

# Class-agnostic Object Detection with Multi-modal Transformer

Muhammad Maaz<sup>1\*</sup>, Hanoona Rasheed<sup>1\*</sup>, Salman Khan<sup>1,2</sup>, Fahad Shahbaz Khan<sup>1,3</sup>, Rao Muhammad Anwer<sup>1,4</sup>, and Ming-Hsuan Yang<sup>5,6,7</sup>

<sup>1</sup>Mohamed bin Zayed University of AI    <sup>2</sup>Australian National University  
<sup>3</sup>Linköping University    <sup>4</sup>Aalto University    <sup>5</sup>University of California, Merced  
<sup>6</sup>Yonsei University    <sup>7</sup>Google Research

**Abstract.** What constitutes an object? This has been a long-standing question in computer vision. Towards this goal, numerous learning-free and learning-based approaches have been developed to score *objectness*. However, they generally do not scale well across new domains and novel objects. In this paper, we advocate that existing methods lack a top-down supervision signal governed by human-understandable semantics. For the first time in literature, we demonstrate that Multi-modal Vision Transformers (MViT) trained with aligned image-text pairs can effectively bridge this gap. Our extensive experiments across various domains and novel objects show the state-of-the-art performance of MViTs to localize generic objects in images. Based on the observation that existing MViTs do not include multi-scale feature processing and usually require longer training schedules, we develop an efficient MViT architecture using multi-scale deformable attention and late vision-language fusion. We show the significance of MViT proposals in a diverse range of applications including open-world object detection, salient and camouflage object detection, supervised and self-supervised detection tasks. Further, MViTs can adaptively generate proposals given a specific language query and thus offer enhanced interactability. Code: <https://git.io/J1HPY>.

**Keywords:** Object detection, Class-agnostic, Vision Transformers

## 1 Introduction

The recent years have witnessed significant advances in object detection (OD) [42] based on developments of large-scale annotated datasets and carefully designed deep learning models. Notably, efforts have been made to tackle more difficult cases such as universal OD [67], long-tailed object distribution modeling [19], open-vocabulary [78] and open-world OD [28]. In contrast, little progress has been made towards a seemingly simpler task of class-agnostic OD [1] in recent years. In the era of fully trainable pipelines, class-agnostic OD is still often

---

\*Equal contribution

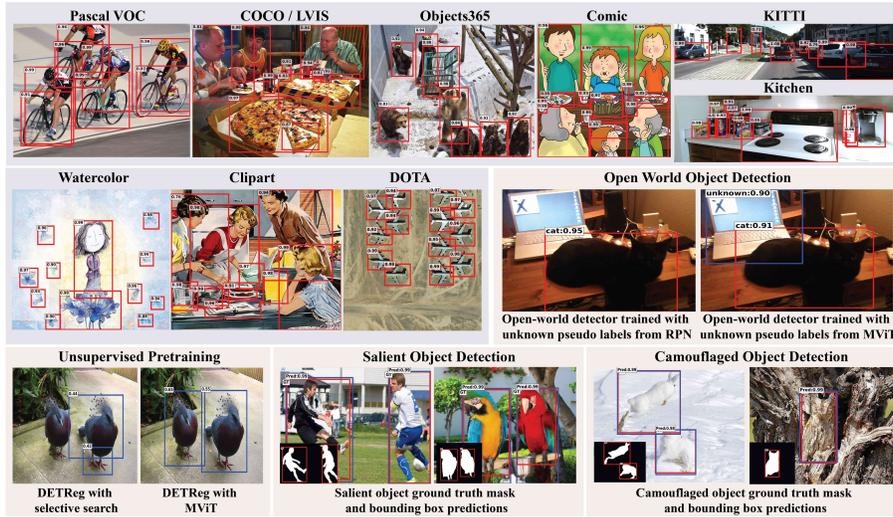


Fig. 1: We show that Multi-modal Vision Transformers (MViTs) excel at Class-agnostic OD across multiple domains: natural images [14,40,17,18], satellite images [72], sketches, cartoons and paintings [26] ( gray background ). The MViTs perform well on diverse datasets (with many classes *e.g.*, LVIS, Object365) using intuitive natural language text queries (*e.g.*, all objects). Further, class-agnostic detectors (MViTs) can be applied to several downstream applications ( pearl background ). In Open-world OD [28], unknown pseudo-labels generated using MDETR [29] can improve novelty detection. For unsupervised object localization, replacing Selective Search proposals [64] in DETReg [3] pretraining with only top-30 MViT proposals leads to improved localization. For Salient and Camouflaged OD, task specific text queries can help perform competitively against fully supervised models without any task specific tuning. Overall, MViTs achieve the state-of-the-art results on various downstream applications.

approached using typical bottom-up approaches such as Selective Search [64], EdgeBox [84], DeepMask [49] and MCG [52].

Despite being an apparently simpler problem in terms of the two-way classification space, the class-agnostic OD task is indeed challenging from the representation learning perspective. The main challenge is to model the vast diversity of *all* valid object classes and delineate such a diverse group from the *background* class which itself has vague semantic definition [2]. Our experiments indicate that this intrinsic complexity of the task makes it difficult to design fully trainable class-agnostic OD models that can work across domains and for novel unseen objects. Although the bottom-up approaches offer proposals for generic objects, they come at the cost of a prohibitively large number of candidate boxes, low-precision, lack of semantic understanding and slow processing, making them less scalable to generic operation in the wild. More recently, self-supervised learning frameworks – based on both ViTs [3,11] and CNNs [74,73] – have focused on promoting better localization of generic objects, however they still show modest performance on class-agnostic OD [3]. Our intuition is that *top-down supervisory*

*signals* are necessary to resolve the ambiguous nature of class-agnostic OD task, which is precisely what is missing from the aforementioned approaches.

