

Visual Modality Prompt for Adapting Vision-Language Object Detectors

Heitor R. Medeiros* Atif Belal Srikanth Muralidharan
Eric Granger Marco Pedersoli
LIVIA, Dept. of Systems Engineering, ETS Montreal, Canada
International Laboratory on Learning Systems (ILLS)

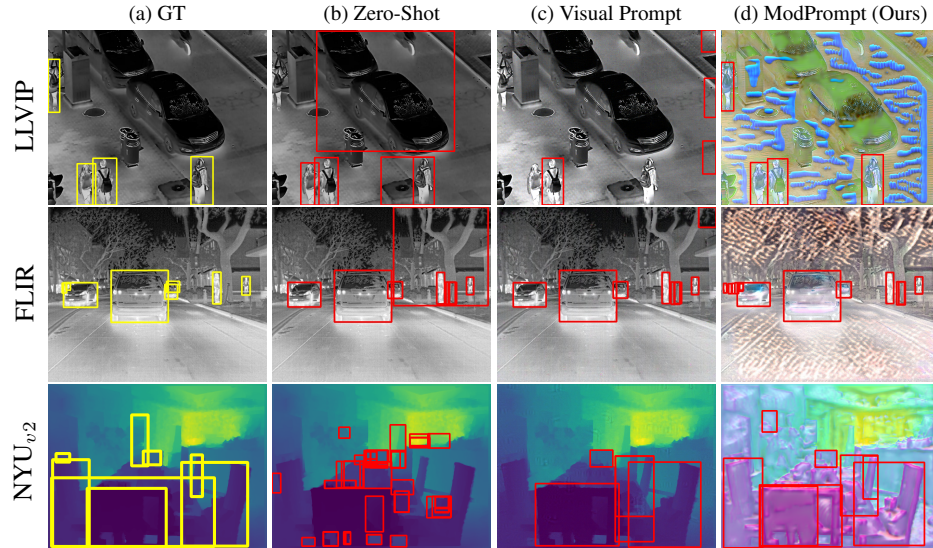


Figure 1. **YOLO-World detections of different approaches across modalities:** LLVIP and FLIR datasets (infrared) and NYU_{v2} (depth). Each column corresponds to a different approach: **(a) Ground Truth (GT):** Shows in yellow the GT bounding boxes for objects. **(b) Zero-Shot (ZS):** Displays detections (in red) from a zero-shot model. It has missed several detections and some inaccurate boxes without tuning. **(c) Visual Prompt:** Illustrates detections with a visual prompt added to the image. It shows improvements over ZS, with more accurate detection in certain areas, but still misses some objects. **(d) ModPrompt (Ours):** Detections from our model. ModPrompt generates artifacts on the image to enhance objects and suppress the background, facilitating detection.

Abstract

The zero-shot performance of object detectors degrades when tested on different modalities, such as infrared and depth. While recent work has explored image translation techniques to adapt detectors to new modalities, these methods are limited to a single modality and traditional detectors. Recently, vision-language detectors (VLDs), such as YOLO-World and Grounding DINO, have shown promising zero-shot capabilities; however, they have not yet been adapted for other visual modalities. Traditional fine-tuning approaches compromise the zero-shot capabilities of the detectors. The visual prompt strategies commonly used for classification with vision-language models apply the same linear prompt translation to each image, making them less effective. To address these limitations, we propose Mod-

Prompt, a visual prompt strategy to adapt VLDs to new modalities without degrading zero-shot performance. In particular, an encoder-decoder visual prompt strategy is proposed, further enhanced by the integration of inference-friendly modality prompt decoupled residual, facilitating a more robust adaptation. Empirical benchmarking results show our method for modality adaptation on YOLO-World and Grounding DINO for challenging infrared (LLVIP, FLIR) and depth (NYU_{v2}) datasets, achieving performance comparable to full fine-tuning while preserving the model’s zero-shot capability. Our code is available at <https://github.com/heitorrapela/ModPrompt>.

1. Introduction

Object detection (OD) is a key challenge in computer vision, aiming to localize and classify objects in images [57].

*Contact: heitor.rapela-medeiros.1@ens.etsmtl.ca

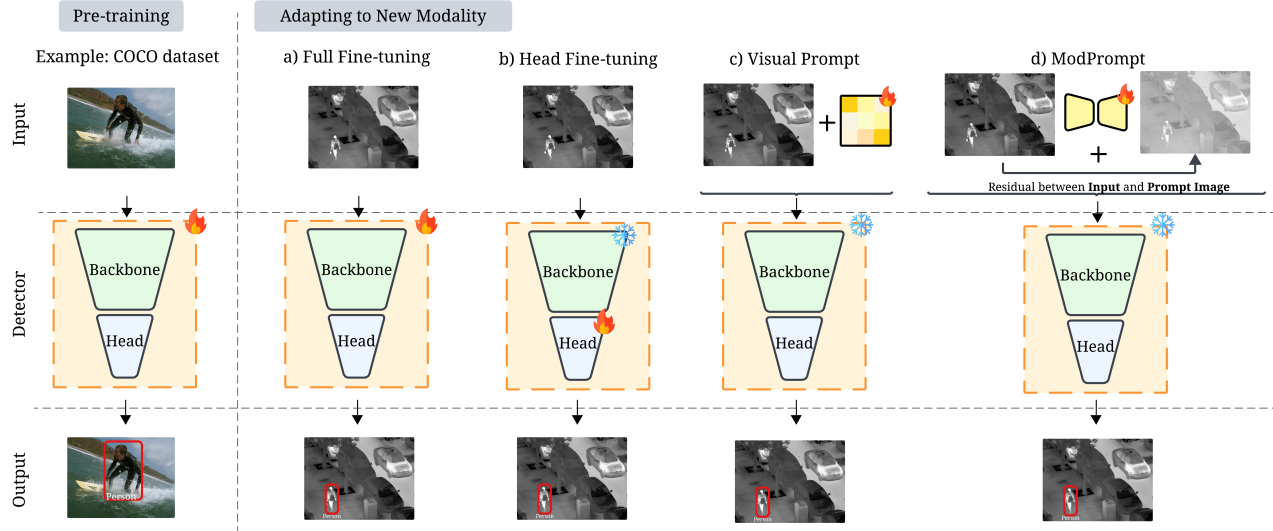


Figure 2. **Strategies to adapt object detectors to new modalities:** (a) **Full Fine-tuning:** Both the backbone (the part of the model responsible for feature extraction) and the head (responsible for the final output, like object detection) are updated with new training data. (b) **Head Fine-tuning:** Only the head is fine-tuned while the backbone remains frozen. (c) **Visual Prompt:** Uses a visual prompt added to the input. The backbone and head remain unchanged, but the visual prompt guides the model to better interpret the new modality. (d) **Our Modality Prompt.** Similarly to a visual prompt, the input image is added to a visual prompt. The main difference is that here the prompt is not static; it is a transformation of the input image.

