Supplemental Material

In this section, we provide additional information 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 MAVL

Similar to MDETR [29], we train MAVL on LMDet dataset containing approximately 1.3M aligned image-text pairs. Unlike MDETR which converges in 40 epochs, our MAVL converges only in 20 epochs with overall better class-agnostic object detection (OD) accuracy. However, the inference for MAVL is approximately 30% slower (see Table 10).

MAVL is trained using a learning rate of $1e^{-3}$ which decays by a factor of 10 after 16 epochs. The vision backbone (ResNet-101 [23]) and language backbone (RoBERTa [43]) use learning rates of $1e^{-4}$ and $1e^{-5}$ respectively. The number of object queries is set to 300. In the late-fusion transformer, a series of six self-attention blocks are used, where a detection head is applied after each block for calculating the individual auxiliary losses which are then summed up (see Fig. 2 in the main paper).

Table 10: Comparison of MDETR [29] and MAVL (ours) in terms of convergence epochs, parameters, inference speed and class-agnostic OD performance on COCO [40] dataset. MAVL converges in half epochs with better accuracy at the cost of slightly slower inference. The frames per second (FPS) are measured on a Quadro RTX 6000 GPU by averaging the time for $1K$ inference passes.

Model	Epochs	Parameters	Inference FPS	COCO AP50
MDETR	40	185M	13.0	40.7
MAVL	20	188M	8.95	43.6

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 semantic space of MViTs using an open-source natural language processing (NLP) library, spacy [24]. Specifically, we find words closer to the keyword '`object`' in the semantic space and construct multiple text queries for

the class-agnostic OD task. The detected boxes from the 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. 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 greater than 0.7.

For Pascal VOC [14], COCO [40], Objects365 [56], LVIS [19], Clipart, Comic and Watercolor [26], we use combined detections from queries '`all objects`', '`all entities`', '`all visible entities and objects`', and '`all obscure entities and objects`'. Additionally, '`all small objects`' text query is included for the evaluation on KITTI [17], Kitchen [18] and DOTA [72] as these datasets have a larger number of small sized objects. Moreover, multi-scale evaluation is used for DOTA dataset due to the significant scale variations in the satellite imagery. Here the original image is split into 8 equal crops and the detections from the individual crops are combined. We observe the multi-scale inference improves the performance on DOTA as it contains 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`'. Quantitative and qualitative results are presented in Fig. 4a (main paper). 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 MAVL are used to generate the pseudo-labels for unknown categories in Open-world Object Detector (ORE) [28] training. To avoid any data leakage, MAVL is trained on a subset of LMDet dataset, 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 1.3M) image-text pairs. MAVL is trained from scratch on this filtered dataset for 20 epochs and then used to produce unknown pseudo-labels using class-agnostic object proposal generation.

To do so, firstly, 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. Rest of the proposals potentially belong to unknown categories and are used as pseudo-labels of unknowns in ORE training. All relevant scripts and annotations will be publicly released.

B Limitations

Although MViTs (GPV-1 [20], MDETR [29] and MAVL) 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.

We evaluate the class-agnostic OD performance of MAVL on DeepLesion [76] dataset (Fig. 6). The ground-truth annotations represented by the green boxes in Fig. 6, indicate that the target lesions do not well represent the concept of an object, and require expert based supervision to identify the abnormalities. 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 direclty used. We observe that the generic class-agnostic detection mechanism of MViTs trained on out-of-domain natural images is not well-suited for generating proposals that can cater the need of specific medical applications.