In this paper, we bring out the capacity of recent Multi-modal Vision Transformers (MViTs) to propose generic class-agnostic OD across different domains. The high-level information provided by the language descriptions helps learn fairly generalizable properties of universal object categories. In turn, the MViTs perform exceptionally well compared to uni-modal object detectors trained for generic object detection as well as the typical bottom-up object proposal generation schemes. Due to the multi-modal nature of these models, we design language-driven queries to discover valid objects in a human-understandable format that can be adapted to explore varied aspects of the object semantic space. With the state-of-the-art performance, an ensuing question is to explore the root cause of such generalization for the ‘*concept of objects*’ embedded in MViTs. Through a series of systematic experiments, we find that it is the language skeleton/structure (rather than the lexicon itself) that defines this strong understanding of generic object definition within MViT models. As an interesting example, when the MViT is trained without actual captions, but just the bounding boxes corresponding to a natural language description, the model still demonstrates strong class-agnostic OD generalization. These insights on the interactive class-agnostic OD mechanism can be deployed in several *downstream* tasks such as novel object discovery, saliency detection, self-supervised learning and open-world detection. The main highlights of this work include:

- We demonstrate the state-of-the-art performance of pre-trained MViTs [29,20] towards class-agnostic OD via a set of human-understandable natural language queries. We also develop an efficient and flexible MViT model, *Multiscale Attention ViT with Late fusion* (MAVL), which performs better in locating generic objects as compared to existing MViTs (Secs. 2 and 3).
- We benchmark generalization of MViT based OD models on diverse domains *e.g.*, natural images, sketches, cartoons, satellite images, paintings and show their favorable performance compared to existing class-agnostic OD models (bottom-up approaches, CNN and ViT based uni-modal pipelines) (Sec. 3).
- Our class-agnostic detectors can benefit various down-stream applications: Open-world OD, Salient OD, Camouflaged OD and Self-supervised learning. Furthermore, when these proposals are combined with RPN proposals in two-stage detectors, it can lead to overall performance improvements due to their rich top-down semantic understanding of the image content (Sec. 4).
- Through an extensive set of systematic experiments, we analyze the factors that majorly contribute to the improved performance of MViTs (Sec. 5).

## 2 Multi-modal ViTs

In this work, we bring out the generalization capacity of Multi-modal ViTs (MViT) to tackle generic OD. The capability of relating natural language with visual features helps MViTs to generalize to novel concepts, achieving state-of-the-art results on class-agnostic OD using human-understandable text queries

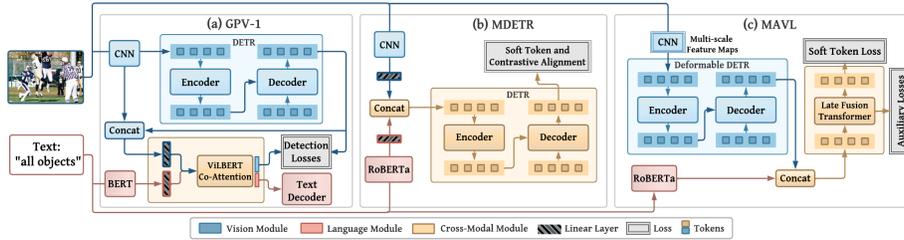


Fig. 2: Architecture overview of MViTs used in this work – GPV-1 [20], MDETR [29] and MAVL (ours). GPV-1 takes image along with a task description as input and outputs relevant region boxes and text. MDETR uses soft token prediction and contrastive alignment in latent space for cross-conceptualization using aligned image-text pairs. MAVL utilizes multi-scale image features with multi-scale deformable attention module (MSDA), and uses late-fusion strategy for vision-language fusion.

(e.g., ‘all objects/entities’). Before a detailed analysis, we provide background on MViTs and propose Multiscale Attention ViT with Late fusion (MAVL).

(a) **GPV**: Gupta *et al.* proposed GPV-I [20], a unified architecture for multi-task learning, where the task is inferred from the text prompt. It takes an image and a task description as input and outputs text with the corresponding bounding boxes. This model uses pretrained BERT [12] to encode the text, concatenates it with the region descriptors from DETR [5] and passes it to ViLBERT [44] co-attention layers for cross-modal conceptualization. It predicts relevance scores for each predicted bounding box indicating the importance of the region for the prompted task. An output text decoder conditioned on the relevance scores is used for better cross-modal understanding (Fig. 2 (a)). GPV is trained on data from five different vision-language tasks.

(b) **MDETR**: Kamath *et al.* [29] proposed a modulated transformer trained to detect objects in an image conditioned on a text query. In MDETR, visual and text features are extracted from a convolutional backbone (*e.g.*, ResNet-101 [23] or EfficientNet [63]) and a language model (RoBERTa [43]) respectively. These features are then concatenated and passed to the DETR [5] model for detection (Fig. 2 (b)). MDETR uses soft token prediction and contrastive alignment in latent space for addressing text-conditioned object detection. In soft token prediction, a uniform probability distribution is predicted over all text tokens for each detected object. In contrastive alignment, the embedded object queries from decoder are aligned with the text representation from encoder. This multi-modal alignment makes the object embeddings closer to the corresponding text embeddings in feature space. The model is pre-trained with 1.3M image-text pairs and achieves the state-of-the-art results on various vision-language downstream tasks including VQA, referring expression and phrase grounding.

(c) **MAVL**: We develop a new multimodal architecture called Multi-scale Attention ViT with Late fusion (MAVL) that improves the class-agnostic OD performance of MDETR using multi-scale spatial context and deformable attention making it efficient to train. Fig. 2 (c) shows our overall design. Below, we high-

light the main features of MAVL:

–*Multi-scale Deformable Attention (MSDA)*. MDETR [29] finds it challenging to scale to high-resolution feature maps due to a fixed self-attention design. Further, it operates on a specified spatial scale which can be sub-optimal for small objects. Our design calculates attention at multiple scales to incorporate better contextual information. However, multiple scales can increase the computational cost, therefore we use Deformable Attention proposed in [83] that employs multi-scale feature processing and dynamically attends to relevant pixel locations for context aggregation. Specifically, it samples a small set of keys around a reference (query) image location. The sparse key sampling in MSDA achieves linear complexity with respect to the size of the image feature maps.

–*Late Multi-modal Fusion*. MSDA module utilizes the spatial structure of an image to sparsely sample keys for each query point. Following the MDETR strategy of concatenating text embeddings with flattened features would destroy the spatial structure of an image. Hence, we fuse text in MAVL model after the images are processed through the Def-DETR encoder-decoder architecture using a *late fusion* mechanism. Specifically, the object query representations from the deformable decoder are concatenated with the text embeddings, and passed through a series of six transformer self-attention (SA) blocks. This design choice is inspired by the recent vision-language fusion works [44,60,62,61]. Using the training procedure of [5], the output head is applied after each SA block and the total loss is calculated by adding all auxiliary losses. We note that no explicit contrastive alignment of object query representation and encoded text is required in our approach. Our experiments show fast convergence (only *half* iterations) and competitive performance of MAVL against MDETR (Tables 1, 2).

