11-22-2021

Multi-modal Transformers Excel at Class-agnostic Object Detection

Muhammad Maaz
Mohamed Bin Zayed University of Artificial Intelligence

Hanoona Bangalath Rasheed
Mohamed Bin Zayed University of Artificial Intelligence

Salman Hameed Khan
Mohamed Bin Zayed University of Artificial Intelligence & Australian National University

Fahad Shahbaz Khan
Mohamed bin Zayed University of Artificial Intelligence

Rao Muhammad Anwer
Mohamed Bin Zayed University of Artificial Intelligence & Aalto University

See next page for additional authors

Multi-modal Transformers Excel at Class-agnostic Object Detection

Muhammad Maaz[1*] Hanoona Rasheed[1*] Salman Khan[1,2] Fahad Shahbaz Khan[1]
Rao Muhammad Anwer[1,3] Ming-Hsuan Yang[4,5,6]
[1]Mohamed bin Zayed University of AI [2]Australian National University [3]Aalto University
[4]University of California, Merced [5]Yonsei University [6]Google Research
{muhammad.maaz, hanoona.bangalath}@mbzuai.ac.ae

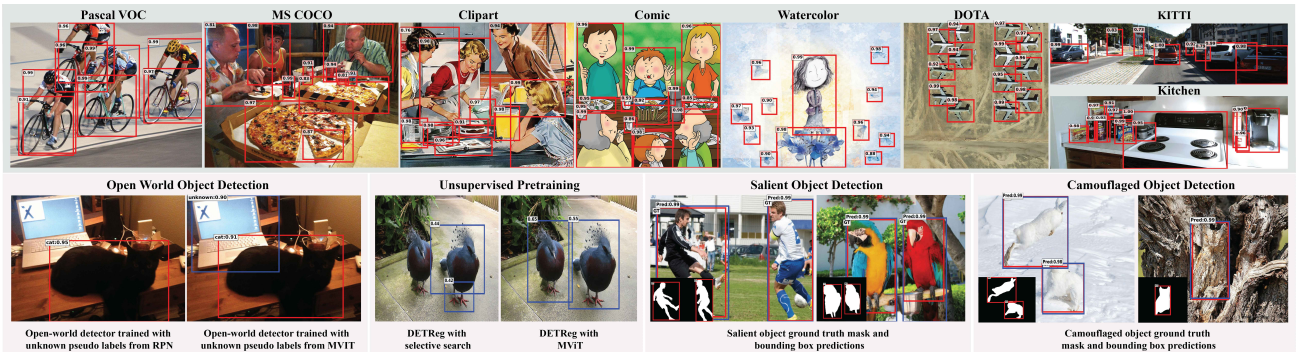

Figure 1. **Top Row:** Class-agnostic object detection (OD) performance of pre-trained Multi-modal Vision Transformer (MViT) [28] across multiple domains (natural images [14, 17, 18, 37], satellite images [67], sketches, cartoons and paintings [25]). The MViTs [20, 28] perform well on diverse datasets using human intuitive natural language text queries (*e.g.*, `all objects`, `all entities`). **Bottom Row:** Class-agnostic detectors (MViTs) can be applied to several downstream applications. In Open-world OD [27], unknown pseudo labels generated using MDETR [28] can improve novelty detection. For unsupervised object localization, replacing Selective Search proposals [60] in DETReg [3] pretraining with only top-30 MViT proposals leads to improved localization. For Salient and Camouflaged OD, task specific text queries like '`all salient objects`' and '`all camouflage objects`' can help perform competitively against fully supervised models without any task specific tuning. Overall, MViTs achieve the state-of-the-art results on various applications.

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 for unseen objects. In this paper, we advocate that existing methods lack a top-down supervision signal governed by human-understandable semantics. To bridge this gap, we explore recent Multi-modal Vision Transformers (MViT) that have been trained with aligned image-text pairs. 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 these findings, we develop an efficient and flexible MViT architecture using multi-scale feature processing and deformable self-attention that can adaptively generate proposals given a specific language query. 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 offer enhanced interactability with intelligible text queries. Code:* https://git.io/J1HPY.

[*]Equal contribution

1. Introduction

The recent years have witnessed significant advances in object detection (OD) [39] 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 [62], long-tailed object distribution modeling [19] and open-world OD [27]. In contrast, little progress has been made towards a seemingly simpler task of class-agnostic OD in recent years. In the era of fully trainable pipelines, class-agnostic object detection [1] is still often approached using traditional bottom-up approaches such as Selective Search [60], EdgeBox [79], DeepMask [46] and MCG [48].

Despite being a seemingly 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 delineating 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 [68, 69] – have solely 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 traditional 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 down-stream 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 [20, 28] towards class-agnostic OD via a set of human-understandable natural language queries. We also develop an efficient and flexible MViT model, *Multi-modal Deformable Detection Transformer* (MDef-DETR), which can effectively locate generic objects. Through an extensive set of systematic experiments, we analyze the factors that majorly contribute to the improved performance of MViTs (Secs. 3 & 4).
- We benchmark the generalization of MViT based object detectors on diverse domains *e.g.*, natural images, sketches, cartoons, satellite imagery and paintings and show their favorable performance compared to existing class-agnostic OD models (bottom-up approaches, CNN and ViT based uni-modal learned pipelines) (Sec. 3).
- We demonstrate applicability of class-agnostic detectors to 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 image content (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 ('`all objects/entities`'). Before a detailed analysis, we provide background on MViTs and propose Modulated Deformable DETR (MDef-DETR).

`(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 is based on DETR [5] and trains on data from five different tasks including visual question answering (VQA), captioning, localization, classification and referring expression. GPV uses pretrained BERT [12] to encode the text, concatenates it with the region descriptors from DETR and passes it to ViLBERT [41] co-attention layers for cross-modal conceptualization. It predicts relevance scores for each predicted box indicating the importance of the region for the prompted task. An output text decoder is used conditioned on the relevance scores for better cross-modal understanding (Fig. 2 (a)).

`(b)` **MDETR:** Kamath *et al*. [28] proposed an end-to-end 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 [59]) and a language model (RoBERTa [40]) 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)` **M-Deformable DETR:** Fig. 2 (c) shows our overall design. Below, we highlight main features of MDef-DETR:

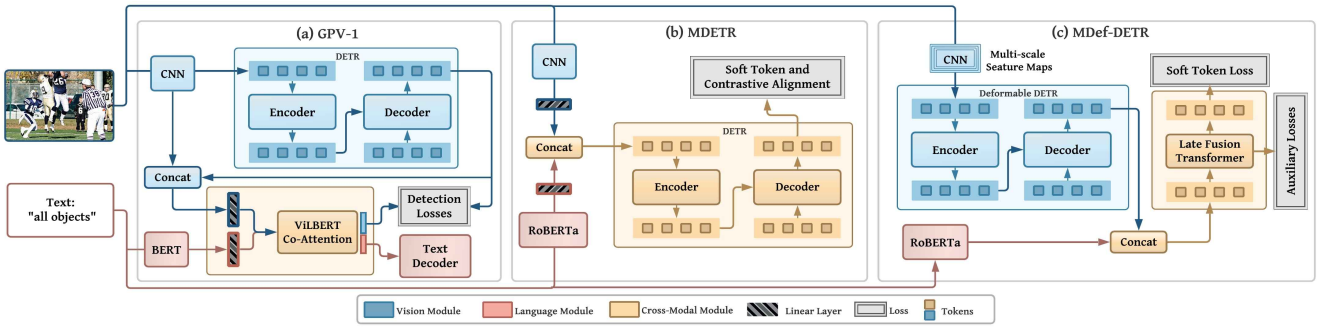

Figure 2. Architecture overview of MViTs used in this work – GPV-1 [20], MDETR [28] and MDef-DETR (ours). GPV-1 takes image along with a task description as input and outputs relevant region boxes and text. MDETR uses soft token prediction and contrastive alignment in latent space for cross-conceptualization using aligned image-text pairs. MDef-DETR utilizes multi-scale image features with multi-scale deformable attention module (MSDA) from [78], and uses late-fusion strategy for multi-modal fusion.

