approach might involve reducing the similarity between object queries in the decoder. This could be achieved through auxiliary loss functions or techniques akin to NMS, optimizing detection performance while maintaining computational efficiency.

**Application Specific Small Object Detectors:** We examined studies that employed transformers for a range of SOD tasks. These include generic detection, detection in aerial images, abnormality detection in medical images, small hidden object detection in active millimeter-wave images for security purposes, underwater object detection, and SOD in videos. Despite their effectiveness in generic and aerial image tasks, transformers remain relatively underexplored in other domains [88]. This observation is particularly notable considering the potential impact transformers could have in critical areas like medical imaging. Our analysis revealed that SOTA techniques in generic applications often fail to deliver consistent performance across these specialized domains. Models tailored for such tasks typically necessitate fine-tuning, adaptation to novel object classes, and leveraging prior knowledge. One potential explanation for the scarcity of specialized models is the insufficient availability of labeled small objects or the presence of noisy annotations within datasets, hindering effective training of transformer models. Strategies such as data augmentation through synthetic data generation and weakly supervised learning approaches may prove beneficial in addressing these challenges. Moreover, integrating additional data modalities, such as metadata commonly found in medical imaging (e.g., patient medical history, demographics), could further enhance SOD performance. By enriching the dataset with supplementary information, models may gain a more comprehensive understanding of the objects of interest, thereby improving detection accuracy and robustness.

## 6 Conclusion

This survey article reviewed over 60 research papers that focus on the development of transformers for the task of SOD, including both purely transformer-based and hybrid techniques that integrate CNNs. These techniques have been examined from seven different perspectives: object representation, fast attention mechanisms for high-resolution or multi-scale feature maps, architecture and block modifications, spatio-temporal information, improved feature representation, auxiliary techniques, and fully transformer-based detection. Each of these categories includes several SOTA techniques, each with its own set of advantages. We also compared these transformer-based approaches to CNN-based frameworks, discussing the similarities and differences between the two. Furthermore, for a range of vision applications, we introduced well-established datasets that serve as benchmarks for future research. Additionally, 12 datasets that have been used in SOD applications are discussed in detail, providing convenience for future research efforts. In future research, the unique challenges associated with the detection of small objects in each application could be explored and addressed. Fields like medical imaging and underwater image analysis stand to gain significantly from the use of transformer models. Additionally, rather than increasing the complexity of transformers using larger models, alternative strategies could be explored to boost performance.

## 7 Acknowledgment

We thank Likun Cai for providing the detection results for CBNet v2 in test images given in Figures 11 and 12.

## References

[1] 2020. Udacity self-driving car driving data, 2017 transformer. Retrieved Sep 2023 from https://github.com/udacity/self-driving-car/tree/master/annotations

[2] Josh Beal, Eric Kim, Eric Tzeng, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. 2020. Toward transformer-based object detection. Retrieved from https://arxiv.org/abs/2012.09958

[3] Alexey Bochkovskiy et al. 2020. Yolov4: Optimal speed and accuracy of object detection. Retrieved from https://arxiv.org/abs/2004.10934

[4] Likun Cai, Zhi Zhang, Yi Zhu, Li Zhang, Mu Li, and Xiangyang Xue. 2022. Bigdetection: A large-scale benchmark for improved object detector pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4777–4787.

[5] Zhaowei Cai and Nuno Vasconcelos. 2021. Cascade R-CNN: High quality object detection and instance segmentation. *TPAMI* 43, 5 (2021), 1483–1498.

[6] Xipeng Cao, Peng Yuan, Bailan Feng, and Kun Niu. 2022. CF-DETR: Coarse-to-fine transformers for end-to-end object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 185–193.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *ECCV 2020: Proceedings of the 16th European Conference on Computer Vision, Part I 16.* Springer, 213–229.

[8] Dong Chen, Duoqian Miao, and Xuerong Zhao. 2023. Hyneter: Hybrid network transformer for object detection. In *ICASSP 2023-Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 1–5.

[9] Gangqi Chen, Zhaoyong Mao, Kai Wang, and Junge Shen. 2023. HTDet: A hybrid transformer-based approach for underwater small object detection. *Remote Sensing* 15, 4 (2023), 1076.