Recent OD advancements, driven by applications like autonomous driving [35, 44], surveillance [10, 40], and robotics [11, 19, 38], have improved real-world systems significantly. Although traditional detectors like YOLO [21] or DINO [52] perform well, they are limited to fixed vocabularies, such as the 80 categories in the COCO [29] dataset, and exhibit poor zero-shot performance [2]. These detectors are trained on predefined categories, restricting their adaptability to varying scenarios. In contrast, Vision-Language Models (VLMs) combine visual representations and semantic text embeddings, allowing them to dynamically understand pixel-level features in context [39]. VLMs enhance accuracy and flexibility by recognizing unseen objects through the integration of language, paving the way for open-vocabulary detection [49], such as Grounding DINO [30] and YOLO-World [7].

Generally, these open-vocabulary detectors are pre-trained on large-scale RGB datasets, making them capable of zero-shot detection [30, 51]. However, their performance degrades if the domain shift is large, as in a different modality, such as infrared or depth. A common approach to mitigate this is by fine-tuning the VLMs on the downstream modality dataset. However, it makes the VLMs lose their zero-shot capabilities due to catastrophic forgetting [24]. Some techniques have been explored to adapt the VLMs to downstream tasks for image classification. These include text prompt tuning (TPT) [54, 55] and visual prompt tuning (VPT) [1, 20]. Unlike traditional fine-tuning, prompt tuning involves adding learnable prompts to the model’s in-

put, allowing it to remain unchanged and preserve its zero-shot capability. VPT methods for classification learn visual prompt tokens to adapt the classifier of transformer-based models to the downstream tasks. Visual prompts are effective when adapting VLMs to a downstream task of the same modality (normally RGB). However, these methods are also less effective if the downstream data is from another visual modality with a large modality shift between pre-trained and downstream data, as shown in Table 1.

In addition to incorporating language into detectors, another major driver of advancements in real-world applications is the integration of diverse visual modalities, such as infrared [26, 34, 45] and depth [27, 46, 47]. These modalities provide additional spatial information, enhancing visibility in low-light and obstructed conditions, resulting in more robust and accurate object detectors [4]. Adapting detectors for different modalities unlocks their potential in specialized contexts by combining the efficiency of transfer learning with modality-specific enhancements. This approach bridges the gap between general-purpose detectors capable of zero-shot detection and the nuanced demands of novel modalities, where labels are still scarce compared to RGB data.

Therefore, to overcome limitations in modality adaptation, image translation [16, 36] methods have been explored for traditional detector-based models. However, these approaches primarily optimize the image translation loss rather than directly focusing on improving detection performance. More recent approaches, such as Hallu-

ciDet [33] and ModTr [32], have demonstrated improvement by emphasizing detection loss optimization during adaptation. However, these methods are limited to traditional detectors without language mechanisms, and they did not explore more challenging modalities such as depth. For HalluciDet, the prior pre-trained knowledge is lost during the adaptation phase. In ModTr, the COCO pre-trained knowledge is preserved, but it lacks the zero-shot powerful capability of VLMs. In contrast, our work leverages recent vision-language detectors, incorporating textual information and allowing for adaptation to different modalities, exploring the benefits of both visual and text information for improving modality adaptation.

In this paper, we propose ModPrompt, a visual prompt strategy to adapt vision-language ODs in the input space to new visual modalities without sacrificing zero-shot knowledge of the vision-language detector. ModPrompt is a detector-agnostic encoder-decoder visual prompt strategy that can easily be integrated with any VLM detector, regardless of the detector’s backbone type. We also introduce an inference-friendly residual mechanism named Modality Prompt Decoupled Residual (MPDR) capable of adapting the learnable text embeddings with the ModPrompt loss, while preserving full knowledge of the model due to the decoupled embedding parameters. We demonstrate that the combination of MPDR with ModPrompt improves the performance of the detector across different modalities while retaining the original zero-shot knowledge, both in the vision and language components. To the best of our knowledge, this is the first work that focuses on adapting state-of-the-art VLM-based OD models to new modalities. Furthermore, ModPrompt is validated on two state-of-the-art VLM ODs, Grounding DINO and YOLO-World, and across four widely used visual modalities, infrared, depth, event-based, and LiDAR.

Our main contributions can be summarized as follows.

- (1) We introduce ModPrompt, a novel pixel-level prompt adaptation strategy based on input translation. Our method is backbone-agnostic, leading to improved adaptability of vision-language detectors across different modalities and preserving the vision encoder knowledge. Additionally, we introduce MPDR, a text-embedding modality prompt decoupled residual mechanism, which preserves full knowledge of the text-embedding while increasing the performance of the model for the novel modality.
- (2) We conduct a detailed investigation of various pixel-level visual prompt techniques, offering insights into why traditional prompt methods struggle with adapting object detectors to new modalities. Our analysis demonstrates how ModPrompt overcomes these limitations to significantly enhance detection performance.
- (3) We provide a comprehensive benchmark on the vi-

sual prompts for modality adaptation across two primary open-vocabulary OD models, YOLO-World and Grounding DINO. We evaluate our method across infrared, depth, event-based, and LiDAR datasets, achieving in some cases a level of performance comparable to full fine-tuning while preserving the model’s zero-shot capability.