−Multi-scale Deformable Attention (MSDA). MDETR [28] 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. Here, we explore a variant based on Deformable DETR (Def-DETR) [5] that employs multi-scale feature processing in its MSDA module and dynamically attends to relevant pixel locations for context aggregation. This design uses Deformable Attention that samples a small set of keys around a reference (query) image location. The sparse key sampling in Def-DETR achieves linear complexity with respect to the size of the image feature maps.

−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 the MDef-DETR model after the inputs are processed through the Def-DETR encoder-decoder architecture using a *late fusion* mechanism. Specifically, the query representations from deformable decoder are concatenated with the text embeddings, and passed through a series of transformer self-attention (SA) blocks. This is consistent with the recent vision-language fusion works [41, 56–58]. 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 this approach. Experimental results show fast convergence (only *half* iterations) and competitive performance of MDef-DETR against MDETR (Table 1 and 2).

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. super-pixels, edges, etc.) or directly learn the mapping between images and generic object locations using fully trainable pipelines learned with bounding box annotations [3, 26, 60, 79]. 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 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 shows 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.

Dataset → / Model ↓	Pascal-VOC AP50	Pascal-VOC R50	COCO AP50	COCO R50	KITTI AP50	KITTI R50
Edge Boxes	0.08	7.14	0.09	5.16	0.09	6.58
Selective Search	0.32	21.35	0.27	12.72	0.03	4.85
Deep Mask	5.92	40.39	2.16	19.22	1.33	15.50
Faster-RCNN	42.88	85.84	26.36	58.66	23.50	53.23
RetinaNet	43.15	86.55	24.59	59.10	30.43	57.60
Def-DETR	30.06	81.04	19.99	53.50	23.69	55.00
GPV-I	61.94	91.12	47.41	70.52	42.98	64.43
MDETR	66.04	90.10	40.66	62.15	46.71	67.24
MDef-DETR	68.59	91.26	43.64	65.03	48.22	63.53

Table 1. Class-agnostic OD performance of MViTs in comparison with traditional bottom-up approaches and uni-modal detectors trained to localize generic objects. We report average precision (AP) and Recall (R) at IoU threshold of 0.5. The MViTs achieve state-of-the-art results using intuitive text queries (Sec. 5.1).

3.1. Class-agnostic Object Detection

Table 1 shows the object proposal generation performance of MViTs with the traditional bottom-up approaches and the end-to-end supervised deep learning methods on three commonly used natural image OD datasets (Pascal

Dataset →	Kitchen		Clipart		Comic		Watercolor		DOTA†	
Model ↓	AP50	R50	AP50	R50	AP50	R50	AP50	R50	AP50	R50
RetinaNet	35.33	89.53	27.03	89.97	33.07	86.14	47.78	91.90	0.72	15.58
GPV-1	24.53	84.79	35.11	86.11	42.29	83.62	50.32	89.54	0.55	9.33
MDETR	38.38	91.38	44.94	90.69	55.82	89.45	63.59	94.32	1.94	21.80
MDef-DETR	45.43	90.99	50.59	92.86	57.72	89.20	63.78	95.16	2.86	24.23

Table 2. Class-agnostic OD performance of MViTs in comparison with RetinaNet [36] on several out-of-domain datasets. MViTs show consistently good results on all datasets. †Proposals on DOTA [67] are generated by multi-scale inference. (Sec. A.2)

VOC [14], MS COCO [37] and KITTI [17]). The bottom-up approaches considered for comparison include Edge-Boxes [79], Selective Search [60] and DeepMask [46] while Faster-RCNN [51], RetinaNet [36] and Deformable-DETR [78] are selected from the deep-learning based methods due to the state-of-the-art performance in class-aware OD. These detectors are trained in a class-agnostic manner using the dataset used for pretraining in MDETR [28]. The MViTs considered are GPV-I [20] and MDETR [28] alongside our proposed MDef-DETR (see Sec. 2 for details).

We report both average precision (AP) and Recall at IoU threshold of 0.5 using the top-50 boxes from each method. The results demonstrate that the detectors trained in class-agnostic fashion perform reasonably well on all datasets, surpassing the bottom-up methods by a large margin. Concurrently, the MViTs perform better than the uni-modal approaches with the use of simple human understandable natural language text queries. This performance shows MViTs' strong understanding of language obtained from the pre-trained language model (BERT [12], RoBERTa [40]) along with the aligned image-text pairs used in pretraining.

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

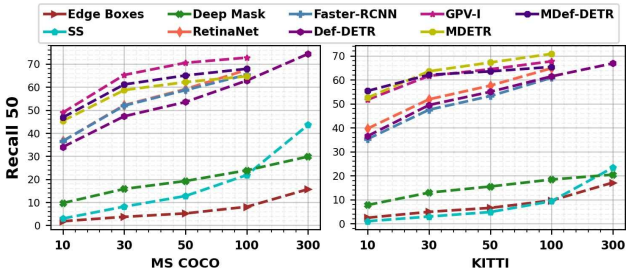

Figure 3. Effect of using different number of top-ranked boxes on multiple class-agnostic OD methods. The MViTs exhibits good recall even with only top-10 and top-30 object proposal.

3.2. How well MViTs 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 [62] belonging to five different domains (Table 2). The studied domains include indoor kitchen scenes [18], cartoon images, watercolor drawings, clipart, comics [25] and satellite/aerial images (DOTA dataset) [67]. The experiments follow the same setting as in Sec. 3.1. These results indicate the generalization capability of MViTs in comparison to the best proposal generation methods earlier evaluated in Table 1.

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, scalabiltiy to rare and novel classes is a desired quality of an object proposal algorithm. To analyze this, the class-agnostic OD mechanism of MDef-DETR is evaluated on rare categories from Open-Images [49] versus frequent categories. The results in Fig. 4 indicate the state-of-the-art recall rates on categories such as *lynx*, *humidifier*, and *armadillo* with as few as zero training instance. Overall, the model generalizes well to rare/unseen categories.

4. 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). 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.

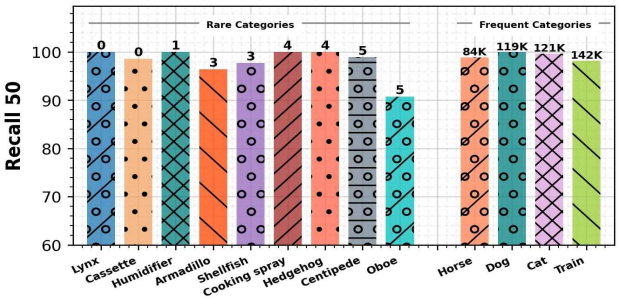

Figure 4. MDef-DETR class-agnostic OD performance on rarely and frequently occurring categories in the pretraining captions. Rare categories, selected from Open Image [49], are shown on left and frequent categories are shown on right. The numbers on top of bars indicate the occurrences of the category in the pretraining dataset. The MViT achieves good recall values even for the classes with no or very few occurrences in the training dataset.

4.1. On the importance of supervision

To understand the role of supervision provided via aligned image-text pairs during training, we evaluate performance of similar architectures trained in an unsupervised manner (Table 3). We consider two recent unsupervised learning models, DETReg [3] and UP-DETR [11]. DETReg trains Deformable DETR [78] to localize objects in class-agnostic fashion, with bounding box pseudo labels from an off-the-shelf region proposal method (Selective Search [60]). 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 ImageNet-1K [52] dataset. Further, we consider a supervised uni-modal (Deformable DETR [78]) trained on MDETR pretraining dataset 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.