[10] Juanjuan Chen, Hansheng Hong, Bin Song, Jie Guo, Chen Chen, and Junjie Xu. 2023. MDCT: Multi-kernel dilated convolution and transformer for one-stage object detection of remote sensing images. *Remote Sensing* 15, 2 (2023), 371.

[11] Peixian Chen, Mengdan Zhang, Yunhang Shen, Kekai Sheng, Yuting Gao, Xing Sun, Ke Li, and Chunhua Shen. 2022. Efficient decoder-free object detection with transformers. In *ECCV 2022: Proceedings of the 17th European Conference on Computer Vision, Part X.* Springer, 70–86.

[12] Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Kun Yao, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. 2023. Group DETR: Fast DETR training with group-wise one-to-many assignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 6633–6642.

[13] Qiang Chen, Jian Wang, Chuchu Han, Shan Zhang, Zexian Li, Xiaokang Chen, Jiahui Chen, Xiaodi Wang, Shuming Han, Gang Zhang, Haocheng Feng, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2022. Group DETR v2: Strong object detector with encoder-decoder pretraining. Retrieved from https://arxiv.org/abs/2211.03594

[14] Xiaokang Chen, Fangyun Wei, Gang Zeng, and Jingdong Wang. 2022. Conditional DETR v2: Efficient detection transformer with box queries. Retrieved from https://arxiv.org/abs/2207.08914

[15] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. 2021. Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 8126–8135.

[16] Yihong Chen, Zheng Zhang, Yue Cao, Liwei Wang, Stephen Lin, and Han Hu. 2020. Reppoints v2: Verification meets regression for object detection. *Advances in Neural Information Processing Systems* 33 (2020), 5621–5631.

[17] Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. 2023. Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 11 (2023), 13467–13488.

[18] Gong Cheng, Peicheng Zhou, and Junwei Han. 2016. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* 54, 12 (2016), 7405–7415.

[19] Cheng Chi, Fangyun Wei, and Han Hu. 2020. Relationnet++: Bridging visual representations for object detection via transformer decoder. *Advances in Neural Information Processing Systems* 33 (2020), 13564–13574.

[20] Se Woon Cho, Na Rae Baek, Min Cheol Kim, Ja Hyung Koo, Jong Hyun Kim, and Kang Ryoung Park. 2018. Face detection in nighttime images using visible-light camera sensors with two-step faster region-based convolutional neural network. *Sensors* 18, 9 (2018), 2995.

[21] Angelo Coluccia, Alessio Fascista, Arne Schumann, Lars Sommer, Anastasios Dimou, Dimitrios Zarpalas, Fatih Cagatay Akyon, Ogulcan Eryuksel, Kamil Anil Ozfuttu, Sinan Onur Altinuc, et al. 2021. Drone-vs-bird detection challenge at IEEE AVSS2021. In *Proceedings of the 2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS).* IEEE, 1–8.

[22] Yiming Cui. 2023. Feature aggregated queries for transformer-based video object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 6365–6376.

[23] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun . 2016. R-FCN: Object detection via region-based fully convolutional networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS)* .

[24] Linhui Dai, Hong Liu, Hao Tang, Zhiwei Wu, and Pinhao Song. 2023. Ao2-detr: Arbitrary-oriented object detection transformer. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 5 (2023), 2342–2356.

[25] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. 2021. Dynamic DETR: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 2988–2997.

[26] Jiangang Ding, Wei Li, Lili Pei, Ming Yang, Chao Ye, and Bo Yuan. 2023. Sw-YoloX: An anchor-free detector based transformer for sea surface object detection. *Expert Systems with Applications* 217 (2023), 119560.

[27] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. 2019. Learning RoI transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2849–2858.

[28] Tonghe Ding, Kaili Feng, Yanjun Wei, Yu Han, and Tianping Li. 2023. DeoT: An end-to-end encoder-only Transformer object detector. *Journal of Real-Time Image Processing* 20, 1 (2023), 1.

[29] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. Retrieved from https://arxiv.org/abs/2010.11929

[30] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. 2023. CenterNet++ for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 5 (2023), 3509–3521.

[31] Shikha Dubey, Farrukh Olimov, Muhammad Aasim Rafique, and Moongu Jeon. 2022. Improving small objects detection using transformer. *Journal of Visual Communication and Image Representation* 89 (2022), 103620.

[32] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. 2021. You only look at one sequence: Rethinking transformer in vision through object detection. *Advances in Neural Information Processing Systems* 34 (2021), 26183–26197.