2. Related Works

Open-Vocabulary Object Detection. Recent advances in open-vocabulary OD exploit vision-language models (VLMs) with region-level multimodal alignment to recognize novel concepts. RegionCLIP [53] uses pseudo labels for contrastive pretraining, while Gu et al. [14] employ knowledge distillation with CLIP-based region embeddings. Lin et al. [28] leverage bipartite matching for image-text alignment. GLIP [25] and GLIPv2 [51] enhance zero-shot detection via multimodal fusion and region-grounded pretraining. Grounding DINO [30] introduces multiphase image-text fusion, and YOLO-World [7] enables efficient open-vocabulary detection with reparametrizable fusion.

Downstream Adaptation of Large Vision Models.

Recent work has focused on parameter-efficient adaptation of large vision models without erasing prior knowledge. Co-op [55] learns continuous text prompts for downstream tasks. Visual Prompt Tuning [20] adds tunable parameters at various network stages. Bahng et al. [1] propose image-space linear probing. CLIP-Adapter [12] introduces feature-space adaptation via bottlenecks and residuals. Yu et al. [48] use task residuals to decouple new parameters for classification. While these methods seem effective, they are not suitable for detection.

Image Translation for Detection.

Image translation methods adapt images across domains by modifying domain-specific characteristics while preserving content [37], often using generative models [13, 23]. Pix2Pix [18] leverages paired supervision, while CycleGAN [56] enables unpaired translation. For OD, HalluciDet [33] performs a two-step adaptation using RGB fine-tuning and translation, whereas ModTr [32] employs zero-shot COCO-trained detectors with fusion strategies to preserve prior knowledge. In contrast, we adapt large VLM-based ODs (e.g., YOLO-World, Grounding DINO) to novel modalities (e.g., IR, depth) while retaining their zero-shot capabilities.

3. Proposed Method

3.1. Preliminary Definitions

Vision Language Object Detection. The training dataset for traditional object detectors can be represented as $\mathcal{D} =$

Dataset	Method	YOLO-World			Grounding DINO		
		AP ₅₀	AP ₇₅	AP	AP ₅₀	AP ₇₅	AP
LLVIP - IR	Zero-Shot (ZS)	81.00 ± 0.00	57.80 ± 0.00	53.20 ± 0.00	85.50 ± 0.00	62.70 ± 0.00	56.50 ± 0.00
	Head Finetuning (HFT)	93.57 ± 0.05	73.83 ± 0.19	64.80 ± 0.08	87.53 ± 0.06	65.57 ± 0.23	58.10 ± 0.20
	Full Finetuning (FT)	97.43 ± 0.05	77.93 ± 0.21	67.73 ± 0.09	97.17 ± 0.31	79.93 ± 0.83	67.83 ± 0.96
	Visual Prompt (Fixed)	70.30 ± 7.89	45.67 ± 6.97	43.53 ± 5.79	83.83 ± 0.06	61.53 ± 0.23	55.13 ± 0.15
	Visual Prompt (Random)	60.13 ± 0.29	38.73 ± 0.17	36.87 ± 0.12	83.87 ± 0.06	61.37 ± 0.06	55.03 ± 0.06
	Visual Prompt (Padding)	79.87 ± 1.00	51.77 ± 0.90	49.30 ± 0.83	82.73 ± 0.31	60.00 ± 0.35	55.13 ± 0.15
	Visual Prompt (WM)	82.00 ± 1.59	53.90 ± 1.06	50.90 ± 0.94	69.57 ± 0.93	41.37 ± 1.27	40.77 ± 0.87
	Visual Prompt (WM _{v2})	74.10 ± 0.43	46.47 ± 0.62	44.70 ± 0.22	69.87 ± 1.12	41.77 ± 1.30	41.13 ± 0.96
	ModPrompt (Ours)	92.80 ± 0.29	70.73 ± 1.02	62.87 ± 0.63	93.13 ± 0.15	67.17 ± 0.78	60.10 ± 0.50
NYU _{v2} - Depth	Zero-Shot (ZS)	04.80 ± 0.00	03.10 ± 0.00	03.00 ± 0.00	08.30 ± 0.00	05.60 ± 0.00	05.30 ± 0.00
	Head Finetuning (HFT)	21.03 ± 0.12	12.03 ± 0.37	12.00 ± 0.14	18.24 ± 0.85	13.23 ± 0.46	12.10 ± 0.21
	Full Finetuning (FT)	49.90 ± 0.08	36.40 ± 0.29	33.57 ± 0.26	51.60 ± 2.09	39.17 ± 1.88	35.77 ± 1.70
	Visual Prompt (Fixed)	04.67 ± 0.05	03.07 ± 0.05	02.90 ± 0.00	08.27 ± 0.06	05.57 ± 0.06	05.27 ± 0.06
	Visual Prompt (Random)	04.23 ± 0.12	02.63 ± 0.05	02.53 ± 0.05	08.33 ± 0.06	05.53 ± 0.06	05.27 ± 0.06
	Visual Prompt (Padding)	03.97 ± 0.05	02.50 ± 0.00	02.43 ± 0.05	07.63 ± 0.06	05.17 ± 0.06	04.63 ± 0.06
	Visual Prompt (WM)	10.67 ± 0.12	06.90 ± 0.22	06.57 ± 0.05	04.87 ± 0.06	02.97 ± 0.06	02.97 ± 0.06
	Visual Prompt (WM _{v2})	10.63 ± 0.12	06.67 ± 0.12	06.50 ± 0.08	05.00 ± 0.10	03.07 ± 0.06	03.03 ± 0.06
	ModPrompt (Ours)	37.17 ± 0.57	27.50 ± 0.64	24.93 ± 0.50	21.70 ± 0.20	15.03 ± 0.29	14.13 ± 0.21

Table 1. **Detection performance (APs) for YOLO-World and Grounding DINO for the two main datasets evaluated: LLVIP-IR and NYU_{v2}-Depth.** The different visual prompt adaptation techniques are compared with our ModPrompt, and the zero-shot (ZS), head finetuning (HFT), and full finetuning (FT) are also reported, where the full finetuning is the upper bound.