Dataset →		Pascal-VOC		COCO		KITTI	
Model ↓	Supervision	AP50	R50	AP50	R50	AP50	R50
UP-DETR	×	0.56	16.61	0.19	6.56	0.001	0.65
DETReg	×	2.58	45.73	2.04	26.00	0.009	2.48
Def-DETR	✓	30.06	81.04	19.99	53.50	23.69	55.00
MDef-DETR	✓	68.59	91.26	43.64	65.03	48.22	63.53

Table 3. class-agnostic OD results of MDef-DETR versus unsupervised object proposal methods (UP-DETR [11] and DETReg [3]) and supervised uni-modal method (Def-DETR [78]). The MViT achieves state-of-the-art class-agnostic OD performance.

4.2. How much does language contribute?

Given the importance of multi-modal supervision towards better performance, we find it pertinent to explore the benefit solely coming from the language supervision. We conduct an ablation study on MDETR [28] and MDef-DETR (ours), by removing all textual inputs corresponding to the caption, but keeping intact the structure introduced by language *i.e.*, learning to localize boxes corresponding to a single caption for each image in an iteration (without any language branch). Both MDETR and MDef-DETR are trained with a large aligned image-text paired dataset consisting of approximately 118K images and corresponding 1.25M captions and their bounding box annotations. 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 1.25M iterations in an epoch. All models use a ResNet-101 backbone and are evaluated after 10 epochs. Evaluations presented in Table 4 indicate that visual branch plays a vital role, however the importance of language can not be ruled out since the boxes related to a caption are still seen together. We analyze the importance of this implicit language structure next.

Dataset→		Pascal-VOC		COCO		KITTI	
Model ↓	Lang.	AP50	R50	AP50	R50	AP50	R50
MDETR	✓	63.87	87.99	38.10	58.50	42.49	60.93
MDef-DETR	✓	65.03	89.09	39.33	62.03	39.04	61.04
MDETR	×	59.74	86.37	33.37	57.94	36.91	54.97
MDef-DETR	×	61.58	86.71	34.43	58.27	36.50	58.89

Table 4. Effect of removing language branch from MViTs keeping the data loader structure intact. All the experiments are run for 10 epochs. The removal of language branch does not effect MViTs' performance largely since the language structure is still intact (boxes related to a caption are seen together).

Ablation on language structure: The above experimental results reveal that removal of textual information does not significantly affect the performance of the model. However, a further ablation on the structure introduced by language in the training pipeline is required for the completeness of this evaluation. As such, we conduct ablations at five levels using Deformable DETR [78], as shown in Table 5. First, all the annotations are combined at an 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 non-maximum suppression (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 each image-text pair during original MDETR training ($\sim$6 boxes). Finally, a longer training schedule is used in the same setting to replicate a scenario closer to the original MDETR 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 4 last row).

This analysis indicates that language structure has a 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 contexual association can be exploited for visual understanding [44]. 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

Experiment	Language Structure	Pascal-VOC AP50	Pascal-VOC R50	MSCOCO AP50	MSCOCO R50	KITTI AP50	KITTI R50
Setting-1	×	16.15	74.50	10.66	46.97	19.44	57.30
Setting-2	×	30.05	81.04	20.00	53.50	23.69	54.99
Setting-3	×	33.81	82.54	19.29	55.77	21.23	52.73
Setting-4	×	35.06	82.72	21.23	56.34	21.50	58.54
Setting-5	✓	**61.58**	**86.71**	**34.43**	**58.27**	**36.50**	**58.89**

Table 5. 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 [78] using MDETR [28] pretraining data. In setting 1, annotations corresponding to same images are combined. Setting 2 has an additional NMS applied to remove duplicate boxes. In setting 3, four to eight boxes are randomly grouped in each iteration. The same model is trained longer in setting 4. In setting 5, the MDETR dataloader structure is kept intact. Results from setting 5 demonstrate the importance of structure introduced by language.

strong generalization achieved using the context encoded within natural language [50, 73, 75, 77].

5. Applications and Use-cases

5.1. Enhanced Interactability

We have observed 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 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`', and '`all obscure entities and objects`', for exploiting the class-agnostic OD performance of MViTs. The detections from the individual text queries are combined, filtered with class-agnostic NMS to remove duplicate detections, and the top N boxes are selected for evaluation.

Table 6 shows the effect of using different text queries in comparison with the combined detections across three datasets (Pascal VOC [14], MS COCO [37] and KITTI [17]). The results indicate that different text queries exploit varying aspects of objectness, and a global context can be captured using combined detections. This also reduces the dataset biasness to a specific query which in turn helps to perform reasonably well across different domains. Moreover, adding '`all small objects`' query improves performance on KITTI which has more small-sized objects.

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. 5 (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. 5 (right) indicates how the use of special queries like '`all long objects`' help improve the detection of irregular shaped objects (without any dataset specific fine-tuning!).

Dataset → / Text Query ↓	Pascal-VOC AP50	Pascal-VOC R50	COCO AP50	COCO R50	KITTI AP50	KITTI R50
`all objects`	51.33	85.51	33.33	58.36	40.19	63.96
`all entities`	65.18	88.38	34.56	54.55	41.48	59.47
`all visible entities & objects`	63.33	88.93	37.94	61.58	41.96	62.95
`all obscure entities & objects`	59.51	86.57	35.15	59.08	42.36	63.53
`all small objects`	40.02	83.90	28.85	58.86	40.42	65.20
combined detections (CD)	63.72	90.97	41.97	**65.13**	**48.22**	**63.53**
CD w/o '`all small objects`'	**68.59**	**91.26**	**43.64**	65.03	45.78	61.64

Table 6. Effect of using different intuitive text queries on the MDef-DETR class-agnostic OD performance. Combining detections from multiple queries captures varying aspects of objectness.

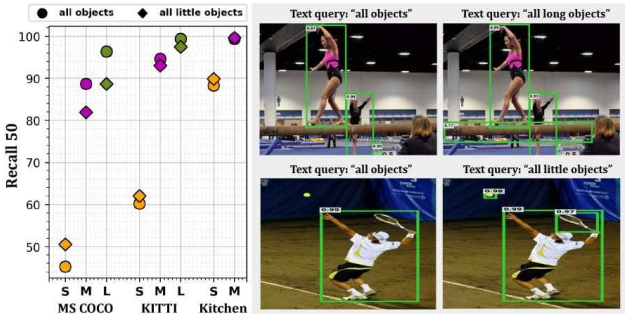

Figure 5. **Left:** MDef-DETR recall for small, medium and large objects across three datasets. The use of specific query like 'all little objects' increases the recall of small objects across different datasets. **Right:** Targeted detections by the relevant text queries.

5.2. 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, 27, 61]. 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) [27].

−ORE Setting: The authors distributed the 80 COCO [37] 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 [51] as the base detector, with contrastive clustering

Task ID	Task 1		Task 2				Task 3				Task 4		
Pseudo Label for Unknown	mAP	R50	mAP			R50	mAP			R50	mAP		
	Current Known	Unknown	Previously Known	Current Known	Both	Unknown	Previously Known	Current Known	Both	Unknown	Previously Known	Current Known	Both
RPN	63.44	14.40	58.28	30.78	45.09	11.32	43.33	23.37	36.68	14.79	37.20	20.73	33.08
MDef-DETR*	64.03	**50.13**	61.57	30.81	46.19	**49.54**	43.77	22.71	36.75	**50.89**	36.22	20.57	32.31

Table 7. Effect of using class-agnostic OD proposals from MDef-DETR for pseudo labelling of unknowns in ORE [27]. MDef-DETR* is a version of MDef-DETR trained on a filtered dataset generated by removing all captions listing any of 60 unknown categories evaluated in ORE. The results indicate a notable improvement in unknown performance when trained using unknown pseudo labels from the MViT.

in latent space and an energy-based classification head for unknown detection. It utilizes example-replay strategy [63] 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 [51] 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 unknowns. The pipeline therefore lacks a good proposal set that generalizes to *all objects*. We propose a variant of ORE, by using class-agnostic proposals for unknown object categories obtained from MDef-DETR. For fair comparison, the MViT is trained on a filtered dataset, generated by explicitly removing all captions that contain any unknown category, leaving 0.76M image-text pairs. Results in Table 7 and Fig. 6 indicate improvement in unknown detection.