[33] Masato Fujitake and Akihiro Sugimoto. 2022. Video sparse transformer with attention-guided memory for video object detection. *IEEE Access* 10 (2022), 65886–65900.

[34] Yuan Gao, Hui Shen, Donghong Zhong, Jian Wang, Zeyu Liu, Ti Bai, Xiang Long, and Shilei Wen. 2019. A solution for densely annotated large scale object detection task. (2019).

[35] Ross Girshick. 2015. Fast R-CNN. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV).* 1440–1448.

[36] Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. 2019. Precise detection in densely packed scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 5227–5236.

[37] Hang Gong, Tingkui Mu, Qiuxia Li, Haishan Dai, Chunlai Li, Zhiping He, Wenjing Wang, Feng Han, Abudusalamu Tuniyazi, Haoyang Li, et al. 2022. Swin-transformer-enabled YOLOv5 with attention mechanism for small object detection on satellite images. *Remote Sensing* 14, 12 (2022), 2861.

[38] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. 2021. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2786–2795.

[39] Khurram Azeem Hashmi, Didier Stricker, and Muhammamd Zeshan Afzal. 2022. Spatio-temporal learnable proposals for end-to-end video object detection. In *Proceedings of the 33rd British Machine Vision Conference.*

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *TPAMI* 37, 9 (2015), 1904–1916.

[41] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick . 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* 2961–2969.

[42] Liqiang He and Sinisa Todorovic. 2022. DESTR: Object detection with split transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 9377–9386.

[43] Lu He, Qianyu Zhou, Xiangtai Li, Li Niu, Guangliang Cheng, Xiao Li, Wenxuan Liu, Yunhai Tong, Lizhuang Ma, and Liqing Zhang. 2021. End-to-end video object detection with spatial-temporal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia.* 1507–1516.

[44] Suzana Herculano-Houzel. 2009. The human brain in numbers: A linearly scaled-up primate brain. *Frontiers in Human Neuroscience* 3 (2009), 31.

[45] Brian KS Isaac-Medina, Chris G. Willcocks, and Toby P. Breckon. 2022. Multi-view vision transformers for object detection. In *Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR).* IEEE, 4678–4684.

[46] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, ChristopherSTAN, Liu Changyu, Laughing, Adam Hogan, lorenzomammana, tkianai, et al. 2020. yolov5. *Code repository.* Retrieved Sep. 2023 from https://github.com/ultralytics/yolov5. (2020).

[47] Sehoon Kim, Coleman Hooper, Thanakul Wattanawong, Minwoo Kang, Ruohan Yan, Hasan Genc, Grace Dinh, Qijing Huang, Kurt Keutzer, Michael W. Mahoney, Yakun Sophia Shao, and Amir Gholami. 2023. Full stack optimization of transformer inference: A survey. Retrieved from https://arxiv.org/abs/2302.14017

[48] Hei Law and Jia Deng. 2018. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV).* 734–750.

[49] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, Yiduo Li, Bo Zhang, Yufei Liang, Linyuan Zhou, Xiaoming Xu, Xiangxiang Chu, Xiaoming Wei, and Xiaolin Wei. 2022. YOLOv6: A single-stage object detection framework for industrial applications. arXiv:2209.02976. Retrieved from https://arxiv.org/abs/2209.02976

[50] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. 2022. DN-DETR: Accelerate DETR training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 13619–13627.

[51] Johann Li, Guangming Zhu, Cong Hua, Mingtao Feng, Basheer Bennamoun, Ping Li, Xiaoyuan Lu, Juan Song, Peiyi Shen, Xu Xu, et al. 2023. A systematic collection of medical image datasets for deep learning. *ACM Computing Surveys* 56, 5 (2023), 51. https://doi.org/10.1145/3615862

[52] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. 2020. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing* 159 (2020), 296–307.

[53] Qingyun Li, Yushi Chen, and Ying Zeng. 2022. Transformer with transfer CNN for remote-sensing-image object detection. *Remote Sensing* 14, 4 (2022), 984.

[54] Sijia Li, Furkat Sultonov, Jamshid Tursunboev, Jun-Hyun Park, Sangseok Yun, and Jae-Mo Kang. 2022. Ghostformer: A GhostNet-based two-stage transformer for small object detection. *Sensors* 22, 18 (2022), 6939.