–*Implementation Details*. Similar to MDETR [29], we train MAVL on approx. 1.3M aligned image-text pairs, using images from Flickr30k [51], MS-COCO (2014) [40] and Visual Genome (VG) [32]. The corresponding annotations are taken from Flickr entities, RefCOCO/+g referring expression [30], VG regions and GQA [25]. In the onward discussion, we refer to this dataset as *Large-scale Modulated Detection* (LMDet) dataset. All MDETR and MAVL models are trained with ImageNet-1K [55] pretrained ResNet-101 [23]. Our MAVL *converges in 20 epochs* (MDETR requires 40 epochs) on LMDet using the same hyperparameters as in MDETR. See Appendix A.1 for more details.

### 3 Multi-modal ViTs as Generic Detectors

The class-agnostic OD seeks to differentiate between generic objects and background in images. This task involves learning the notion of *objectness*. Existing approaches typically explore low-level visual cues (i.e. superpixels, edges, etc.) or directly learn the mapping between images and generic object locations using fully trainable pipelines learned with bounding box annotations [64,84,27,3]. We note that these procedures lack high-level semantic information necessary to relate objects across diverse scenes to derive a comprehensive and general notion of universal objects. In this work, We explore the class-agnostic OD capacity of

Table 1: Class-agnostic OD results of MViTs in comparison with bottom-up approaches (row 3-5) and uni-modal detectors (row 6-8) trained to localize generic objects. Bottom row shows gain of MAVL over the best uni-modal method. In general, MViTs achieve state-of-the-art performance using intuitive text queries (details in Sec. 4.1).

Dataset → Model ↓	Pascal-VOC		COCO		KITTI		Objects365		LVIS	
	AP50	R50								
Edge Boxes	0.08	7.14	0.09	5.16	0.09	6.58	0.07	3.27	0.05	3.00
Selective Search	0.32	21.4	0.27	12.7	0.03	4.85	0.38	10.7	0.24	9.31
Deep Mask	5.92	40.4	2.16	19.2	1.33	15.5	1.31	14.5	0.51	8.17
Faster-RCNN	42.9	85.8	26.4	58.7	23.5	53.2	24.8	54.6	8.91	35.6
RetinaNet	43.2	86.6	24.6	59.1	30.4	57.6	24.3	54.8	8.57	35.7
Def-DETR	30.1	81.0	20.0	53.5	23.7	55.0	17.0	45.9	6.60	30.7
GPV-I	61.9	91.1	38.0	64.4	43.0	64.4	25.6	50.2	9.18	27.5
MDETR	66.0	90.1	40.7	62.2	46.7	<b>67.2</b>	30.4	54.0	10.7	32.8
MAVL (Ours)	<b>68.6</b>	<b>91.3</b>	<b>43.6</b>	<b>65.0</b>	<b>48.2</b>	63.5	<b>33.2</b>	<b>57.9</b>	<b>11.7</b>	<b>37.0</b>
	+25.4	+4.7	+19.0	+5.9	+17.8	+5.9	+8.4	+3.1	+2.8	+1.3

MViTs trained using aligned image-text pairs (Sec. 2). We observe these models can produce high quality object proposals by using intuitive text queries like ‘all objects’ and ‘all entities’. This demonstrates their capability to relate natural language with visual concepts to model generic objectness, enabling them to discover novel categories and generalize across different domains while offering human interaction with intelligible text queries.

### 3.1 Class-agnostic Object Detection

**Settings:** Table 1 shows the object proposal generation performance of MViTs with the typical bottom-up approaches and the end-to-end supervised deep learning methods on five challenging natural image OD datasets (Pascal VOC [14], MS COCO [40], KITTI [17], Objects365 [56] and LVIS [19]). The bottom-up approaches considered for comparison include EdgeBoxes [84], Selective Search [64] and DeepMask [49] while Faster-RCNN [54], RetinaNet [39] and Deformable-DETR [83] are selected from the deep-learning based methods due to the state-of-the-art performance in class-aware OD. The MViTs considered are GPV-I [20] and MDETR [29] alongside our proposed MAVL (see Sec. 2 for details).

For fairness, all the uni-modal detectors considered for evaluation are trained with ResNet-101 backbone using box-level supervision on LMDet dataset. Faster-RCNN and RetinaNet follow the standard Detectron2 [70] training setting with FPN at  $1\times$  schedule. The combined detections from the text queries in Table 3 are used for evaluating MViTs (see Sec. 4.1 and Appendix A.2 for details). Moreover, images used in the evaluation *do not* have any overlap with LMDet.

**Results:** We report both average precision (AP) and Recall at IoU threshold of 0.5 using the top-50 boxes from each method. Overall, the detectors trained in class-agnostic fashion perform reasonably well on all datasets, surpassing the bottom-up methods by a large margin. Furthermore, the MViTs perform better

Table 2: Class-agnostic OD performance of MViT<sub>s</sub> in comparison with RetinaNet [39] on several out-of-domain datasets. MViT<sub>s</sub> show consistently good results on all datasets. <sup>†</sup>Proposals on DOTA [72] are generated by multi-scale inference (see Sec. A.2).

Dataset →	Kitchen		Clipart		Comic		Watercolor		DOTA <sup>†</sup>	
Model ↓	AP50	R50	AP50	R50	AP50	R50	AP50	R50	AP50	R50
RetinaNet	35.3	89.5	27.0	90.0	33.1	86.1	47.8	91.9	0.72	15.6
GPV-1	24.5	84.8	35.1	86.1	42.3	83.6	50.3	89.5	0.55	9.33
MDETR	38.4	<b>91.4</b>	44.9	90.7	55.8	<b>89.5</b>	63.6	94.3	1.94	21.8
MAVL (Ours)	<b>45.4</b>	91.0	<b>50.6</b>	<b>92.9</b>	<b>57.7</b>	89.2	<b>63.8</b>	<b>95.6</b>	<b>2.86</b>	<b>24.2</b>

than the uni-modal approaches with the use of simple human understandable natural language text queries. This performance shows MViT<sub>s</sub>’ strong understanding of language content obtained from the pretrained NLP model (BERT [12], RoBERTa [43]) along with the aligned image-text pairs used in pretraining.

For MViT<sub>s</sub>, interestingly a relatively small number of boxes match the quality achieved by a much larger proposal set from competing methods. Fig. 3a shows the recall obtained by varying the number of top object proposals for all methods on two datasets. MViT<sub>s</sub> achieve competitive recall with only top-10 proposals.