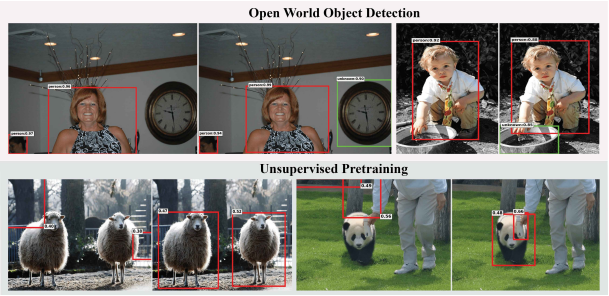

Figure 6. **Top:** Qualitative results of ORE [27] unknown detections when trained with RPN [51] versus MDef-DETR unknown pseudo labels. **Bottom:** class-agnostic OD of DETReg [3] when trained using Selective Search [60] versus MDef-DETR proposals.

5.3. Pretraining for Class-aware Object Detection

The recent progress in self-supervised learning (SSL) [6, 21, 43, 74] has minimized the need for large *labelled* datasets to achieve good performance on downstream tasks. These techniques primarily 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 [68] and DetCo [69] 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 [60] and achieves improvement over the previous methods.

Dataset→ Model ↓	Pascal-VOC 10% AP	AP50	AP75	Pascal-VOC 100% AP	AP50	AP75
DETReg - SS	51.40	72.20	56.60	63.50	83.30	70.30
DETReg - MDef-DETR	**58.78**	**80.46**	**65.65**	**64.51**	**84.16**	**71.29**

Table 8. Effect of using MDef-DETR proposals for pre-training of DETReg [3] instead of Selective Search [60] proposals. Pretraining using MDef-DETR proposal increases DETReg downstream performance on VOC dataset using both 10% and 100% data.

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 8). Following DETReg, we select top 30 proposals from MDef-DETR and pretrain the model for 50 epochs on ImageNet [52] dataset, followed by fine-tuning on 10% and 100% Pascal VOC [14] data for 150 and 100 epochs respectively. The results show a gain of ~ 7 and ~ 1 in AP in the two respective cases.

5.4. 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, MDef-DETR proposals generated with queries like '`all salient objects`' are compared with PoolNet [38] and CPD [66] models that are specifically trained for predicting saliency maps. We evaluate the models on salient object datasets DUT-OMRON [72] and ECSSD [53]. 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 [65]. In the case of DUT-OMRON, the provided ground truth bounding boxes are used by computing an average across the five human annotations.

Table 9 indicates the effectiveness of MDef-DETR in detecting the foreground salient objects. It is also interesting to note how the task specific†† query provides better prediction of salient parts of the image in comparison to a more generic† query like 'all objects' (Fig. 7).

Dataset →		DUT-OMRON		ECSSD	
Model ↓	Text Query	AP50	R50	AP50	R50
CPD [66]	-	64.45	77.40	87.10	92.70
PoolNet [38]	-	66.49	78.80	87.37	93.07
MDef-DETR	General†	66.95	89.05	84.52	95.69
MDef-DETR	Task specific††	75.52	93.26	85.70	96.07

Table 9. Proposals from MDef-DETR for Salient OD (SOD) task in comparison with the state-of-the-art saliency approaches. The MViT achieves top performance using task specific†† query, that combines detections from 'all salient objects' and 'all foreground objects' despite not explicitly trained on SOD task.

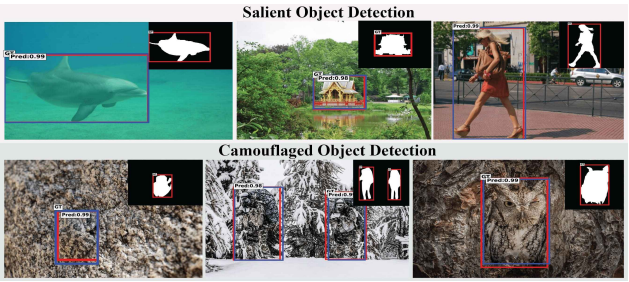

Figure 7. Qualitative results of MDef-DETR for Salient OD (top row) and Camouflaged OD (bottom row) tasks. The ground truth masks along with the generated ground truth bounding boxes are shown on top right of the images. The MViT shows good performance on the challenging salient and camouflaged OD tasks.

5.5. 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 compared to salient or generic objects. Here, we explore the interactive OD capacity of MViTs on COD task by evaluating the performance of MDef-DETR against the state-of-the-art model (SINET-V2 [15]) on CHAMELEON [55], CAMO [32] and COD10K [16] datasets (Table 10). Similar to salient OD setting, we convert camouflage ground truth masks and masks predicted by SINET-V2 to bounding boxes using connected components labelling [65]. However, the available bounding box ground truths have been used for COD10K dataset. We note favorable performance of MDef-DETR proposals for the challenging COD task, although the model is not specifically trained on camouflaged objects (Fig. 7).

5.6. 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 networks (RPN) [51]. We observe an improvement in accuracy when off-the-shelf MDef-DETR proposals are combined with RPN proposals in Faster RCNN [51] during inference (Fig. 8). This indicates the complimentary nature of these proposals that is based on a rich top-down perception of the image content.

Fig. 8 shows the results of replacing RPN proposals in Faster RCNN with DETReg [3] and MDef-DETR proposals. The results indicate that the supervised proposal generation methods (RPN and MDef-DETR) perform well compared to the unsupervised method (DETReg). However, off-the-shelf MDef-DETR proposals show better performance than RPN when using a small number of proposals (*e.g.*, 10 proposals). Moreover, combining RPN and MDef-DETR proposals improves the overall detection accuracy.

Dataset →		CHAMELEON		CAMO		COD10K	
Model ↓	Text Query	AP50	R50	AP50	R50	AP50	R50
SINET-V2 [15]	-	67.29	76.67	56.51	77.21	44.44	66.55
MDef-DETR	General†	30.24	53.33	46.49	75.37	39.55	67.78
MDef-DETR	Task specific††	36.16	61.11	48.04	78.31	41.98	69.13

Table 10. Proposals from MDef-DETR on the Camouflaged OD task. Despite being off-the-shelf proposals, the MViT proposals show good class-agnostic OD performance. For task specific†† inference, the detections are combined from 'all camouflage objects' and 'all disguise objects' text queries, while the general† query corresponds to the detections from 'all objects' query.

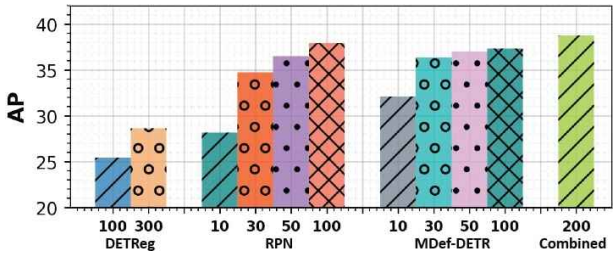

Figure 8. Complimentary effect of using off-the-shelf proposals from MDef-DETR in Faster RCNN [51] trained on COCO [37], indicated as 'combined' (*i.e.*, RPN + MDef-DETR). The x-axis shows the number of proposals used during inference by the corresponding methods. MDef-DETR generates good quality proposals, which perform well even with small proposal set sizes.