[55] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. 2019. Scale-aware trident networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6054–6063.

[56] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. 2022. Exploring plain vision transformer backbones for object detection. In *ECCV 2022: Proceedings of the 17th European Conference on Computer Vision, Part IX*. Springer, 280–296.

[57] Tingting Liang, Xiaojie Chu, Yudong Liu, Yongtao Wang, Zhi Tang, Wei Chu, Jingdong Chen, and Haibin Ling. 2022. Cbnet: A composite backbone network architecture for object detection. *IEEE Transactions on Image Processing* 31 (2022), 6893–6906.

[58] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI Open* 3 (2022), 111–132.

[59] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2117–2125.

[60] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2980–2988.

[61] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*. Springer, 740–755.

[62] Wei-Hong Lin, Jia-Xing Zhong, Shan Liu, Thomas Li, and Ge Li. 2020. RoIMix: Proposal-fusion among multiple images for underwater object detection. In *ICASSP 2020-Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2588–2592.

[63] Chang Liu, Sheng Xu, and Baochang Zhang. 2021. Aerial small object tracking with transformers. In *Proceedings of the 2021 IEEE International Conference on Unmanned Systems (ICUS)*. IEEE, 954–959.

[64] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. 2023. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *ICLR*.

[65] Ting Liu, Yao Zhao, Yunchao Wei, Yufeng Zhao, and Shikui Wei. 2019. Concealed object detection for activate millimeter wave image. *IEEE Transactions on Industrial Electronics* 66, 12 (2019), 9909–9917.

[66] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 21–37.

[67] Yang Liu, Peng Sun, Nickolas Wergeles, and Yi Shang. 2021. A survey and performance evaluation of deep learning methods for small object detection. *Expert Systems with Applications* 172 (2021), 114602.

[68] Zhigang Liu, Juan Du, Feng Tian, and Jiazheng Wen. 2019. MR-CNN: A multi-scale region-based convolutional neural network for small traffic sign recognition. *IEEE Access* 7 (2019), 57120–57128.

[69] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

[70] Wanjie Lu, Chaozhen Lan, Chaoyang Niu, Wei Liu, Liang Lyu, Qunshan Shi, and Shiju Wang. 2023. A CNN-transformer hybrid model based on CSWin transformer for UAV image object detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16 (2023), 1211–1231.

[71] Teli Ma, Mingyuan Mao, Honghui Zheng, Peng Gao, Xiaodi Wang, Shumin Han, Errui Ding, Baochang Zhang, and David Doermann. 2021. Oriented object detection with transformer. Retrieved from https://arxiv.org/abs/2106.03146

[72] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. 2022. Class-agnostic Object detection with multi-modal transformer. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)*. Springer.

[73] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. 2022. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8844–8854.

[74] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. 2021. Conditional DETR for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3651–3660.

[75] Matthias Mueller, Neil Smith, and Bernard Ghanem. 2016. A benchmark and simulator for UAV tracking. In *ECCV 2016: Proceedings of the 14th European Conference on Computer Vision, Part I 14*. Springer, 445–461.

[76] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. 2020. Imbalance problems in object detection: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 10 (2020), 3388–3415.

[77] Jeffrey Ouyang-Zhang, Jang Hyun Cho, Xingyi Zhou, and Philipp Krähenbühl. 2022. NMS strikes back. Retrieved from https://arxiv.org/abs/2212.06137

[78] F. Ozge Unel, Burak O. Ozkalayci, and Cevahir Cigla. 2019. The power of tiling for small object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.*

[79] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. 2019. Libra R-CNN: Towards balanced learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 821–830.

[80] Zhiliang Peng, Zonghao Guo, Wei Huang, Yaowei Wang, Lingxi Xie, Jianbin Jiao, Qi Tian, and Qixiang Ye. 2023. Conformer: Local features coupling global representations for recognition and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 8 (2023), 9454–9468.

[81] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. 2021. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 367–376.

[82] Shima Rashidi, Krista Ehinger, Andrew Turpin, and Lars Kulik. 2020. Optimal visual search based on a model of target detectability in natural images. *Advances in Neural Information Processing Systems* 33 (2020), 9288–9299.