### 3.2 How well MViT<sub>s</sub> generalize?

**Generalization to New Domains:** We extend our analysis from natural image datasets (Sec. 3.1) to rule out if MViT representations are biased towards natural images, for which these models are originally trained on. To this end, we evaluate on universal OD datasets [67] belonging to five different domains (Table 2). The studied domains include indoor kitchen scenes [18], cartoon images, watercolor drawings, clipart, comics [26] and satellite/aerial images (DOTA dataset) [72]. The experiments follow the same setting as in Sec. 3.1. These results indicate the generalization capability of MViT<sub>s</sub> in comparison to the best proposal generation methods earlier evaluated in Table 1 (RetinaNet trained for class-agnostic OD).

**Generalization to Rare/Novel Classes:** With the notion of objectness, humans are capable of identifying novel and rare objects, although they may not recognize their specific category. Similarly, scalability to rare and novel classes is a desired quality of an object detector. To analyze this, the class-agnostic OD mechanism of MAVL is evaluated on rare categories from Open-Images [34] versus frequent categories and compared with Deformable DETR and Deep Mask trained for class agnostic OD. Fig. 3b indicate state-of-the-art recall on rare categories such as *lynx*, *humidifier*, and *armadillo* with as few as zero training instance. Overall, we note the model generalizes well to rare/unseen categories.

## 4 Applications and Use-cases

The high-quality class-agnostic object proposals obtained from MViT<sub>s</sub> can be helpful towards several downstream applications, as we demonstrate next.

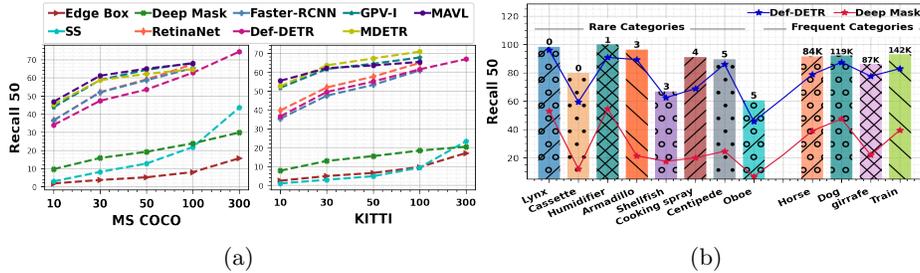


Fig. 3: (a) Effect of using different number of top-ranked boxes on multiple class-agnostic OD methods. The MViTs exhibits good recall even with only top-10 proposals. (b) MAVL class-agnostic OD performance on rarely and frequently occurring categories in LMDet. Rare categories are selected from Open Images [34]. The MAVL recall rates (represented by the bars) are compared with those of Def-DETR [83] and DeepMask [49] (represented by the lines). The numbers on top of the bars indicate the total occurrences of the category in LMDet captions. The MViT achieves good recall even for the classes with no or very few occurrences in the training dataset.

#### 4.1 Enhanced Interactability

We have observed that MViTs can generate high quality object proposals with intuitive human understandable queries such as ‘all objects’. This motivates us to explore the language semantic space of such models to construct a set of queries that can well capture the generic concept of objectness. We filter words from captions in LMDet that are semantically close to the word ‘object’ in the linguistic feature space. We then utilize these words to construct intuitive text queries such as ‘all objects’, ‘all entities’, ‘all visible entities and objects’, ‘all obscure entities and objects’, and ‘all small objects’, for exploiting the class-agnostic OD performance of MViTs. The detections from the individual text queries are combined, filtered with class-agnostic non-maximum suppression (NMS) to remove duplicate detections, and top-N boxes are selected for evaluation. We use N=50 in all of our experiments.

Table 3: Using different intuitive text queries with MAVL. Combining detections from multiple queries captures varying aspects of objectness.

Dataset → Text Query ↓	Pascal-VOC		COCO		KITTI	
	AP50	R50	AP50	R50	AP50	R50
all objects	51.3	85.5	33.3	58.4	40.2	64.0
all entities	65.2	88.4	34.6	54.6	41.9	59.5
all visible entities & objects	63.3	89.0	37.9	61.6	42.0	63.0
all obscure entities & objects	59.5	86.6	35.2	59.1	42.4	63.5
all small objects	40.0	83.9	28.9	58.9	40.4	65.2
combined detections (CD)	63.7	91.0	42.0	<b>65.0</b>	<b>48.2</b>	<b>63.5</b>
CD w/o ‘all small objects’	<b>68.6</b>	<b>91.3</b>	<b>43.6</b>	<b>65.0</b>	45.8	61.6

**Task specific queries:** The detection of small and irregular sized objects has remained a long-standing challenge. In our case, the flexible nature of MViTs facilitates using a range of human-understandable text queries. The queries can be chosen that best describe the special requirements needed in a given detection task. We demonstrate certain scenarios of how this feature can be exploited for better predictions. Fig. 4a (left) shows an interesting case of how the text query ‘all little objects’ improves recall for small objects as compared to a rather general text query. Similarly, Fig. 4a (right) indicates how the use of special

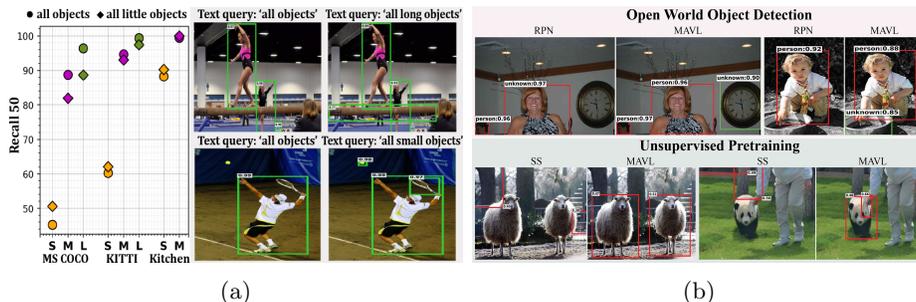


Fig. 4: (a) MAVL recall for small (S), medium (M) and large (L) objects across three datasets. The use of specific query (‘all little objects’) increases recall of small objects across different datasets (*left*). Targeted detections by the relevant text queries (*right*). (b) Visualizations of ORE [28] unknown detections when trained with RPN versus MAVL unknown pseudo-labels (*top*). Class-agnostic OD of DETReg [3] when trained using Selective Search (SS) [64] versus MAVL proposals (*bottom*).

queries like ‘all long objects’ helps improve the detection of irregular shaped objects (without any dataset specific fine-tuning!).

## 4.2 Open-world Object Detection

The open-world setting assumes a realistic paradigm where a model can experience *unknown objects* during training and inference [4,13,65,28]. The goal is to identify unknowns and incrementally learn about them as and when new annotations are provided about a subset of unknowns. This stands in contrast to generic OD where models are trained to label unknown objects as background and only focus on the known objects. Here, we explore how a generic class-agnostic OD model can help with the open-world task to identify unknowns. As a case study, we apply our approach to a recent open-world detector (ORE) [28].