$\{(X, Y)\}_{j=1}^M$. Here, $X \in \mathbb{R}^{W \times H \times C}$ represents an image with dimensions $W \times H$ and C channels, and $Y = \{(b_i, c_i)\}_{i=1}^N$ consist of bounding boxes b_i and object category c_i . For the visual language model, the annotations are reformulated as a region-text pair $Y = \{(b_i, t_i)\}_{i=1}^N$, where t_i corresponds to the text for the region b_i . Specifically, t_i is the name of the object category in b_i . The objective of the object detection task is to accurately identify all objects in a given image. The average precision (AP) metric across classes is used as the standard evaluation protocol. An object detector is formally represented as the mapping $f_\theta : \mathbb{R}^{W \times H \times C} \rightarrow \hat{Y}$, where θ denotes the parameter vector. The detection cost function $\mathcal{C}_{det}(\theta)$, a differentiable proxy for the AP metric, is computed as an average loss of the detection loss $\mathcal{L}_{det}$ over the entire dataset $\mathcal{D}$, described mathematically as:

$$\mathcal{C}_{det}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(x, Y) \in \mathcal{D}} \mathcal{L}_{det}(f_\theta(x), Y). \quad (1)$$

Visual Prompt for OD. Given a frozen pre-trained detector, the objective of the Visual Prompt method is to learn a task-specific visual prompt v_ϕ , parametrized by ϕ , which can be combined with the input x to improve the final detection performance. During test time, this visual prompt is added to the test images to adapt them for the desired detection task. As described in the following equation:

$$\mathcal{C}_{vp}(\phi) = \frac{1}{|\mathcal{D}|} \sum_{(x, Y) \in \mathcal{D}} \mathcal{L}_{det}(f_\theta(x + v_\phi), Y). \quad (2)$$

3.2. ModPrompt

The visual prompt methods discussed above learn simple linear transformations, like adding a fixed patch to each image. These transformations are learned during the training process and are not conditioned on the input image at inference, making them less effective. In ModPrompt, we incorporate a function h_ϑ , an encoder-decoder inspired by U-Net [41], dependent on the input image x , which is trained conditioned on labels Y . Our encoder-decoder is flexible in terms of the encoder backbone initialized from pre-trained weights. We adapt the last layer of the decoder to constrain it to be 3-channels with values ranging between 0 and 1 to simulate a pseudo-RGB image. Note that this mechanism is completely different from traditional U-Nets, which are trained to reconstruct and generate c -channels for semantic segmentation tasks, requiring both input and ground truth segmentation map during training. However, ModPrompt requires access only to the pre-trained RGB detector and final target modality data, so it is based on guidance detection loss instead of reconstruction. The ModPrompt training cost is defined by the following equation:

$$\mathcal{C}_{mp}(\vartheta) = \frac{1}{|\mathcal{D}|} \sum_{(x, Y) \in \mathcal{D}} \mathcal{L}_{det}(f_\theta(x + h_\vartheta(x)), Y). \quad (3)$$

This translation function (f_θ) is responsible for adapting the input image to a modality representation optimally suited to address detection in the target modality. Our

idea of a visual ModPrompt is inspired by text conditional prompt [54] work, which incorporates a small network for the text encoder conditional adaptation. In our approach, the visual prompt depends on the input image.

3.3. Modality Prompt Decoupled Residual with Knowledge Preservation (MPDR)

The vision-language models take both text and image as input. ModPrompt aims to transform the input image to adapt to the new modality. For the adaptation of the textual part, text prompt tuning methods [54, 55] have been explored for classification. But these methods require back-propagating through the whole text backbone, making them computationally expensive. Recently, task residual [48] was proposed that tunes the precomputed text embedding of the vision-language classifiers. Inspired by this, we design a strategy to efficiently tune the embedding of the vision-language detectors. As illustrated in Figure 3, offline text embeddings are generated for each object category. Then, the MPDR, which are learnable embedding tokens, are tuned to adapt to a specific modality end-to-end with ModPrompt loss. The MPDR is crucial for decoupling the knowledge from the original embedding, which preserves prior text-embedding knowledge and is able to incorporate new knowledge during the adaptation stage. During the training, the loss of ModPrompt and MPDR work in synergy to adapt to any new modality. In the test phase, the MPDR can be deactivated easily by zero-masking its parameters, which is the same as full zero-shot embedding knowledge, or enabled, which is the adapted embedding for this modality (that’s why decoupling the knowledge is important) and doesn’t introduce any inference overhead. The following equation describes the final training cost:

$$\mathcal{C}_{\text{mp-tp}}(\vartheta, \phi) = \mathcal{C}_{\text{mp}}(\vartheta) + \mathcal{C}_{\text{tp}}(\phi), \quad (4)$$

where $\mathcal{C}_{\text{mp}}(\vartheta)$ is the optimization of parameters ϑ of the encoder-decoder function, and $\mathcal{C}_{\text{tp}}(\phi)$ is the text-prompt adaptation on the embedding space by training new parameters ϕ for the class embeddings in the offline vocabulary.

3.4. Training Summary

For training the ModPrompt with MPDR, we first pre-compute the text embeddings for the target classes, using the text encoder. Then, we sum the learnable embedding residuals with the frozen embeddings and train them together with the ModPrompt. Note that, different from other embedding adaptation approaches, our MPDR can be used with ModPrompt for full zero-shot knowledge preservation due to the residual decoupled embedding learning strategy. The cost function is based on Equation 4, but instead of the online cost function for text $\mathcal{C}_{\text{tp}}(\phi)$, we work on the text em-

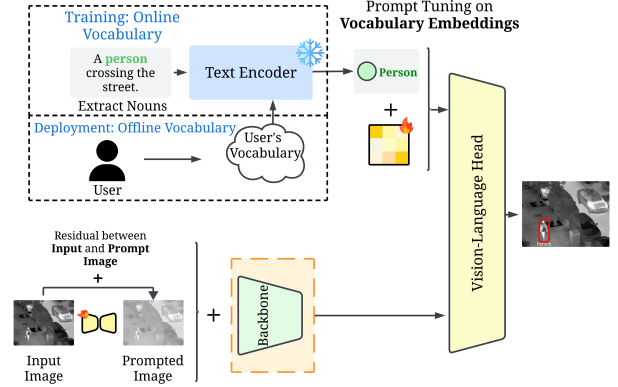


Figure 3. **Our proposed strategy for text-prompt tuning: an inference-friendly and knowledge-preserving decoupled embedding tuning method.** An offline embedding is generated for each object category, and then a novel decoupled residual trainable parameters and the ModPrompt are integrated into the detector to adapt it to new modalities.

bedding space, which is better suited for efficient training and inference.