[83] Shima Rashidi, Ruwan Tennakoon, Aref Miri Rekavandi, Papangkorn Jessadatavornwong, Amanda Freis, Garret Huff, Mark Easton, Adrian Mouritz, Reza Hoseinnezhad, and Alireza Bab-Hadiashar. 2025. IT-RUDA: Information theory assisted robust unsupervised domain adaptation. *ACM Transactions on Intelligent Systems and Technology* (2025).

[84] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 779–788.

[85] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, faster, stronger. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 7263–7271.

[86] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. Retrieved from https://arxiv.org/abs/1804.02767

[87] Aref Miri Rekavandi, Abd-Krim Seghouane, and Robin J. Evans. 2021. Robust subspace detectors based on $\alpha$-divergence with application to detection in imaging. *IEEE Transactions on Image Processing* 30 (2021), 5017–5031.

[88] Aref Miri Rekavandi, Lian Xu, Farid Boussaid, Abd-Krim Seghouane, Stephen Hoefs, and Mohammed Bennamoun. 2025. A guide to image-and video-based small object detection using deep learning: Case study of maritime surveillance. *IEEE Transactions on Intelligent Transportation Systems* 26, 3 (2025), 2851–2879.

[89] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the 29th International Conference on Neural Information Processing Systems (NeurIPS)* .

[90] Si-Dong Roh and Ki-Seok Chung. 2022. DAFA: Diversity-aware feature aggregation for attention-based video object detection. *IEEE Access* 10 (2022), 93453–93463.

[91] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115 (2015), 211–252.

[92] Ruoyue Shen, Nakamasa Inoue, and Koichi Shinoda. 2023. Text-guided object detector for multi-modal video question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.* 1032–1042.

[93] Yuntao Shou, Tao Meng, Wei Ai, Canhao Xie, Haiyan Liu, and Yina Wang. 2022. Object detection in medical images based on hierarchical transformer and mask mechanism. *Computational Intelligence and Neuroscience* 2022 (2022), 1–12.

[94] Hwanjun Song, Deqing Sun, Sanghyuk Chun, Varun Jampani, Dongyoon Han, Byeongho Heo, Wonjae Kim, and Ming-Hsuan Yang. 2022. ViDT: An efficient and effective fully transformer-based object detector. In *Proceedings of the International Conference on Learning Representations (ICLR).*

[95] Russell Stewart, Mykhaylo Andriluka, and Andrew Y. Ng. 2016. End-to-end people detection in crowded scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2325–2333.

[96] Peng Sun, Ting Liu, Xiaotong Chen, Shiyin Zhang, Yao Zhao, and Shikui Wei. 2022. Multi-source aggregation transformer for concealed object detection in millimeter-wave images. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 9 (2022), 6148–6159.

[97] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris M. Kitani. 2021. Rethinking transformer-based set prediction for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 3611–3620.

[98] Yudi Tang, Bing Wang, Wangli He, and Feng Qian. 2023. Pointdet++: An object detection framework based on human local features with transformer encoder. *Neural Comput & Applic* 35 (2023), 10097–10108.

[99] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9627–9636.

[100] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *Proceedings of the International Conference on Machine Learning*. PMLR, 10347–10357.

[101] Leon Amadeus Varga and Andreas Zell. 2021. Tackling the background bias in sparse object detection via cropped windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2768–2777.

[102] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017), 1–11.

[103] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7464–7475.

[104] Han Wang, Jun Tang, Xiaodong Liu, Shanyan Guan, Rong Xie, and Li Song. 2022. PTSEFormer: Progressive temporal-spatial enhanced transformer towards video object detection. In *ECCV 2022: Proceedings of the 17th European Conference on Computer Vision, Part VIII*. Springer, 732–747.

[105] Jinwang Wang, Chang Xu, Wen Yang, and Lei Yu. 2021. A normalized Gaussian Wasserstein distance for tiny object detection. Retrieved from https://arxiv.org/abs/2110.13389

[106] Liya Wang and Alex Tien. 2023. Aerial image object detection with vision transformer detector (ViTDet). In *IGARSS 2023-Proceedings of the 2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 6450–6453.

[107] Wen Wang, Yang Cao, Jing Zhang, and Dacheng Tao. 2022. Fp-detr: Detection transformer advanced by fully pre-training. In *Proceedings of the International Conference on Learning Representations*.

[108] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. 2022. Anchor DETR: Query design for transformer-based object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* .

[109] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. 2019. Sequence level semantics aggregation for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9217–9225.