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] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 73–80. IEEE, 2010. 1

[2] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the Objectness of Image Windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2189–2202, 2012. 1, 16

[3] Amir Bar, Xin Wang, Vadim Kantorov, Colorado J Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. DETReg: Unsupervised Pretraining with Region Priors for Object Detection. *arXiv preprint arXiv:2106.04550*, 2021. 1, 2, 3, 5, 7, 8, 13, 15, 17

[4] Abhijit Bendale and Terrance Boult. Towards Open World Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1893–1902, 2015. 6

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *The European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2, 3, 17

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Advances in Neural Information Processing Systems*, 2020. 7, 17

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. *arXiv preprint arXiv:2104.14294*, 2021. 16

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 17

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt Representation Learning. In *The European Conference on Computer Vision*, pages 104–120. Springer, 2020. 16

[10] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip Torr. BING: Binarized Normed Gradients for Objectness Estimation at 300fps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3286–3293, 2014. 16

[11] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. UP-DETR: Unsupervised Pre-training for Object Detection with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1601–1610, 2021. 2, 5, 17

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 2019. 2, 4, 16

[13] Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The Overlooked Elephant of Object Detection: Open Set. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1021–1030, 2020. 6

[14] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 1, 4, 6, 7, 12, 13, 14

[15] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 8, 13, 14

[16] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2777–2787, 2020. 8, 14

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 1, 4, 6, 12, 13, 14

[18] Georgios Georgakis, Md Alimoor Reza, Arsalan Mousavian, Phi-Hung Le, and Jana Košecká. Multiview RGB-D Dataset for Object Instance Detection. In *CoRR*, pages 426–434. IEEE, 2016. 1, 4, 12, 13, 14, 16

[19] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019. 1

[20] Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. Towards General Purpose Vision Systems. *arXiv preprint arXiv:2104.00743*, 2021. 1, 2, 3, 4, 13, 17

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 7, 17

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2961–2969, 2017. 13, 14, 16

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2, 12, 16

[24] Matthew Honnibal and Ines Montani. spaCy: Industrial-strength Natural Language Processing in Python. 2020. 12

[25] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-Domain Weakly-Supervised Object Detection Through Progressive Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition, pages 5001–5009, 2018. 1, 4, 12, 13, 14, 16

[26] Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and Premkumar Natarajan. Class-agnostic Object Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 919–928, 2021. 3, 16

[27] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards Open World Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5830–5840, 2021. 1, 6, 7, 12, 15

[28] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR–Modulated Detection for End-to-End Multi-Modal Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 1, 2, 3, 4, 5, 6, 12, 13, 14, 16, 17

[29] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021. 16

[30] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning Open-World Object Proposals without Learning to Classify. *arXiv preprint arXiv:2108.06753*, 2021. 16

[31] Weicheng Kuo, Bharath Hariharan, and Jitendra Malik. DeepBox: Learning Objectness with Convolutional Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2479–2487, 2015. 16

[32] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer Vision and Image Understanding*, 184:45–56, 2019. 8

[33] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv preprint arXiv:1908.03557*, 2019. 16

[34] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *The European Conference on Computer Vision*, pages 121–137. Springer, 2020. 16

[35] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 16

[36] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2980–2988, 2017. 4

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *The European Conference on Computer Vision*, pages 740–755. Springer, 2014. 1, 4, 6, 8, 12, 13, 14

[38] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A Simple Pooling-Based Design for Real-Time Salient Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3917–3926, 2019. 7, 8, 13

[39] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep Learning for Generic Object Detection: A Survey. *International Journal of Computer Vision*, 128(2):261–318, 2020. 1

[40] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019. 2, 4

[41] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks . In *Advances in Neural Information Processing Systems*, 2019. 2, 3, 16

[42] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-Task Vision and Language Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020. 17

[43] Ishan Misra and Laurens van der Maaten. Self-Supervised Learning of Pretext-Invariant Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 7

[44] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527, 2007. 5

[45] Gabriel Peyré and Marco Cuturi. Computational Optimal Transport, 2020. 16

[46] Pedro O. Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to Segment Object Candidates. In *Advances in Neural Information Processing Systems*, 2015. 1, 4, 16

[47] Pedro O. Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to Refine Object Segments. In *The European Conference on Computer Vision*, 2016. 16

[48] Jordi Pont-Tuset, Pablo Arbelaez, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale Combinatorial Grouping for Image Segmentation and Object Proposal Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):128–140, 2016. 1, 16

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. *Image*, 2:T2, 2021. 4

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, 2021. 6

[51] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems*, 28:91–99, 2015. 4, 6, 7, 8, 16

[52] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 5, 7, 12, 16

[53] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical Image Saliency Detection on Extended CSSD. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):717–729, 2015. 7, 13

[54] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing Objects with Self-Supervised Transformers and no Labels. In *British Machine Vision Conference*, 2021. 16

[55] Przemysław Skurowski, Hassan Abdulameer, J Błaszczyk, Tomasz Depta, Adam Kornacki, and P Kozieł. Animal Camouflage Analysis: CHAMELEON Database. *Unpublished Manuscript*, 2(6):7, 2018. 8

[56] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *International Conference on Learning Representations*, 2019. 3, 16

[57] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A Joint Model for Video and Language Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7473, 2019. 3

[58] Hao Tan and Mohit Bansal. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Conference on Empirical Methods in Natural Language Processing*, 2019. 3, 16

[59] Mingxing Tan and Quoc Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 2, 16

[60] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective Search for Object Recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013. 1, 3, 4, 5, 7, 13, 15, 16, 17

[61] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified Video Objects: A Benchmark for Dense, Open-World Segmentation. *arXiv preprint arXiv:2104.04691*, 2021. 6

[62] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. Towards Universal Object Detection by Domain Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7289–7298, 2019. 1, 4

[63] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly Simple Few-Shot Object Detection. *arXiv preprint arXiv:2003.06957*, 2020. 7

[64] Ross Wightman. PyTorch Image Models. https://github.com/rwightman/pytorch-image-models, 2019. 16

[65] Kesheng Wu, Ekow Otoo, and Arie Shoshani. Optimizing connected component labeling algorithms. In *Medical Imaging 2005: Image Processing*, volume 5747, pages 1965–1976. International Society for Optics and Photonics, 2005. 7, 8

[66] Zhe Wu, Li Su, and Qingming Huang. Cascaded Partial Decoder for Fast and Accurate Salient Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3907–3916, 2019. 7, 8, 13

[67] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A Large-scale Dataset for Object Detection in Aerial Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018. 1, 4, 12, 14

[68] Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region Similarity Representation Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2, 7, 17

[69] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. DetCo: Unsupervised Contrastive Learning for Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8392–8401, 2021. 2, 7, 17

[70] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with Noisy Student improves ImageNet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 16

[71] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging*, 5(3):036501, 2018. 13

[72] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency Detection via Graph-Based Manifold Ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3166–3173, 2013. 7, 13

[73] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-Vocabulary Object Detection Using Captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 6

[74] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In *International Conference on Machine Learning*, 2021. 7

[75] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12985–12994, 2020. 6

[76] Ziming Zhang, Yun Liu, Xi Chen, Yanjun Zhu, Ming-Ming Cheng, Venkatesh Saligrama, and Philip H.S. Torr. BING++: A Fast High Quality Object Proposal Generator at 100fps. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 40, pages 1209–1223, 2018. 16

[77] Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. UC2: Universal Cross-lingual Cross-modal Vision-and-Language Pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4155–4165, 2021. 6

[78] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*, 2021. 3, 4, 5, 6

[79] C Lawrence Zitnick and Piotr Dollár. Edge Boxes: Locating Object Proposals from Edges. In *The European Conference on Computer Vision*, pages 391–405. Springer, 2014. 1, 3, 4, 16

Appendix

In this section, we provide additional details regarding,

- Implementation details (Appendix A)
- Limitations (Appendix B)
- Qualitative results (Appendix C)
- Additional results (Appendix D)
- Related works (Appendix E)

A. Implementation Details

A.1. MDef-DETR

Similar to MDETR [28], we train MDef-DETR on approximately 1.25M aligned image-text pairs. The dataset is the same as used to pretrain MDETR on the modulated detection task. We train MDef-DETR for 20 epochs following the same hyper-parameters as in MDETR. Moreover, learning rates of 1e-4 and 1e-5 are used for the backbone and transformer respectively, which decays by a factor of 10 after 16 epochs. All the MDETR and MDef-DETR models are trained with ImageNet [52] pretrained ResNet-101 [23] backbone, if not mentioned explicitly.