4. Results and Discussion

4.1. Experimental Methodology

(a) **Datasets: LLVIP:** LLVIP is a surveillance dataset composed of 12,025 IR images in the training set and 3,463 IR images in the test set with only pedestrians annotated. **FLIR ALIGNED:** We used the sanitized and aligned paired sets provided by Zhang et al. [50]. It has 4,129 training IR images and 1,013 IR test images. FLIR images are captured from the perspective of a camera in the front of a car, with a resolution of 640 by 512. It contains the bicycles, dogs, cars, and people classes. Following Cao et al. [5], we decided to remove “dog” objects because the number of annotations is inadequate for training with the FLIR dataset; results of FLIR can be seen in the supp. material due to space constraints. **NYU_{v2}:** The NYU-Depth V2 [43] dataset provides a valuable collection of indoor video sequences captured with a Microsoft Kinect camera. In this dataset, we used the depth information, which is composed of 795 training images and 654 test images with a resolution of 640 by 480. The dataset comprised 19 different classes, such as bathtubs, beds, bookshelves, boxes, and chairs. In all our experiments, only IR and depth images were used.

(b) **Implementation Details:** The YOLO-World models were trained on an A100 NVIDIA GPU, and Grounding DINO on V100, and were implemented using PyTorch. For YOLO-World, we used the original work [7], which is based on MMDet [6], and for Grounding DINO, we used the open-source version from the MMDet library. For the YOLO-World, we use both YOLO-World-S-v1 and YOLO-

Detector	Method	LLVIP-IR			NYU _{v2} -DEPTH		
		AP ₅₀	AP ₇₅	AP	AP ₅₀	AP ₇₅	AP
YOLO-World	Fixed	86.60 ± 0.00 (+16.3)	63.53 ± 0.05 (+17.8)	57.67 ± 0.05 (+14.1)	12.47 ± 0.05 (+7.80)	07.50 ± 0.00 (+4.43)	07.37 ± 0.05 (+4.47)
	Random	86.43 ± 0.05 (+26.3)	63.40 ± 0.08 (+24.6)	57.50 ± 0.00 (+20.6)	12.23 ± 0.05 (+8.00)	07.30 ± 0.08 (+4.67)	07.20 ± 0.00 (+4.67)
	Padding	83.57 ± 0.12 (+3.70)	59.03 ± 0.05 (+7.26)	54.23 ± 0.05 (+4.93)	10.63 ± 0.09 (+6.66)	05.60 ± 0.08 (+3.10)	05.97 ± 0.05 (+3.54)
	WeightMap	87.47 ± 0.17 (+5.47)	62.23 ± 0.31 (+8.33)	57.43 ± 0.05 (+6.53)	13.43 ± 0.09 (+2.76)	08.07 ± 0.12 (+1.17)	07.90 ± 0.08 (+1.33)
	ModPrompt	96.60 ± 0.37 (+3.80)	77.27 ± 0.33 (+6.54)	67.37 ± 0.05 (+4.50)	44.67 ± 0.17 (+7.50)	32.53 ± 0.66 (+5.03)	29.93 ± 0.12 (+5.00)
Grounding DINO	Fixed	88.70 ± 0.66 (+4.87)	67.73 ± 0.50 (+6.20)	60.13 ± 0.45 (+5.00)	10.00 ± 0.00 (+1.73)	06.40 ± 0.28 (+0.83)	06.40 ± 0.28 (+1.13)
	Random	88.80 ± 0.95 (+4.93)	68.00 ± 1.21 (+6.63)	60.37 ± 0.90 (+5.33)	09.90 ± 0.14 (+1.57)	06.50 ± 0.00 (+0.97)	06.15 ± 0.07 (+0.88)
	Padding	86.93 ± 0.55 (+4.20)	65.23 ± 1.24 (+5.23)	58.07 ± 1.10 (+4.17)	09.27 ± 0.15 (+1.64)	06.00 ± 0.17 (+0.83)	05.57 ± 0.15 (+0.87)
	WeightMap	74.90 ± 0.92 (+5.33)	46.80 ± 0.70 (+5.43)	45.77 ± 0.46 (+5.00)	06.17 ± 0.12 (+1.30)	03.53 ± 0.15 (+0.56)	03.60 ± 0.10 (+0.63)
	ModPrompt	93.73 ± 0.25 (+0.60)	68.27 ± 0.40 (+1.10)	60.87 ± 0.50 (+0.77)	26.63 ± 0.38 (+4.93)	18.53 ± 0.32 (+3.50)	17.27 ± 0.46 (+3.14)

Table 2. **Detection performance (APs) for YOLO-World and Grounding DINO on LLVIP-IR and NYU_{v2}-Depth datasets.** Each visual prompt adaptation strategy is compared with the learnable MPDR (results in parentheses are the gain with the MPDR module), which is responsible for updating the new modality embeddings and not changing the original embedding knowledge.

World-S-v2, with the YOLOv8 backbone and CLIP [39] as the language model. For Grounding DINO, we use Swin-Transformer Tiny [31] as the detector backbone and BERT [9] for the language encoder. Our models are evaluated in terms of AP performance using COCO API [29]. The batch size for YOLO-World models was 8, and the models were trained with a single GPU. For Grounding DINO, the batch size was 16, and 2 GPUs were used for training using DDP. Yolo-World and Grounding Dino were trained with the maximum training epoch of 80 and 60, respectively. For offline embedding extraction, we used CLIP-ViT-base-patch32 and Bert-base-uncased for YOLO-World and Grounding DINO, respectively. Additional details are provided in the supp. materials.