– *ORE Setting*: The authors distribute the 80 COCO [40] classes in four incremental learning tasks where 20 classes have been added to the known categories in each subsequent task. At each stage, the model must learn from the given subset of 20 newly introduced known classes, should not forget the previous known classes and must be able to detect unknown classes whose labelled examples have not been provided so far as the unknowns. ORE uses Faster-RCNN [54] as the base detector, with contrastive clustering in latent space and an energy-based classification head for unknown detection. It utilizes example-replay strategy [66] for alleviating forgetting, when progressively learning the unknown categories once their labels become available.

– *Unknown Pseudo-labels with MViTs*: ORE exploits the two-stage mechanism of Faster-RCNN [54] and uses proposals from the class-agnostic region proposal network (RPN) for pseudo-labelling of unknowns. The foreground object proposals with high objectness score which do not overlap with any ground-truth are labelled as unknowns. We note that since RPN is only trained on the objects of interest, its detections are overly sparse and lead to a low recall for

Table 4: MViT proposals are used for pseudo-labelling of unknowns in ORE [28]. MAVL represents the model trained on a filtered dataset generated by *removing* all captions from LMDet listing any of the 60 unknown categories evaluated in ORE. The results indicate a notable improvement in unknown detection.

Task ID	Task 1		Task 2				Task 3				Task 4		
	mAP	R50	mAP			R50	mAP			R50	mAP		
	Current	Unknown	Previous	Current	Both	Unknown	Previous	Current	Both	Unknown	Previous	Current	Both
Pseudo-label for Unknown	Known	Unknown	Known	Known	Both	Unknown	Known	Known	Both	Unknown	Known	Known	Both
RPN	63.4	14.4	58.3	30.8	45.1	11.3	43.3	23.4	36.7	14.8	37.2	20.7	33.1
MAVL*	64.0	<b>50.1</b>	61.6	30.8	46.2	<b>49.5</b>	43.8	22.7	36.8	<b>50.9</b>	36.2	20.6	32.3

unknowns. The pipeline therefore lacks a good proposal set that generalizes to *novel objects*. We propose a variant of ORE, by using class-agnostic proposals for unknown object categories obtained from MAVL. For a fair comparison, the MViT is trained on a filtered dataset, generated by explicitly removing all captions from LMDet that contain any unknown category, leaving 0.76M image-text pairs (see Appendix A.4 for further details). The results in Table 4 and Fig. 4b indicate significant improvements in unknown detection. See Fig. 10 in Appendix C for more qualitative results.

### 4.3 Pretraining for Class-aware Object Detection

The recent progress in self-supervised learning (SSL) [46,21,6,79] has minimized the need for large *labelled* datasets to achieve good performance on downstream tasks. These techniques encode the global image representation and achieve competitive generalization on various downstream tasks. However, these methods are suboptimal for class-aware OD, where the classification needs to be performed at local image patches (i.e. bounding boxes). Several recent efforts have been reported to address this challenge. ReSim [73] and DetCo [74] only pretrain the backbone to encode local and global representations. Whereas, DETReg [3] pretrains both the backbone and detection network using off-the-shelf proposals from selective search [64] and achieves improvement over the previous methods.

However, the proposals from heuristic selective search method, used in DETReg pretraining, are overly noisy and contain redundant boxes. We show that replacing these noisy pseudo-labels with MViT proposals can improve the downstream performance on OD task (Table 5). Following DETReg, we select top-30 proposals from MAVL and pretrain the model for 50 epochs on ImageNet [55] dataset, followed by fine-tuning on 10% and 100% data from Pascal VOC [14] for 150 and 100 epochs respectively. The results show an absolute gain of  $\sim 7$  and  $\sim 1$  in AP in the two respective cases.

Table 5: Effect of using MAVL proposals for pre-training of DETReg [3] instead of Selective Search [64] proposals.

Dataset → Model ↓	Pascal-VOC 10%			Pascal-VOC 100%		
	AP	AP50	AP75	AP	AP50	AP75
DETReg - SS	51.4	72.2	56.6	63.5	83.3	70.3
DETReg - MAVL	<b>58.8</b>	<b>80.5</b>	<b>65.7</b>	<b>64.5</b>	<b>84.2</b>	<b>71.3</b>

Table 6: Proposals from MAVL are evaluated against state-of-the-art SOD and COD approaches. The **general**<sup>†</sup> represents ‘all objects’ text query.

Dataset → Model ↓	Text Query	DUT-OMRON		ECSSD		Dataset → Model ↓	Text Query	CHAMELEON		CAMO		COD10K	
		AP50	R50	AP50	R50			AP50	R50	AP50	R50	AP50	R50
CPD [71]	-	64.5	77.4	87.1	92.7	SINET-V2 [15]	-	67.3	76.7	56.5	77.2	44.4	66.6
PoolNet [41]	-	66.5	78.8	87.4	93.1	MAVL	<b>General</b> <sup>†</sup>	30.2	53.3	46.5	75.4	39.6	67.8
MAVL	<b>General</b> <sup>†</sup>	67.0	89.1	84.5	95.7	MAVL	<b>Task specific</b> <sup>††</sup>	36.2	61.1	48.0	78.3	42.0	69.1
MAVL	<b>Task specific</b> <sup>††</sup>	75.5	93.3	85.7	96.1								

(a) Salient OD (SOD). Here **task specific**<sup>††</sup> query combines proposals from ‘all salient objects’ and ‘all foreground objects’ text queries.

(b) Camouflaged OD (COD) on three datasets. Here **task specific**<sup>††</sup> query combines proposals from ‘all camouflaged objects’ and ‘all disguised objects’ text queries.

#### 4.4 Salient Object Detection

Given the generalized class-agnostic performance of MViTs on multiple domains, we evaluate their ability to distinguish between salient and non-salient parts of an image. We exploit the interactive nature of MViTs by passing specific queries to detect the salient objects. To this end, MAVL proposals generated with queries like ‘all salient objects’ are compared with PoolNet [41] and CPD [71] models that are specifically trained for predicting saliency maps. We evaluate the models on the DUT-OMRON [77] and ECSSD [57] datasets. These datasets are only used for MViT evaluation and are not used during training. Since MViTs generate bounding boxes, we convert the saliency ground-truths and the saliency maps predicted by CPD and PoolNet to bounding boxes using connected components labelling [69]. In the case of DUT-OMRON, the provided ground-truth bounding boxes are used by computing an average across the five human annotations.