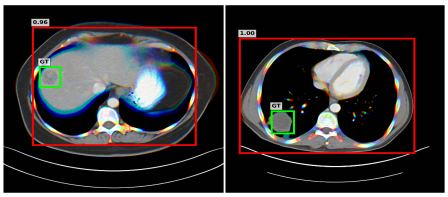
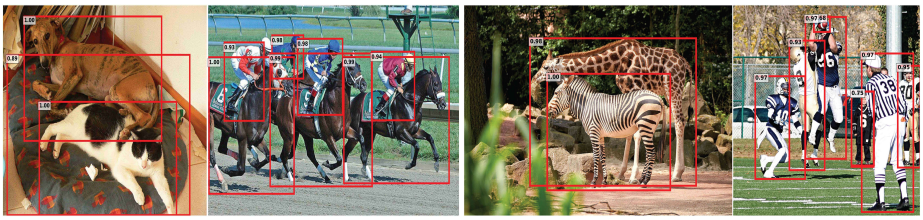


Fig. 6: Illustration of MAVL detections on the DeepLesion [76] dataset. The green boxes indicate the ground-truth bounding box enclosing the lesion on the CT images and the red boxes are the class-agnostic predictions. The samples indicate a failure case of class-agnostic detection of MViT's on lesion detection.

C Qualitative Results

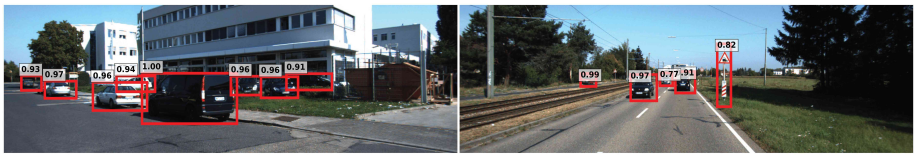
We present examples of class-agnostic predictions of MDETR and MAVL across natural image dataset Pascal VOC [14], COCO/LVIS [40,19], autonomous driving dataset KITTI [17] and indoor Kitchen dataset [18] in Fig. 7 and out-of-domain datasets that include sketches, painting, cartoons [26] and satellite images [72] in Fig. 8. The detections are generated using the natural language text query, '`all objects`'. In Fig. 9, we present some qualitative examples of class-agnostic OD with DETReg [3] trained using off-the-shelf proposals from Selective Search [64] in comparison with DETReg trained using MAVL proposals.

Fig. 10 shows some examples of improved Open-world detector (ORE) trained with MAVL 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 MAVL. The visualizations indicate that the improved model is better capable of detecting unknowns. Additionally, it reduces the misclassifications of unknown categories with other known categories. For example, the second sample in Fig. 10 (top row - right side), corresponds to a sample in task 3 where 'laptop' belongs to the unknown categories set, was misclassified as 'TV', which is however correctly classified as an unknown with the improved model. This is advantageous as it



(a) Pascal VOC [14] (b) MS COCO [40] / LVIS [19]

(c) Kitchen [18]



(d) KITTI [17]

Fig. 7: Class-agnostic detections of MViTs (MDETR [29] and MAVL) on natural image datasets, Pascal VOC, MS COCO/LVIS, Kitchen and KITTI.

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. 11, we present examples of qualitative results obtained for salient OD and camouflaged OD with specific queries, '`all salient objects`' and '`all camouflaged objects`' respectively, along with the bounding box annotations from the ground-truth masks.