(c) Baseline Methods:

- **Zero-shot Model.** The zero-shot model refers to performing the inference with the original pre-trained open-vocabulary detector model, but on our final detection modality. Normally, these models are trained with large RGB datasets, such as Object365 [42] or GoldG [22].

- **Finetuning.** The finetuning strategy refers to adapting a pre-trained model by changing its parameters directly for the final task, but it destroys the pre-training knowledge of the model. The most common strategies for adapting detectors are the *head finetuning*, which adapts only the head of the detector, and *full finetuning*, which adapts all the weights of the detector.

- **Visual Prompt.** For this baseline, we define a family of different visual prompt strategies explored in our work. Specifically, we have *fixed*, *random*, *padding*, and *weight map*. In *fixed*, we add a fixed patch in a fixed position for all training. In *random*, we have a random padding that is sampled per image. In *padding*, we add a padding prompt and combine it with the image. In *weight map*, we consider the learnable weights to be a mask with the dimensions of the input image.

- **Modality Prompt.** Here, we refer to our proposed encoder-decoder visual prompt, in which we learn how to translate the input, guided by the final detection loss, and then we combine its output with the original input.

Method	Variation	LLVIP - IR		
		AP ₅₀	AP ₇₅	AP
Fixed	30	61.60 ± 0.75	39.93 ± 0.52	37.97 ± 0.56
	300	70.30 ± 7.89	45.67 ± 6.97	43.53 ± 5.79
Random	30	60.13 ± 0.29	38.73 ± 0.17	36.87 ± 0.12
	300	56.27 ± 0.46	33.73 ± 0.62	33.13 ± 0.42
Padding	30	79.87 ± 1.00	51.77 ± 0.90	49.30 ± 0.83
	300	39.53 ± 2.36	15.90 ± 1.02	19.07 ± 1.18
ModPrompt	MB	92.80 ± 0.29	70.73 ± 1.02	62.87 ± 0.63
	RES	91.03 ± 0.12	68.40 ± 1.10	61.43 ± 0.58

Method	Variation	NYU _{v2} - Depth		
		AP ₅₀	AP ₇₅	AP
Fixed	30	04.67 ± 0.05	03.07 ± 0.05	02.90 ± 0.00
	300	03.43 ± 0.05	02.00 ± 0.08	02.10 ± 0.00
Random	30	04.23 ± 0.12	02.63 ± 0.05	02.53 ± 0.05
	300	01.53 ± 0.17	00.77 ± 0.12	00.87 ± 0.12
Padding	30	03.97 ± 0.05	02.50 ± 0.00	02.43 ± 0.05
	200	00.37 ± 0.12	00.10 ± 0.08	00.17 ± 0.05
ModPrompt	MB	35.37 ± 0.12	25.20 ± 0.24	23.27 ± 0.17
	RES	37.17 ± 0.57	27.50 ± 0.64	24.93 ± 0.50

Table 3. **Detection performance (APs) for YOLO-World under the two main datasets evaluated: LLVIP-IR and NYU_{v2}-Depth.** We compared the main visual prompt strategies *fixed*, *random*, *padding*, and ModPrompt. The variations consist of the number of prompt pixels ($p_s = 30, 200, \text{ or } 300$) and for ModPrompt, the MobileNet (MB) or ResNet (RES).

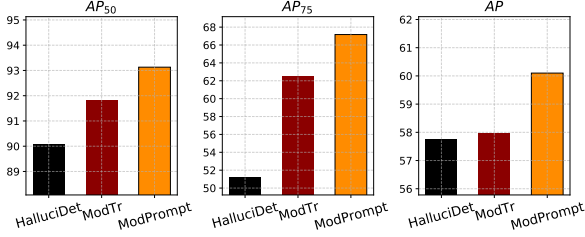


Figure 4. Detection performance on LLVIP dataset of different SOTA Modality Translation OD methods in terms of APs.

4.2. Visual Modality Adaptation

This section explores two main open-vocabulary detectors, YOLO-World and Grounding DINO, for the input space visual prompt adaptation on IR and depth modalities. In the Table 1, we tabulate the results using YOLO-World and Grounding DINO. We compare zero-shot (ZS), full-finetuning (FT) of the visual backbone, visual prompt with fixed patch, random patch, padding patch, weight map (WM) on all pixels, and weight map with scale shift (WM_{v2}), and our ModPrompt. Every visual prompt strategy adapts at the input level, so we keep all other parameters frozen to avoid catastrophic forgetting. In this setup, the YOLO-World with LLVIP-IR dataset, we observe that ModPrompt outperforms the other prompt strategies with AP₅₀ 92.80 and the second best was WM with 82.00, similar trends for the AP₇₅ with ModPrompt reaching 70.73 and WM with 53.90. Additionally, in terms of AP, our method has 62.87 AP, while the second-best approach has 50.90 AP. For Grounding DINO, in LLVIP-IR, our ModPrompt had an 93.13 AP₅₀, 67.17 AP₇₅, and 60.10 AP, and the second best (fixed patch) had an AP₅₀ of 83.87, but close to random and padding prompts. We observe that ModPrompt performs better when objects are well-defined in the image and when objects are not too small, otherwise, like all other input-level pixel strategies it struggles, especially on refined bounding-box localization, which can be seen with AP₇₅ and AP, whereas in AP₅₀ it shows good results. Performance on the FLIR dataset is reported in the supp. material.

Comparison with SOTA Modality Translation OD methods. Our ModPrompt technique is compared with recent works on modality translators for ODs: HallucDet [33] and ModTr [32]. In Fig. 4, we observe that our results are better in all APs for the LLVIP dataset. Additional results for the FLIR-IR dataset are presented in the supplementary material.