Table 6a indicates the effectiveness of MAVL in detecting the foreground salient objects. It is also interesting to note how the **task specific**<sup>††</sup> query (e.g., ‘all salient/foreground objects’) provides better prediction of salient parts of the image in comparison to a more generic<sup>†</sup> query like ‘all objects’ (Fig. 5a). See Appendix D.5 and Fig. 11 in Appendix C for additional details.

#### 4.5 Camouflaged Object Detection

Camouflaged object detection (COD) involves identifying objects that are *seamlessly* embedded in their background. The objects have a similar texture to their surroundings and are difficult to locate as compared to salient or generic objects. Here, we explore the interactive OD capacity of MViTs on COD task by evaluating the performance of MAVL against the state-of-the-art model (SINET-V2 [15]) on CHAMELEON [59], CAMO [35] and COD10K [16] datasets (Table 6b). Similar to salient OD setting, we convert camouflage ground-truth masks and masks predicted by SINET-V2 to bounding boxes using connected components labelling [69]. However, the available bounding box ground-truths have been used for COD10K dataset. We note favorable performance of MAVL proposals, although the model is not specifically trained on camouflaged objects (Fig. 5a). This affirms the generality of MAVL proposals. See Appendix D.6 and Fig. 11.

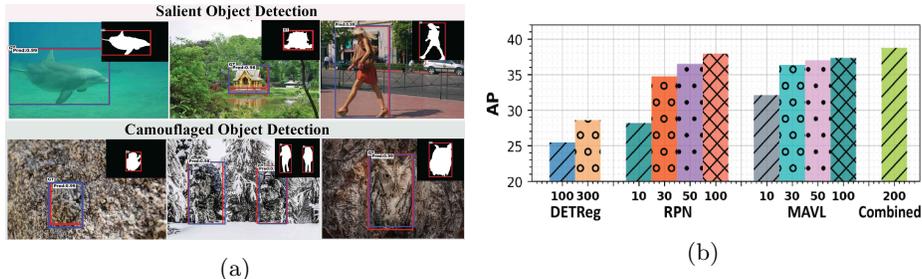


Fig. 5: (a) Qualitative results of Saliency (Top) and Camouflaged OD (Bottom). The ground-truth masks and boxes are shown on top right of the images. (b) Complimentary effect of using off-the-shelf proposals from MAVL in Faster RCNN [54] trained on COCO [40], indicated as ‘combined’ (*i.e.*, RPN + MAVL). The x-axis shows the number of proposals. MAVL generates good quality proposals, which perform well even with small proposal set sizes and demonstrate complimentary advantage to RPN.

#### 4.6 Improving Two-stage Object Detection

The class-agnostic object proposals from MViTs have strong understanding of semantics and can be deployed along with the region proposal network (RPN) [54]. We observe an improvement in accuracy when off-the-shelf MAVL proposals are combined with RPN proposals in Faster RCNN [54] during inference (Fig. 5b). This indicates the complimentary nature of these proposals that is based on a rich top-down perception of the image content.

Fig. 5b shows the results of replacing RPN proposals in Faster RCNN with DETReg [3] and MAVL proposals. The results indicate that the supervised proposal generation methods (RPN and MAVL) perform well compared to the unsupervised method (DETReg). However, off-the-shelf MAVL proposals show better performance than RPN when using a small proposal set (*e.g.*, 10 proposals). Combining RPN and MAVL proposals improves the overall detection accuracy.

## 5 What makes MViTs a Generic Detector?

Our empirical analysis shows the state-of-the-art performance of MViTs towards class-agnostic OD across different domains (Sec. 3) which positively impacts a number of downstream applications (Sec. 4). Having established this, we conduct a series of systematic experiments to explore the contributing factors for representational learning of the general ‘*objectness measure*’ in MViTs. Specifically, we identify the role of supervision and multi-modal learning as crucial factors.

### 5.1 On the importance of supervision

We consider two recent unsupervised learning models, DETReg [3] and UP-DETR [11]. DETReg trains Deformable DETR [83] to localize objects in class-agnostic fashion, with bounding box pseudo-labels from an off-the-shelf region

proposal method (Selective Search [64]). Meanwhile, UP-DETR performs unsupervised pretraining on random query patches in an image for class-agnostic OD. Both the unsupervised models, DETReg and UP-DETR, are trained on unimodal (Deformable DETR [83]) trained on LMDet in class-agnostic fashion, to evaluate the performance contributed by language supervision. We note that the image-level supervision with only box labels improves the performance in comparison with unsupervised methods. However, the use of caption texts aligned with input images proves to be vital and improves the performance approximately by *two* times, highlighting the importance of multi-modal supervision.

Table 7: MAVL proposals perform well compared to unsupervised methods (UP-DETR [11] and DETReg [3]) and supervised unimodal method (Def-DETR [83]).

Dataset → Model ↓	Supervision	Pascal-VOC		COCO		KITTI	
		AP50	R50	AP50	R50	AP50	R50
UP-DETR	unsupervised	0.56	16.6	0.19	6.56	0.01	0.65
DETReg	self-supervised	2.58	45.7	2.04	26.0	0.01	2.48
Def-DETR	box-level	30.1	81.0	20.0	53.5	23.7	55.0
MAVL	box + text	68.6	91.3	43.6	65.0	48.2	63.5

## 5.2 How much does language contribute?

Given the importance of multi-modal supervision towards better performance, we find it pertinent to explore the benefit solely from the language supervision. We conduct an ablation study on MDETR and MAVL, by removing all textual inputs corresponding to captions, but keeping intact the structure introduced by language *i.e.*, learning to localize boxes corresponding to a caption for each image in an iteration (without any language branch). Both MDETR and MAVL are trained on LMDet containing aligned image-text pairs. Here, the structure in which the information is fed during training is of high importance to us. Each image may have multiple captions, and hence it may be seen multiple times in the same iteration, but with varying contexts. The experimental setup removes all captions during training and evaluations, however keeps the described data loader structure intact, thus having approximately 1.3M iterations in an epoch. All models use ResNet-101 backbone and are evaluated after 10 epochs for ablation (instead of total 20 epochs). Table 8 indicate that visual branch plays a vital role, however the importance of language cannot be ruled out since the boxes related to a caption are still seen together. We analyze the importance of this implicit language structure next.

**Ablation on language structure:** The above experimental results reveal that removal of textual information does not significantly affect model performance. However, a further ablation on the structure introduced by language is required for the completeness of this evaluation. As such, we conduct ablations at five levels using Deformable DETR [83], as shown in Table 9. First, all the annotations

Table 8: Effect of removing language branch from MVITs keeping the data loader structure intact. The performance is not affected largely as the language structure is still intact (boxes from caption are seen together).