Unlike MDETR which requires 40 training epochs for convergence, our MDef-DETR converges only in 20 epochs with better class-agnostic object detection (OD) accuracy. However, the inference for MDef-DETR is approximately 30% slower (see Table 11).

Model	Epochs	Parameters	Inference FPS	COCO AP50
MDETR	40	185M	13.02	40.66
MDef-DETR	20	188M	8.95	43.64

Table 11. Comparison of MDETR [28] and MDef-DETR (ours) in terms of convergence epochs, parameters, inference speed and class-agnostic OD performance on COCO [37] dataset. MDef-DETR converges in half epochs with better accuracy at the cost of slightly slower inference. The FPS are measured on a Quadro RTX 6000 GPU by averaging the time for $1K$ inference passes.

A.2. MViTs as Class Agnostic Object Detectors

We explore the interactive nature of multi-modal vision transformers (MViTs) for class-agnostic OD task. We construct intuitive natural language text queries by exploring the language semantic space of MViTs using an open-source natural language processing (NLP) library, spacy [24]. Specifically, we found words closer to the word '`object`' in semantic language space and constructed multiple queries for the class-agnostic OD task. The detected boxes from multiple text queries are combined, a class-agnostic non-maximum suppression (NMS) at IoU threshold of 0.5 is applied and top 'N' boxes are selected for evaluation. We use $N = 50$ and report average precision and recall at IoU threshold of 0.5 in all experiments. For the salient and camouflaged object detection (SOD and COD) tasks, we only consider boxes with objectness scores of greater than 0.7.

For Pascal VOC [14], COCO [37], Clipart, Comic and Watercolor [25], we use combined detections from queries '`all objects`', '`all entities`', '`all visible entities and objects`', and '`all obscure entities and objects`'. Whereas, we also include '`all small objects`' text query for the evaluation on KITTI [17], Kitchen [18] and DOTA [67] because these datasets have a large number of small sized objects. Additionally multi-scale evaluation is used for DOTA dataset due to a very significant scale variations in satellite imagery. Here the original image is split into 8 equal crops and the detections are combined by running inference on each crop. We observe that performing multi-scale inference improves the performance on DOTA where there are more tiny objects as compared to other datasets.

A.3. Detection of Small Objects

We observe that the targeted queries like '`all small objects`' and '`all little objects`' can improve the detection accuracy of small objects as compared to a rather general text query '`all objects`'. For quantitative comparison, all objects covering less than 5% of the image area are considered small, between 5% and 20% are considered medium and greater than 20% are considered large.

A.4. Open-world Object Detection

The proposals from MDef-DETR are used to generate the pseudo labels for unknown categories in Open-world Object Detector (ORE) [27] training. To avoid any data leakage, MDef-DETR is trained with a filtered dataset which is obtained by removing all the captions that contain any of the 60 unknown categories in ORE task-1. This filtering leaves us with a dataset having approximately 0.76M (out of 1.25M) image-text pairs. MDef-DETR is trained from scratch on this filtered dataset for 20 epochs and then used to produce unknown pseudo labels using class-agnostic object proposals.

To do so, firstly, the proposals with objectness score less than 0.7 are discarded. Secondly, all proposals having an IoU greater than 0.5 with any ground truth bounding box of a known category are removed. The rest of the proposals potentially belong to unknown categories and are used to supervise unknown detections in ORE training.

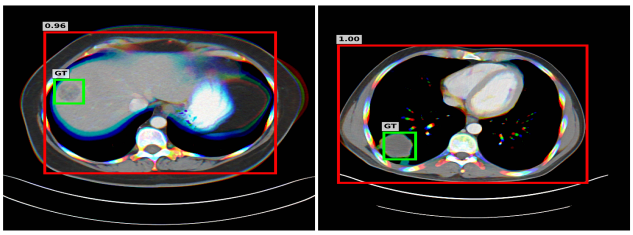

Figure 9. Illustration of MDef-DETR detections on the DeepLesion [71] dataset. The green boxes indicate the ground truth bounding box enclosing the lesion on the CT images and the red boxes are the class-agnostic predictions. The samples indicate a failure case of class-agnostic detection of MViT's on lesion detection dataset.

All the relevant codes and annotations will be available at https://git.io/J1HPY.

B. Limitations

Although MViTs (GPV-1 [20], MDETR [28] and MDef-DETR) show state-of-the-art class-agnostic OD performance across various dataset domains, they cannot be directly adapted to specialized out-of-domain detection tasks such as in medical imaging. In medical domain, lesion detection task involves locating the congenital malformations in different types of medical images including X-rays, CT scans, MRI scans and Ultrasoud. These applications require specialized data along with expert supervision (obtained from well-trained domain specialists) to perform well. Hence, in most cases, the general class-agnostic OD methods (*e.g.* MViTs) cannot be used in these applications directly. We evaluate the class-agnostic OD performance of MDef-DETR on DeepLesion [71] dataset (Fig. 9). We observe that such problems in medical imaging are not well addressed by the generic class-agnostic detection mechanism of MViTs trained on out-of-domain natural images. The ground truth annotations represented by the green boxes in Fig. 9, indicate that the target lesions do not well represent the concept of an object, and require expert based supervision to identify the abnormalities.

C. Qualitative Results

We present examples of class-agnostic predictions of MDETR and MDef-DETR across multiple domains, including natural image dataset Pascal VOC [14], MS COCO [37], autonomous driving dataset KITTI [17], sketches, painting and cartoons [25] and indoor Kitchen dataset [18] in Fig. 10. The detections are generated using the natural language text query, '`all objects`'.

In Fig. 11, we present some qualitative examples of class-agnostic OD with DETReg [3] trained using off-the-shelf proposals from Selective Search [60] in comparison with DETReg trained using MDef-DETR proposals. Fig. 12 shows some examples of improved Open-world detector (ORE) trained with MDef-DETR unknown pseudo labels. The images on the left of each example correspond to the ORE trained with unknown pseudo labels from RPN and on the right correspond to the ORE trained with unknown pseudo labels from MDef-DETR. The visualizations indicate that the improved model is better capable of detecting unknowns. Additionally, it reduces the miss-classifications of unknown categories with other known categories. For example, the second sample in Fig. 12 (top row - right side), corresponds to a sample in task 3 where 'laptop' belongs to the unknown categories set, was miss-classified as 'TV', which is however correctly classified as an unknown with the improved model. This is advantageous as it can better aid continual learning, *i.e.*, the model can learn about the unknown categories when additional information about the unknowns are obtained via supervision. In Fig. 13, we present examples of qualitative results obtained for salient OD and camouflaged OD with specif queries, '`all salient objects`' and '`all camouflaged objects`' respectively, along with the bounding box annotations from the ground truth masks.

D. Additional Results

D.1. Salient Object Detection

A common formulation of deep learning based Salient Object Detection (SOD) approaches is to predict a saliency map for each input image. We evaluate MDef-DETR against state-of-the art SOD approaches by converting the bounding box predictions of the the MViT model to masks using a COCO [37] trained Mask-RCNN [22] mask head. These converted masks are evaluated against the saliency predictions of PoolNet [38] and CPD [66] models on DUT-OMRON [72] and ECSSD [53] datasets (Table 12).