[110] Jialian Wu, Chunluan Zhou, Qian Zhang, Ming Yang, and Junsong Yuan. 2020. Self-mimic learning for small-scale pedestrian detection. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2012–2020.

[111] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. 2018. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3974–3983.

[112] Ruiyang Xia, Guoquan Li, Zhengwen Huang, Yu Pang, and Man Qi. 2022. Transformers only look once with nonlinear combination for real-time object detection. *Neural Computing and Applications* 34, 15 (2022), 12571–12585.

[113] Shoukun Xu, Jianan Gu, Yining Hua, and Yi Liu. 2023. DKTNet: Dual-key transformer network for small object detection. *Neurocomputing* 525 (2023), 29–41.

[114] Wenyu Xu, Chaofan Zhang, Qi Wang, and Pangda Dai. 2022. FEA-swin: Foreground enhancement attention swin transformer network for accurate UAV-based dense object detection. *Sensors* 22, 18 (2022), 6993.

[115] Xiangkai Xu, Zhejun Feng, Changqing Cao, Mengyuan Li, Jin Wu, Zengyan Wu, Yajie Shang, and Shubing Ye. 2021. An improved swin transformer-based model for remote sensing object detection and instance segmentation. *Remote Sensing* 13, 23 (2021), 4779.

[116] Jingqian Xue, Da He, Mengwei Liu, and Qian Shi. 2022. Dual network structure with interweaved global-local feature hierarchy for transformer-based object detection in remote sensing image. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15 (2022), 6856–6866.

[117] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M. Summers. 2018. DeepLesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging* 5, 3 (2018), 036501–036501.

[118] Hao Yang, Zihan Yang, Anyong Hu, Che Liu, Tie Jun Cui, and Jungang Miao. 2023. Unifying convolution and transformer for efficient concealed object detection in passive millimeter-wave images. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 8 (2023), 3872–3887.

[119] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. 2019. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9657–9666.

[120] Tao Ye, Wenyang Qin, Zongyang Zhao, Xiaozhi Gao, Xiangpeng Deng, and Yu Ouyang. 2023. Real-time object detection network in UAV-vision based on CNN and transformer. *IEEE Transactions on Instrumentation and Measurement* 72 (2023), 1–13.

[121] Dmitry Yudin and Dmitry Slavioglo. 2018. Usage of fully convolutional network with clustering for traffic light detection. In *Proceedings of the 2018 7th Mediterranean Conference on Embedded Computing (MECO)*. IEEE, 1–6.

[122] Kai Zeng, Qian Ma, Jiawen Wu, Sijia Xiang, Tao Shen, and Lei Zhang. 2022. NLFFTNet: A non-local feature fusion transformer network for multi-scale object detection. *Neurocomputing* 493 (2022), 15–27.

[123] Chi Zhang, Lijuan Liu, Xiaoxue Zang, Frederick Liu, Hao Zhang, Xinying Song, and Jindong Chen. 2022. DETR++: Taming your multi-scale detection transformer. Retrieved from https://arxiv.org/abs/2206.02977

[124] Hongkai Zhang, Hong Chang, Bingpeng Ma, Naiyan Wang, and Xilin Chen. 2020. Dynamic R-CNN: Towards high quality object detection via dynamic training. In *ECCV 2020: Proceedings of the 16th European Conference on Computer Vision, Part XV 16*. Springer, 260–275.

[125] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. 2023. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *Proceedings of the 11th International Conference on Learning Representations*.

[126] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. 2020. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9759–9768.

[127] Yuanlin Zhang, Yuan Yuan, Yachuang Feng, and Xiaoqiang Lu. 2019. Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing* 57, 8 (2019), 5535–5548.

[128] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. 2023. TransVOD: End-to-end video object detection with spatial-temporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 6 (2023), 7853–7869.

[129] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. 2019. Objects as points. Retrieved from https://arxiv.org/abs/1904.07850

[130] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2021. Deformable DETR: Deformable transformers for end-to-end object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)* .

[131] Yuan Zhu, Qingyuan Xia, and Wen Jin. 2022. SRDD: A lightweight end-to-end object detection with transformer. *Connection Science* 34, 1 (2022), 2448–2465.

[132] Zhuofan Zong, Guanglu Song, and Yu Liu. 2023. DETRs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6748–6758.

Received 26 December 2024; revised 11 May 2025; accepted 24 July 2025