Dataset → Model ↓	Lang.	Pascal-VOC		COCO		KITTI	
		AP50	R50	AP50	R50	AP50	R50
MDETR	✓	63.9	88.0	38.1	58.5	42.5	60.9
MAVL	✓	65.0	89.1	39.3	62.0	39.0	61.0
MDETR	×	59.7	86.4	33.4	57.9	36.9	55.0
MAVL	×	61.6	86.7	34.4	58.3	36.5	58.9

in LMDet are combined at image level by concatenating the bounding boxes of all captions corresponding to an image (Setting-1). This removes any prior information introduced by the language structure. Then, class-agnostic NMS is applied at a threshold of 0.9 to filter boxes that have high overlaps (Setting-2). To imitate the repetitive pattern introduced during training, bounding box annotations corresponding to an image are randomly sampled and grouped (Setting-3). The number of samples in a combination is kept close to the average number of boxes in image-text pairs in original MAVL training ( $\sim 6$  boxes). Finally, a longer training schedule is used in the same setting to replicate a scenario closer to the original MAVL training (Setting-4). These four settings are then compared with a model that is trained without any captions, but maintains the structure introduced by language (Setting-5, same as Table 8 last row). This analysis indicates that language structure has significant impact in learning a general notion of objectness. With the use of aligned image-text pairs, additional contextual information is provided to the model. As objects generally tend to co-occur with other objects and certain scenes, such contextual association can be exploited for visual understanding [47]. Use of captions that describe a scene conveys such a notion of co-occurring objects and their mutual relationships, indicating that the structure introduced by language provides rich semantic and spatial context. Consistent with our findings, other recent efforts also indicate strong generalization achieved using the context encoded within natural language [80,53,78,82].

Table 9: Experimental analysis to explore the contribution of language by removing all textual inputs, but maintaining the structure introduced by captions. Experiments are performed on Def-DETR [83] using LMDet.

Experiment	Language	Pascal-VOC		MSCOCO		KITTI	
	Structure	AP50	R50	AP50	R50	AP50	R50
Setting-1	×	16.2	74.5	10.7	47.0	19.4	57.3
Setting-2	×	30.1	81.0	20.0	53.5	23.7	55.0
Setting-3	×	33.8	82.5	19.3	55.8	21.2	52.7
Setting-4	×	35.1	82.7	21.2	56.3	21.5	58.5
Setting-5	✓	<b>61.6</b>	<b>86.7</b>	<b>34.4</b>	<b>58.3</b>	<b>36.5</b>	<b>58.9</b>

## 6 Conclusion

This paper demonstrates intriguing performance of MViTs, trained only on natural images, for generic OD across a diverse set of domains. We systematically study the main reasons for this generalization, and note that the language structure available in image-caption pairs used to train MViTs plays a key role. Based on these insights, we develop a more flexible and efficient MViT for off-the-shelf class-agnostic OD, that can be instantiated with different text queries to generate desired proposal sets. Furthermore, we show various use-cases where class-agnostic proposals can be used to improve performance *e.g.*, open-world OD, camouflaged and salient OD, supervised and self-supervised OD.

**Acknowledgements.** Ming-Hsuan Yang is supported by the NSF CAREER grant 1149783. Fahad Shahbaz Khan is supported by the VR starting grant (2016-05543).

## References

1. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 73–80. IEEE (2010)
2. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the Objectness of Image Windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**(11), 2189–2202 (2012)
3. Bar, A., Wang, X., Kantorov, V., Reed, C.J., Herzig, R., Chechik, G., Rohrbach, A., Darrell, T., Globerson, A.: DETReg: Unsupervised Pretraining with Region Priors for Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
4. Bendale, A., Boulton, T.: Towards Open World Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1893–1902 (2015)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End Object Detection with Transformers. In: The European Conference on Computer Vision. pp. 213–229. Springer (2020)
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In: Advances in Neural Information Processing Systems (2020)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers. arXiv preprint arXiv:2104.14294 (2021)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. In: International Conference on Machine Learning. pp. 1597–1607. PMLR (2020)
9. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: UNiversal Image-TExt Representation Learning. In: The European Conference on Computer Vision. pp. 104–120. Springer (2020)
10. Cheng, M.M., Zhang, Z., Lin, W.Y., Torr, P.: BING: Binarized Normed Gradients for Objectness Estimation at 300fps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3286–3293 (2014)
11. Dai, Z., Cai, B., Lin, Y., Chen, J.: UP-DETR: Unsupervised Pre-training for Object Detection with Transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1601–1610 (2021)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL (2019)
13. Dhamija, A., Gunther, M., Ventura, J., Boulton, T.: The Overlooked Elephant of Object Detection: Open Set. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1021–1030 (2020)
14. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* **88**(2), 303–338 (2010)
15. Fan, D.P., Ji, G.P., Cheng, M.M., Shao, L.: Concealed Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021)
16. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2777–2787 (2020)

17. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3354–3361. IEEE (2012)
18. Georgakis, G., Reza, M.A., Mousavian, A., Le, P.H., Košecká, J.: Multiview RGB-D Dataset for Object Instance Detection. In: CoRR. pp. 426–434. IEEE (2016)
19. Gupta, A., Dollar, P., Girshick, R.: LVIS: A Dataset for Large Vocabulary Instance Segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5356–5364 (2019)
20. Gupta, T., Kamath, A., Kembhavi, A., Hoiem, D.: Towards General Purpose Vision Systems. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16399–16409 (2022)
21. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum Contrast for Unsupervised Visual Representation Learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
22. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2961–2969 (2017)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
24. Honnibal, M., Montani, I.: spaCy: Industrial-strength Natural Language Processing in Python (2020)
25. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6700–6709 (2019)
26. Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K.: Cross-Domain Weakly-Supervised Object Detection Through Progressive Domain Adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5001–5009 (2018)
27. Jaiswal, A., Wu, Y., Natarajan, P., Natarajan, P.: Class-agnostic Object Detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 919–928 (2021)
28. Joseph, K., Khan, S., Khan, F.S., Balasubramanian, V.N.: Towards Open World Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5830–5840 (2021)
29. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: MDETR—Modulated Detection for End-to-End Multi-Modal Understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1780–1790 (2021)
30. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Conference on Empirical Methods in Natural Language Processing
31. Kim, D., Lin, T.Y., Angelova, A., Kweon, I.S., Kuo, W.: Learning Open-World Object Proposals without Learning to Classify. arXiv preprint arXiv:2108.06753 (2021)
32. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* **123**(1), 32–73 (2017)