Dataset →	DUT-OMRON		ECSSD	
Model	MAE ↓	F-b ↑	MAE ↓	F-b ↑
CPD [66]	0.057	0.794	0.040	0.936
PoolNet [38]	0.054	0.866	0.038	0.954
MDef-DETR(Ours)	0.206	0.639	0.235	0.656

Table 12. Segmentation based evaluation of MDef-DETR on salient object detection in comparison with the state-of-the-art saliency approaches. The bounding box predictions of MDef-DETR for text query, '`all salient objects`' are converted to masks using COCO [37] trained mask head of Mask-RCNN [22].

D.2. Camouflaged Object Detection

In this section we compare camouflaged masks predictions of SINET-V2 [15] with MDef-DETR. Similar to SOD task, the bounding box predictions from the MViT are converted to object masks using the mask head of COCO

(a) Pascal VOC [14]

(b) MS COCO [37]

(c) Comic [25]

(d) Clipart [25]

(e) Watercolor [25]

(f) DOTA [67]

(g) Kitchen [18]

(h) KITTI [17]

Figure 10. class-agnostic detections of MViTs (MDETR [28] and MDef-DETR) on Pascal VOC (10a), MS COCO (10b), Comic (10c), Clipart (10d), Watercolor (10e), DOTA (10f), Kitchen (10g), KITTI (10h) dataset.

[37] trained Mask-RCNN [22] model. Following [16], S-measure (S_α), E-measure (E_ϕ), weighted F-measure (F_β^w) and mean absolute error (MAE) of mask predictions are reported in Table 13.

Model	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$
SINET-V2 [15]	0.783	0.867	0.660	0.042
MDef-DETR(ours)	0.491	0.533	0.275	0.267

Table 13. Comparison of masks prediction results of state-of-the-art COD model [15] with MDef-DETR. MDef-DETR proposals generated using 'all `camouflage objects`' query are converted to masks using COCO [37] trained mask head of Mask-RCNN [22].

D.3. Effect of Various Backbones

ResNet vs. EfficientNet: We explore the class-agnostic OD performance of MViTs for different convolutional

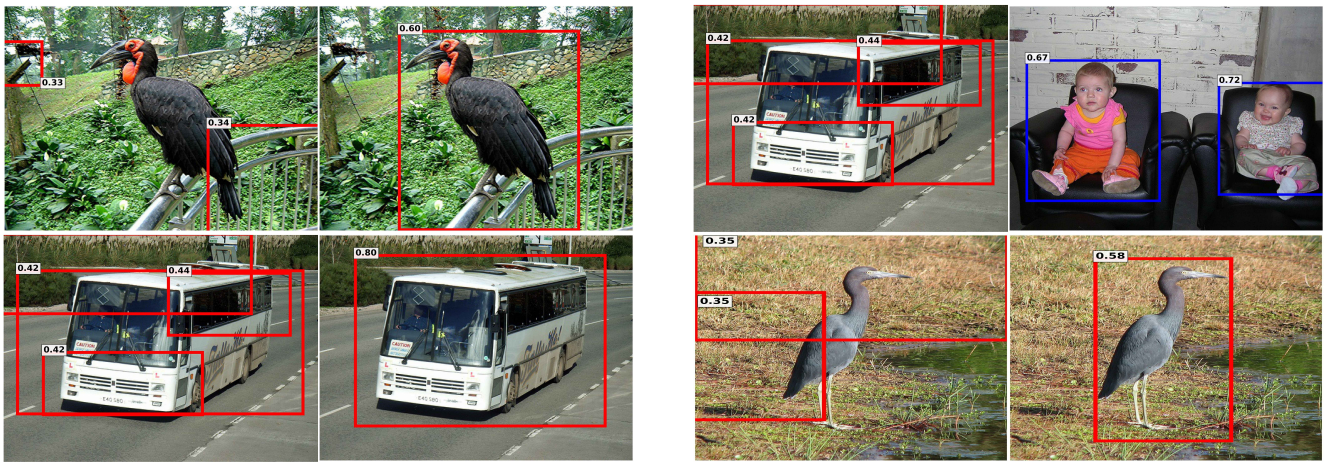

Figure 11. Class-agnostic OD performance of DETReg [3] trained using Selective Search [60] versus MDef-DETR proposals. The images on the left side of each example correspond to DETreg trained with Selective search and the images on the right side correspond to the one trained with MDef-DETR that results in better localized predictions.

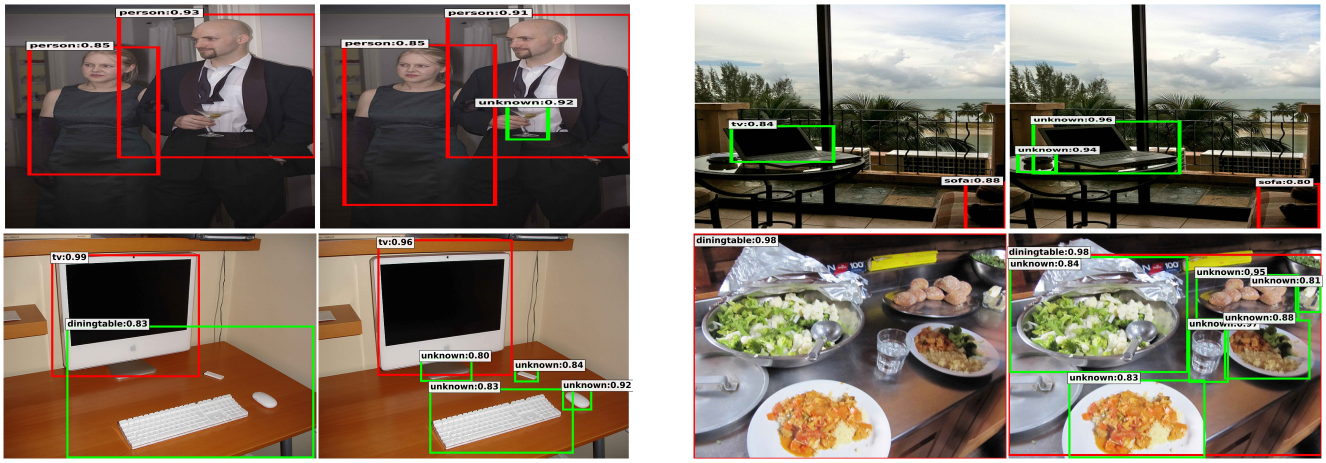

Figure 12. Qualitative results of unknown detections in ORE [27] when trained using RPN (**left**) versus MDef-DETR (**right**) unknown pseudo labels. Using proposals from MDef-DETR as unknown pseudo labels improves the prediction of unknowns.

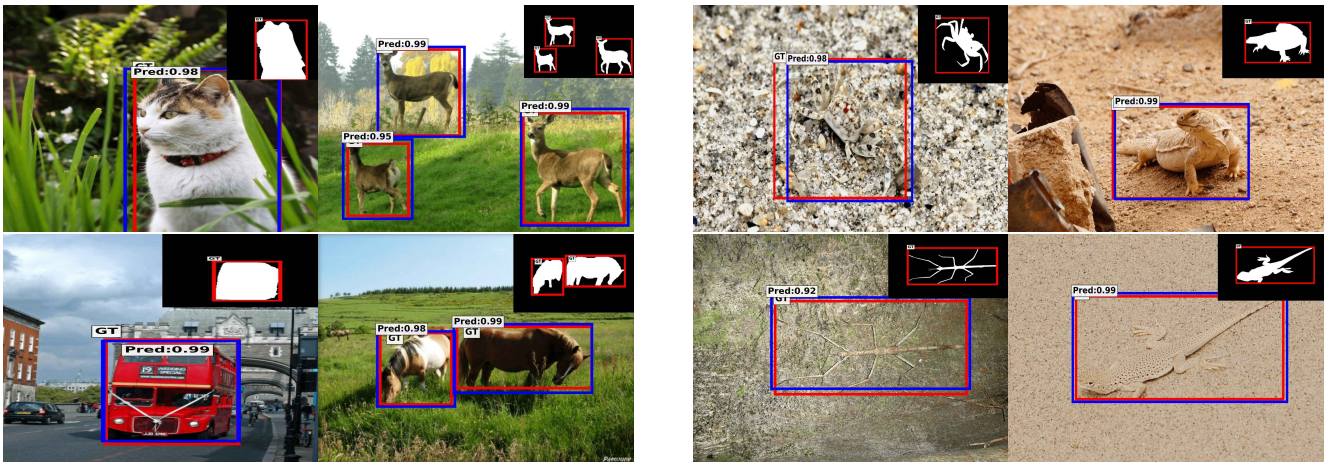

Figure 13. Qualitative results of MDef-DETR for Salient OD (**left**) and Camouflaged OD (**right**) tasks. The ground truth masks along with the generated bounding boxes are shown on top right of the images.