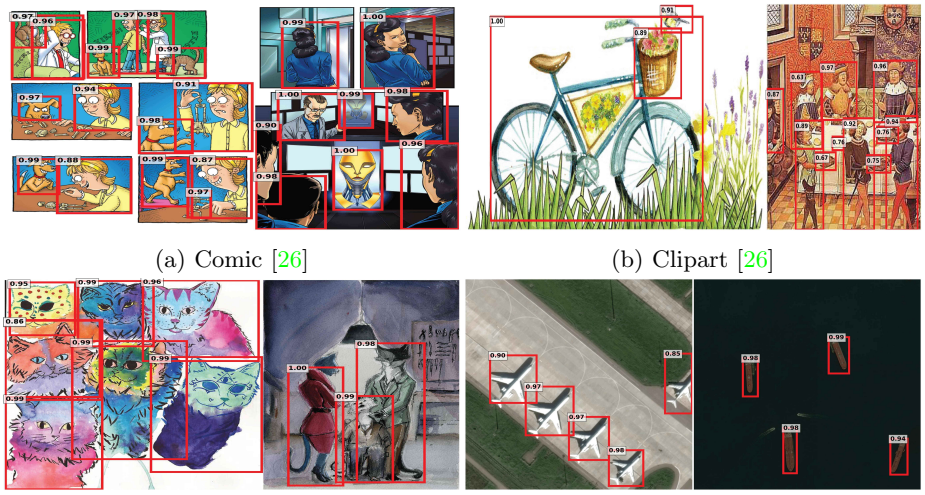
D Additional Results

D.1 Gains from MSDA in MAVL

We ablate the contribution of MSDA in Table 11 for our MAVL model. The class-agnostic OD results show the significance of MSDA.

D.2 Impact of Late Fusion in MAVL

The late fusion is crucial to our MAVL since it enables an efficient MViT design while keeping the multi-scale spatial information intact. Notably, early fusion (as in MDETR) ignores the spatial structure of images which makes it infeasible



(a) Comic [26] (b) Clipart [26]

(c) Watercolor [26] (d) DOTA [72]

Fig. 8: Class-agnostic detections of MViTs (MDETR [29] and MAVL) on out-of-domian datasets, Comic, Clipart, Watercolor and DOTA.

Fig. 9: Class-agnostic OD performance of DETReg [3] trained using Selective Search [64] versus MAVL 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 MAVL that results in better localized predictions

to operate with MSDA (that requires spatial information for deformable attention). Thus, MAVL effectively combines MSDA with late vision-text fusion and provides gain over MDETR in class-agnostic OD benchmarks. Unlike MDETR, our MAVL does not rely on contrastive alignment and thus removing MSDA significantly affects the results (Table 11).

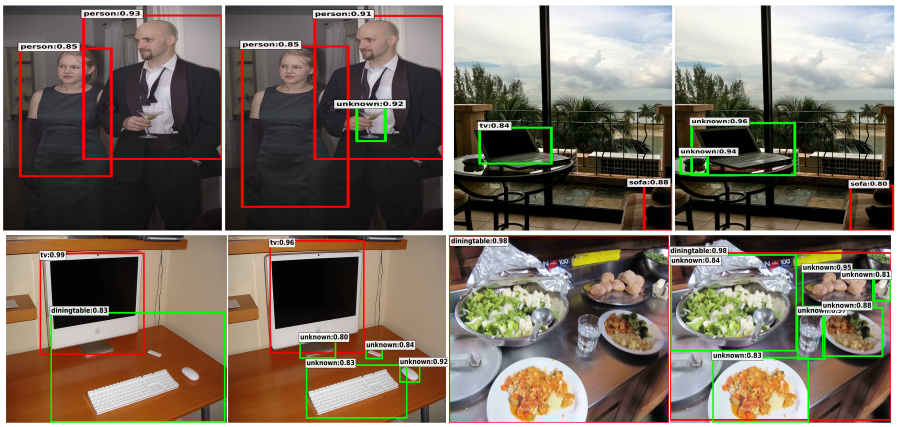


Fig. 10: Qualitative results of unknown detections in ORE [28] when trained using RPN (left) versus MAVL (right) unknown pseudo-labels. Using proposals from MAVL as unknown pseudo-labels improves the prediction of unknowns. It reduces the misclassifications of unknown categories with other known categories. The second example (shown in top row - right side), corresponds to a sample in task 3 where 'laptop' belongs to the unknown categories set, was misclassified as 'TV', which is however correctly classified as an unknown with the improved model. This better aids in continual learning.