Comparison of the number of Train Parameters and Knowledge Preservation. In Table 4 we compare the

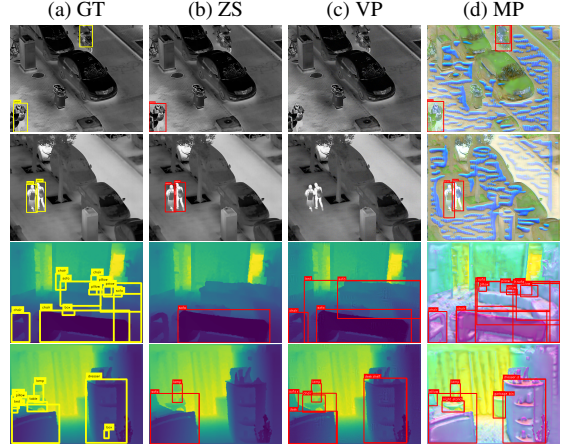


Figure 5. Detections for YOLO-World for the different approaches: First two rows for LLVIP (infrared), and last two rows for NYU_{v2} (depth). Each column corresponds to a different approach: (a) GT (Ground Truth): Shows in yellow the ground-truth bounding boxes for objects. (b) Zero-Shot (ZS): Displays detections (in red) from a zero-shot model. (c) Visual Prompt (VP): Illustrates detections with a visual prompt added to the image. (d) ModPrompt (MP) (Ours): Detections from our proposed model.

number of training parameters and show the catastrophic forgetting in the compared baseline methods. As we can see, ModPrompt preserves the zero-shot performance on the COCO dataset, while the performance drops significantly in HFT and FT. Other visual prompt strategies, such as WM and WM₂, can preserve the zero-shot performance, but our average performance is much better than theirs while requiring fewer trainable parameters.

Method	Params (M)	LLVIP	COCO	Avg.
ZS	0.00	81.00 ± 0.00	51.90 ± 0.00	66.45 (00.00)
HFT	2.31	93.57 ± 0.05	00.66 ± 0.04	47.12 (-19.33)
FT	76.81	97.43 ± 0.05	00.10 ± 0.00	48.77 (-17.68)
WM	3.93	87.47 ± 0.17	51.90 ± 0.00	69.69 (+3.24)
WM _{v2}	7.86	85.00 ± 0.08	51.90 ± 0.00	68.45 (+2.00)
ModPrompt	3.08	95.63 ± 0.04	51.90 ± 0.00	73.77 (+7.32)

Table 4. AP₅₀ of YOLO-World on LLVIP-IR and COCO data. We compare the number of trainable parameters and show the catastrophic forgetting in HFT and FT baselines.

4.3. Ablation Studies

Visual Prompts. We evaluate different variations of the visual prompt adaptation methods introduced in our paper. Specifically, we compare the performance when different input patch sizes are used; for instance, $p_s = 30$ refers to a patch size of 30 pixels. In this study, we tested various patch sizes for each of the visual prompt methods and reported the performance in Table 3. We

evaluate ModPrompt using two different translators with U-Net based backbones- MobileNet (MB) [17] and ResNet (RES) [15]. Empirically, we observe minimal variation in results when increasing the prompt patch size. In some cases, a larger prompt could even degrade the performance of the detector on the new modality, as seen with $p_s = 200$ on NYU_{v2} datasets. A similar trend can be observed with ModPrompt, where sometimes the MB version, with far fewer parameters, can outperform the RES version. This finding suggests that a lighter MB translator can be a viable choice over a heavier RES translator, making it better suited for real-time performance, with minimal degradation in detection metrics.

MPDR Knowledge Preservation mechanism. Table 2 shows that MPDR can be beneficial for improving the majority of the visual prompt strategies while preserving the original knowledge of the detector. For the text adaptation setting, we pre-compute the embeddings for the target classes using our language model, avoiding forward/backpropagation on it. In this setting, we also tried performing adaptations directly on the embeddings without using MPDR parameters, similar to text/embedding prompt tuning; in this case, the performance on the final target modality was close to MDPR. However, the zero-shot knowledge of the model was lost; thus, we opted to report only the MDPR in the main manuscript.

Broader Modality Generalization. To further demonstrate the effectiveness of ModPrompt, we extended our evaluation to include Event-based (PEDRo) [3] and LiDAR (STCcrowd) [8] modalities. As shown in Table 5, ModPrompt achieves strong performance, outperforming HFT and the visual prompt baseline. These results demonstrate that our method generalizes effectively across four distinct modalities (IR, Depth, Event-based, and LiDAR), confirming its robustness beyond the original experimental scope.

Method	Event-based (PEDRo)			LiDAR (STCcrowd)		
	AP ₅₀	AP ₇₅	AP	AP ₅₀	AP ₇₅	AP
ZS	23.9	08.9	11.4	00.0	00.0	00.0
HFT	62.1	27.7	31.4	56.8	21.8	27.0
FFT	94.1	78.7	66.9	88.5	61.2	55.7
WM	49.9	21.3	24.7	34.8	10.0	14.6
ModPrompt	92.9	74.2	63.8	85.7	54.8	51.2

Table 5. **Results on PEDRo (event-based) and STCcrowd (LiDAR) datasets on YOLO-World (YW).**

Study on Computational Cost. Our MPDR module is explicitly designed for inference-friendliness. Since text embeddings are precomputed, no forward pass through the text encoder is required either during inference or training. As shown in Figure 6, the MPDR module reduces the train-

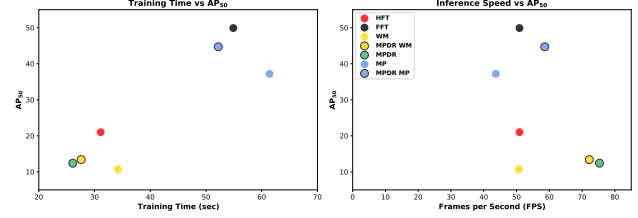


Figure 6. **Comparison of Training Time and Inference Speed vs. Detection Performance on NYU_{v2}.** Left: shows training time per epoch (in seconds) vs. AP₅₀. Right: presents inference time (FPS) vs. AP₅₀. Each point is a method, with HFT, FFT, WM, Modality Prompt (MP), and MP with Decoupled Residuals (MPDR) variants shown in distinct colors. Higher values in both FPS and AP₅₀ indicate better runtime and accuracy trade-offs.

ing time and increases the inference FPS. Although MPDR and MPDR WM require less computation, our MPDR ModPrompt achieves substantially higher detection performance while maintaining reasonable efficiency.