Dataset	Pascal VOC		COCO		KITTI		Kitchen		Clipart		Comic		Watercolor		DOTA	
Model	AP50	R-50	AP50	R-50	AP50	R-50	AP50	R-50	AP50	R-50	AP50	R-50	AP50	R-50	AP50	R-50
MDETR-R101	66.04	**90.10**	40.66	**62.15**	46.71	**67.24**	38.38	91.38	44.94	90.69	55.82	89.45	63.59	94.32	1.94	21.80
MDETR-E5	**69.61**	90.01	**42.34**	61.28	**48.06**	65.21	**53.26**	**91.50**	**62.34**	**92.73**	**69.86**	**90.46**	**74.40**	**94.98**	**3.71**	**24.88**

Table 14. Class-agnostic object detection performance of MDETR [28] for different convolutional backbones. The results indicate that the use of strong backbone improves the results especially on the out-of-domain (Kitchen [18], Calipart, Comic, Watercolor [25]) datasets.

backbones. Following [28], we compare the ResNet-101 [23] taken from Torchvision with EfficientNet-E5 [59] taken from Timm Library [64]. The ResNet model is trained on ImageNet [52] and achieves 77.4% top-1 accuracy on ImageNet validation, while the EfficientNet model is trained using Noisy-Student [70] on an additional 300M unlabelled images achieving 85.1% top-1 accuracy on ImageNet validation.

Table 14 indicates that using a stronger backbone improves the class-agnostic OD accuracy across different dataset domains. The performance boost is significant for out of domain datasets, Kitchen [18], Clipart, Comic and Watercolor [25], indicating better generalization of MViT when trained using a stronger backbone.

E. Related work

Class-Agnostic Detection: The class-agnostic OD is relatively less studied compared to class-aware detection. However, many object proposal generation algorithms have been proposed, since it remains a critical step in many applications like recognition and detection. The proposal generation algorithms can be categorized into three categories: (a) bottom-up segmentation based, (b) edge information based and (c) data-driven approaches based on deep neural network (DNN) architectures. In the first category that uses segmentation to derive proposals, multiple pixel groupings (superpixels) are merged according to various heuristics. Alexe *et al*. proposed an objectness [2] scoring method that combines various low-level features such as edges, color and superpixels to score object proposals. Selective Search [60] uses multiple hierarchical segmentations based on superpixels for object proposals. Similarly, MCG [48] uses segment hierarchy to group regions. Among the second category approaches, EdgeBoxes [79] scores bounding box proposals based on contours that the boxes enclose. BING algorithm [10, 76] generates binary features based on edge information for fast objectness estimation.

DNNs have also been investigated for generating object proposals. DeepBox [31] proposes a network that can be used to rerank any bottom-up proposals, *e.g*. the ones generated by EdgeBox [79]. DeepMask [46] generates rich object segmentations and an associated score of the likelihood of the patch to fully contain a centered object. A refinement of this method is proposed in SharpMask [47]. Alternatively, Ren *et al*. proposed region proposal network (RPN) [51] for generating object proposals, that identifies a set of regions that potentially contain objects along with corresponding objectness score. These are then refined for classification and localization for class-aware object detection. These are widely used in many two-stage objects detectors *e.g*. RCNN variants [22, 35, 51]. Jaiswal *et al*. proposed an adversarial framework [26] for class-agnostic object detection which replaces object type classification head with a binary classifier for class-agnostic detection. Another recent work proposes an Object Localization Network (OLN) [30] that replaces the classifier head in Faster-RCNN [51] with localization quality estimators such as centerness and IoU score for objectness estimation. Alternatively, Siméoni *et al*. proposed a method [54] that extracts features from a DINO [7] self supervised pre-trained transformer that uses patch correlations in an image to propose object proposals.

Multi-modal Transformers: Multi-modal Vision Transformers (MViT) typically involve learning task agnostic vision-language (V+L) representations using millions of image-text pairs and then transferring the knowledge to downstream tasks [9, 28, 34]. Inspired from the success of BERT [12] in natural language processing (NLP), VisualBERT [33], ViLBERT [41] and LXMERT [58] jointly learn V+L representations using image-caption pairs. They utilize a pretrained region proposal method [51] and learn the V+L correlation using self-supervised tasks such as mask language modeling and sentence image alignment. In a concurrent work, VL-BERT [56] performs pretraining on both text-only and visual-linguistic datasets and achieve an improved performance on multiple downstream visual comprehension tasks. UNITER [9] introduces Word-Region Alignment (WRA) pretraining task using Optimal Transport (OT) [45] which facilitates the alignment between text and image regions. It only masks one modality at a time while keeping the other modality intact which helps it to better capture the V+L relationships. We refer the readers to a recent survey [29] for a detailed treatment on ViTs.

All these methods utilize an off-the-shelf region proposal method [51] which usually produces noisy regions. OSCAR [34] tries to mitigate this problem by using object detector tags for modeling V+L understanding. It relies on the fact that the salient objects in the image are easy to detect and are typically mentioned in the caption. Alternatively, MDETR [28] leverages explicit alignment between text and ground-truth bounding boxes to learn visual-

language alignment. It builds on-top-of recently proposed DETR [5] model, generalizes to unseen concepts and outperforms the previous methods on many V+L downstream tasks. Going further, 12-in-1 [42] utilizes the pretrained V+L representations and performs a joint training of a single model on 12 datasets. This learning paradigm improves the single task performance as compared to the traditional task-wise training by achieving superior results on 11 out of 12 tasks. 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. It is also based on DETR [5] and uses aligned text-image pairs during its training, similar to [28]. We observe that these [20, 28] multi-modal transformers, which are trained using aligned image-text pairs, produce high quality object proposals by using simple text queries e.g., '`all objects`'.

Unsupervised Approaches: Recently, many unsupervised pretraining methods are proposed for the object detection task. Xiao *et al*. introduced ReSim [68] to encode both the region and global representations during self-supervised pretraining. In addition to the standard contrastive learning objective [8, 21], it slides a window in the overlapping region of the different views of an image and maximizes the feature similarity of the corresponding features across all convolutional layers. DetCo [69] approaches this problem by generating both the global views and local patches from an image and defines hierarchical global-to-global, local-to-local and global-to-local contrastive objectives. UP-DETR [11] proposes '`random query patch detection`' pretext task for pretraining of DETR [5]. The random patches from the image are generated and the model is trained on a large-scale dataset to locate these patches. DETReg [3] argues that it is necessary to pre-train both the backbone and the detection network for learning good representations for object detection downstream tasks. It utilizes an off-the-shelf selective search [60] proposal generation algorithm for acquiring pseudo labels for localization and pretrained contrastive clustering based SwAV [6] model for separating categories in the feature space. All these methods can be used for generating class-agnostic object proposals after the unsupervised pretraining. However, as shown in our analysis, the unsupervised approaches do not perform as well as the proposed class-agnostic OD framework based on supervised MViTs.