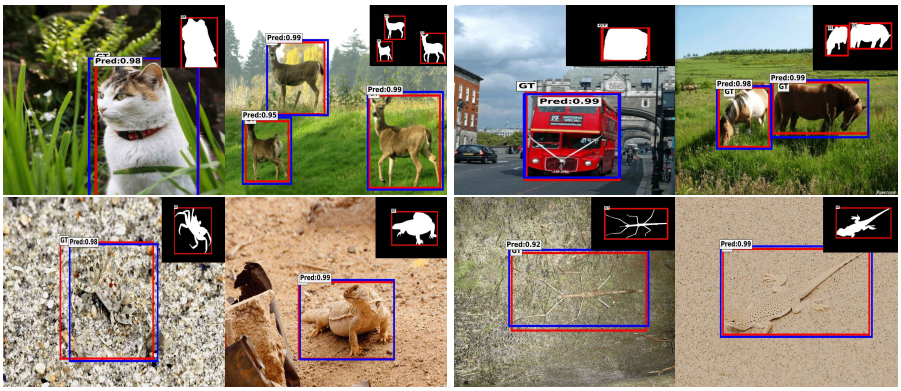


Fig. 11: **Top Rows**: Qualitative results of MAVL for Salient OD. **Bottom Rows**: Camouflaged OD (right) tasks. The ground-truth masks along with the generated bounding boxes are shown on top right of the image

D.3 Generalization Ability onto Novel/Rare Classes

Table 12 shows quantitative results on LVIS rare, common and frequent categories. (1) Similar to frequent and common, our MAVL provides good recall rates for rare LVIS categories, indicating its robust class-agnostic behavior. We note that most of the rare category instances in LVIS are of tiny size (area $<7\times7$ pixels) and have low recall ($\sim$19%) as compared to the medium/large instances

Model	Pascal-VOC AP50	Pascal-VOC R50	MSCOCO AP50	MSCOCO R50	KITTI AP50	KITTI R50
MAVL w/o MSDA	59.9	82.4	33.3	51.6	33.2	50.1
MAVL	65.0	89.1	39.3	62.0	39.0	61.0

Table 11: Effect of removing MSDA from MAVL. It decreases the class-agnostic OD performance, indicating the importance of MSDA. The models are evaluated after 10 epochs for ablation.

with much high recall (∼86%). (2) MAVL-ORE is trained by removing 60/80 common COCO categories from LMDet leaving only 0.76M/1.3M image-text pairs. This strict setting with much less training data also shows favorable rare class recall.

Model	Lang.	Rare	Common	Frequent	All
1:MAVL	×	30.0	31.6	32.4	32.1
2:MAVL	✓	38.0	40.5	37.1	37.0
3:MAVL-ORE	✓	33.4	36.7	33.2	33.1

Table 12: Class-agnostic recall (R50) of MAVL on LVIS rare, common and frequent categories. MAVL-ORE is trained on a filtered dataset generated by removing all captions listing any of 60 unknown COCO categories evaluated in ORE [28].

D.4 Querying All Class Names

Table 13 shows the class-agnostic OD results of MAVL when queried using a general (e.g., combination of queries in Table 3) versus combining detections from all category specific queries. Specifically, we use query '`every` < `category name` >' for each category of a dataset and combine proposals using class-agnostic NMS. We note that MAVL generates better *class-agnostic* detections with *general* text queries.

Model	Pascal-VOC AP50	Pascal-VOC R50	MSCOCO AP50	MSCOCO R50	KITTI AP50	KITTI R50
MAVL (ours)	68.6	91.3	43.6	65.0	48.2	63.5
MAVL (cat-wise)	61.7	91.2	36.7	64.6	47.7	59.8

Table 13: Comparison of using general versus category-specific queries for class-agnostic OD on three datasets.

D.5 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 MAVL

against state-of-the art SOD approaches by converting the bounding box predictions of the the MViT model to masks using a COCO [40] trained Mask-RCNN [22] mask head. These converted masks are evaluated against the saliency predictions of PoolNet [41] and CPD [71] models on DUT-OMRON [77] and ECSSD [57] datasets (Table 14a). Following [41] and [71], F-measure (F_b) and mean absolute Error (MAE) are reported.

Table 14: Segmentation based evaluation of MAVL on salient and comouflaged object detection in comparison with the corresponding state-of-the art approaches. The MAVL proposals are converted to masks using COCO [40] trained mask head of Mask-RCNN [22].