33. Kuo, W., Hariharan, B., Malik, J.: DeepBox: Learning Objectness with Convolutional Networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2479–2487 (2015)
34. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallocci, M., Kolesnikov, A., et al.: The open images dataset v4. *IJCV* **128**(7), 1956–1981 (2020)
35. Le, T.N., Nguyen, T.V., Nie, Z., Tran, M.T., Sugimoto, A.: Anabranched network for camouflaged object segmentation. *Computer Vision and Image Understanding* **184**, 45–56 (2019)
36. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv preprint arXiv:1908.03557 (2019)
37. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In: The European Conference on Computer Vision. pp. 121–137. Springer (2020)
38. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature Pyramid Networks for Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2117–2125 (2017)
39. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal Loss for Dense Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2980–2988 (2017)
40. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: The European Conference on Computer Vision. pp. 740–755. Springer (2014)
41. Liu, J.J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A Simple Pooling-Based Design for Real-Time Salient Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3917–3926 (2019)
42. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M.: Deep Learning for Generic Object Detection: A Survey. *International Journal of Computer Vision* **128**(2), 261–318 (2020)
43. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692 (2019)
44. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In: Advances in Neural Information Processing Systems (2019)
45. Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S.: 12-in-1: Multi-Task Vision and Language Representation Learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10437–10446 (2020)
46. Misra, I., Maaten, L.v.d.: Self-Supervised Learning of Pretext-Invariant Representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6707–6717 (2020)
47. Oliva, A., Torralba, A.: The role of context in object recognition. *Trends in Cognitive Sciences* **11**(12), 520–527 (2007)
48. Peyré, G., Cuturi, M.: Computational Optimal Transport (2020)
49. Pinheiro, P.O., Collobert, R., Dollár, P.: Learning to Segment Object Candidates. In: Advances in Neural Information Processing Systems (2015)
50. Pinheiro, P.O., Lin, T.Y., Collobert, R., Dollár, P.: Learning to Refine Object Segments. In: The European Conference on Computer Vision (2016)

51. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2641–2649 (2015)
52. Pont-Tuset, J., Arbelaez, P., Barron, J.T., Marques, F., Malik, J.: Multiscale Combinatorial Grouping for Image Segmentation and Object Proposal Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(1), 128–140 (2016)
53. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: International Conference on Machine Learning (2021)
54. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems* **28**, 91–99 (2015)
55. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* **115**(3), 211–252 (2015)
56. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8430–8439 (2019)
57. Shi, J., Yan, Q., Xu, L., Jia, J.: Hierarchical Image Saliency Detection on Extended CSSD. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**(4), 717–729 (2015)
58. Siméoni, O., Puy, G., Vo, H.V., Roburin, S., Gidaris, S., Bursuc, A., Pérez, P., Marlet, R., Ponce, J.: Localizing Objects with Self-Supervised Transformers and no Labels. In: British Machine Vision Conference (2021)
59. Skurowski, P., Abdulameer, H., Błaszczuk, J., Depta, T., Kornacki, A., Koziel, P.: Animal Camouflage Analysis: CHAMELEON Database. Unpublished Manuscript **2**(6), 7 (2018)
60. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In: International Conference on Learning Representations (2019)
61. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: VideoBERT: A Joint Model for Video and Language Representation Learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7464–7473 (2019)
62. Tan, H., Bansal, M.: LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In: Conference on Empirical Methods in Natural Language Processing (2019)
63. Tan, M., Le, Q.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: International Conference on Machine Learning. pp. 6105–6114. PMLR (2019)
64. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective Search for Object Recognition. *International Journal of Computer Vision* **104**(2), 154–171 (2013)
65. Wang, W., Feiszli, M., Wang, H., Tran, D.: Unidentified Video Objects: A Benchmark for Dense, Open-World Segmentation. arXiv preprint arXiv:2104.04691 (2021)
66. Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly Simple Few-Shot Object Detection. arXiv preprint arXiv:2003.06957 (2020)

67. Wang, X., Cai, Z., Gao, D., Vasconcelos, N.: Towards Universal Object Detection by Domain Attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7289–7298 (2019)
68. Wightman, R.: PyTorch Image Models. <https://github.com/rwightman/pytorch-image-models> (2019). <https://doi.org/10.5281/zenodo.4414861>
69. Wu, K., Otoo, E., Shoshani, A.: Optimizing connected component labeling algorithms. In: Medical Imaging 2005: Image Processing. vol. 5747, pp. 1965–1976. International Society for Optics and Photonics (2005)
70. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. <https://github.com/facebookresearch/detectron2> (2019)
71. Wu, Z., Su, L., Huang, Q.: Cascaded Partial Decoder for Fast and Accurate Salient Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3907–3916 (2019)
72. Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: DOTA: A Large-scale Dataset for Object Detection in Aerial Images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3974–3983 (2018)
73. Xiao, T., Reed, C.J., Wang, X., Keutzer, K., Darrell, T.: Region Similarity Representation Learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
74. Xie, E., Ding, J., Wang, W., Zhan, X., Xu, H., Sun, P., Li, Z., Luo, P.: DetCo: Unsupervised Contrastive Learning for Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8392–8401 (2021)
75. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with Noisy Student improves ImageNet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10687–10698 (2020)
76. Yan, K., Wang, X., Lu, L., Summers, R.M.: DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging* **5**(3), 036501 (2018)
77. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency Detection via Graph-Based Manifold Ranking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3166–3173 (2013)
78. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-Vocabulary Object Detection Using Captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14393–14402 (2021)
79. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In: International Conference on Machine Learning (2021)
80. Zhang, M., Tseng, C., Kreiman, G.: Putting visual object recognition in context. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12985–12994 (2020)
81. Zhang, Z., Liu, Y., Chen, X., Zhu, Y., Cheng, M.M., Saligrama, V., Torr, P.H.: BING++: A Fast High Quality Object Proposal Generator at 100fps. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 40, pp. 1209–1223 (2018)
82. Zhou, M., Zhou, L., Wang, S., Cheng, Y., Li, L., Yu, Z., Liu, J.: UC2: Universal Cross-lingual Cross-modal Vision-and-Language Pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4155–4165 (2021)

83. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable Transformers for End-to-End Object Detection. In: International Conference on Learning Representations (2021)
84. Zitnick, C.L., Dollár, P.: Edge Boxes: Locating Object Proposals from Edges. In: The European Conference on Computer Vision. pp. 391–405. Springer (2014)