Dataset →	DUT-OMRON		ECSSD	
Model	MAE ↓	F-b ↑	MAE ↓	F-b ↑
CPD [71]	0.06	0.79	0.04	0.94
PoolNet [41]	0.05	0.87	0.04	0.95
MAVL(Ours)	0.21	0.64	0.24	0.66

(a) MAVL proposals from text query, '`all salient objects` 'are used.

Model	S_α ↑	E_ϕ ↑	F_β^w ↑	MAE ↓
SINET-V2 [15]	0.78	0.87	0.66	0.04
MAVL(ours)	0.49	0.53	0.28	0.27

(b) MAVL proposals generated using '`all camouflaged objects`' query are used.

D.6 Camouflaged Object Detection

In this section, we compare camouflaged masks predictions of SINET-V2 [15] with MAVL. Similar to SOD task, the bounding box predictions from the MViT are converted to object masks using the mask head of COCO [40] trained Mask-RCNN [22] model. Following [16], S-measure (S_α), E-measure (E_ϕ), weighted F-measure (F_β^w) and MAE of mask predictions are reported in Table 14b.

D.7 Effect of Various Backbones

ResNet vs. EfficientNet: We explore the class-agnostic OD performance of MViTs for different convolutional backbones. Following [29], we compare the ResNet-101 [23] taken from Torchvision with EfficientNet-E5 [63] taken from Timm Library [68]. The ResNet model is trained on ImageNet [55] and achieves 77.4% top-1 accuracy on ImageNet validation, while the EfficientNet model is trained using Noisy-Student [75] on an additional 300M unlabelled images achieving 85.1% top-1 accuracy on ImageNet validation.

Table 15 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 [26], indicating better generalization of MViT when trained using a stronger backbone model.

Table 15: Class-agnostic object detection performance of MDETR [29] for different convolutional backbones. The results indicate that the use of strong backbone improves the results especially on the out-of-domain (Kitchen [18], Clipart, Comic, Watercolor [26]) datasets.

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	66.0	90.1	40.7	62.2	46.7	67.2	38.4	91.4	44.9	90.7	55.8	89.5	63.6	94.3	1.94	21.8
MDETR-E5	**69.6**	90.0	**42.3**	61.3	**48.1**	65.2	**53.3**	91.5	**62.3**	92.7	**69.9**	90.5	**74.4**	95.0	**3.71**	24.9

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 [64] uses multiple hierarchical segmentations based on superpixels for object proposals. Similarly, MCG [52] uses segment hierarchy to group regions. Among the second category approaches, EdgeBoxes [84] scores bounding box proposals based on contours that the boxes enclose. BING algorithm [10,81] generates binary features based on edge information for fast objectness estimation.

DNNs have also been investigated for generating object proposals. DeepBox [33] proposes a network that can be used to rerank any bottom-up proposals, *e.g.,* the ones generated by EdgeBox [84]. DeepMask [49] 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 [50]. Alternatively, Ren *et al.* proposed region proposal network (RPN) [54] 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 [54,22,38]. Jaiswal *et al.* proposed an adversarial framework [27] 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) [31] that replaces the classifier head in Faster-RCNN [54] with localization quality estimators such as centerness and IoU score for objectness estimation. Alternatively, Siméoni *et al.* proposed a method [58] 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 [37,9,29]. Inspired from the success of BERT [12] in natural language processing (NLP), VisualBERT [36], ViLBERT [44] and LXMERT [62] jointly learn V+L representations using image-caption pairs. They utilize a pretrained region proposal method [54] 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 [60] 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) [48] 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.

All these methods utilize an off-the-shelf region proposal method [54] which usually produces noisy regions. OSCAR [37] 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 [29] 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 [45] 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 typical 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]. We observe that these [29,20] 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 [73] to encode both the region and global representations during self-supervised pretraining. In addition to the standard contrastive learning objective [21,8], 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 [74] 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 [64] 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.