4.4. Qualitative Results

Figure 5 shows the visualization of ModPrompt on YOLO-World compared with zero-shot and the best visual prompt baseline. Thus, we can see that visual prompt-based methods, in general, struggle a bit to provide attached bounding boxes, but they do a good job of adapting. The VP method could not detect people in the LLVIP image well, while ModPrompt did a good job. Additional visualizations are provided in the supplementary material.

5. Conclusion

In this work, we presented ModPrompt, a visual prompt strategy that adapts open-vocabulary object detectors to new modalities (e.g., infrared, depth, event-based, and LiDAR) while preserving zero-shot capabilities. Our encoder-decoder design yields strong performance across challenging benchmarks. We also introduced MPDR, a decoupled text-prompt mechanism that retains language knowledge during adaptation and can match full fine-tuning performance. ModPrompt outperforms prior pixel-level prompts and generalizes across both CNN and transformer backbones, offering a robust and efficient solution for modality-adaptive object detection. By expanding the potential of vision-language models in different applications, ModPrompt offers a compelling solution for future advancements in adaptable object detection frameworks, especially in settings that require full knowledge preservation.

Acknowledgments. This work was supported by Distech Controls Inc., the Natural Sciences and Engineering Research Council of Canada, the Digital Research Alliance of Canada, and MITACS.

References

- [1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 2, 3
- [2] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 384–400, 2018. 2
- [3] Chiara Boretti, Philippe Bich, Fabio Pareschi, Luciano Prono, Riccardo Rovatti, and Gianluca Setti. Pedro: an event-based dataset for person detection in robotics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023. 8
- [4] Nicolas Bustos, Mehrra Mashhadi, Susana K Lai-Yuen, Sudeep Sarkar, and Tapas K Das. A systematic literature review on object detection using near infrared and thermal images. *Neurocomputing*, page 126804, 2023. 2
- [5] Yue Cao, Junchi Bin, Jozsef Hamari, Erik Blasch, and Zheng Liu. Multimodal object detection by channel switching and spatial attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 403–411, 2023. 5
- [6] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5
- [7] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 2, 3, 5
- [8] Peishan Cong, Xinge Zhu, Feng Qiao, Yiming Ren, Xidong Peng, Yuenan Hou, Lan Xu, Ruigang Yang, Dinesh Manocha, and Yuexin Ma. Stcrowd: A multimodal dataset for pedestrian perception in crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19608–19617, 2022. 8
- [9] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 6
- [10] Thomas Dubail, Fidel Alejandro Guerrero Peña, Heitor Rapela Medeiros, Masih Aminbeidokhti, Eric Granger, and Marco Pedersoli. Privacy-preserving person detection using low-resolution infrared cameras. In *European Conference on Computer Vision*, pages 689–702. Springer, 2022. 2
- [11] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust rgb-d object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687. IEEE, 2015. 2
- [12] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 3
- [13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3
- [14] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. 3
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 8
- [16] Christian Herrmann, Miriam Ruf, and Jürgen Beyerer. Cnn-based thermal infrared person detection by domain adaptation. In *Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*, page 1064308. International Society for Optics and Photonics, 2018. 2
- [17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 8
- [18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017. 3
- [19] Eugenio Ivorra, Mario Ortega, Mariano Alcañiz, and Nicolás García-Aracil. Multimodal computer vision framework for human assistive robotics. In *2018 Workshop on Metrology for Industry 4.0 and IoT*, pages 1–5. IEEE, 2018. 2
- [20] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 2, 3
- [21] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8. <https://github.com/ultralytics/ultralytics>, 2023. 2
- [22] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetrm: modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021. 6
- [23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. 3
- [24] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 2
- [25] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu

- Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 3
- [26] Shasha Li, Yongjun Li, Yao Li, Mengjun Li, and Xiaorong Xu. Yolo-firi: Improved yolov5 for infrared image object detection. *IEEE access*, 9:141861–141875, 2021. 2
- [27] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. 2
- [28] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. In *The Eleventh International Conference on Learning Representations*, 2023. 3
- [29] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 6
- [30] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 2, 3
- [31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6
- [32] Heitor Rapela Medeiros, Masih Aminbeidokhti, Fidel Alejandro Guerrero Peña, David Latortue, Eric Granger, and Marco Pedersoli. Modality translation for object detection adaptation without forgetting prior knowledge. In *European Conference on Computer Vision*, pages 51–68. Springer, 2024. 3, 7
- [33] Heitor Rapela Medeiros, Fidel A Guerrero Pena, Masih Aminbeidokhti, Thomas Dubail, Eric Granger, and Marco Pedersoli. Hallucidet: Hallucinating rgb modality for person detection through privileged information. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1444–1453, 2024. 3, 7
- [34] Heitor R Medeiros, David Latortue, Eric Granger, and Marco Pedersoli. Mixed patch visible-infrared modality agnostic object detection. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 9023–9032. IEEE Computer Society, 2025. 2
- [35] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 2
- [36] Mehmet Akif Özkanoglu and Sedat Ozer. Infragan: A gan architecture to transfer visible images to infrared domain. *Pattern Recognition Letters*, 155:69–76, 2022. 2
- [37] Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia*, 24:3859–3881, 2021. 3
- [38] Harry A Pierson and Michael S Gashler. Deep learning in robotics: a review of recent research. *Advanced Robotics*, 31(16):821–835, 2017. 2
- [39] 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. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 6
- [40] Anitha Ramachandran and Arun Kumar Sangaiah. A review on object detection in unmanned aerial vehicle surveillance. *International Journal of Cognitive Computing in Engineering*, 2:215–228, 2021. 2
- [41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4
- [42] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 6
- [43] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 5
- [44] Jack Stilgoe. Machine learning, social learning and the governance of self-driving cars. *Social studies of science*, 48(1): 25–56, 2018. 2
- [45] Qingwang Wang, Yongke Chi, Tao Shen, Jian Song, Zifeng Zhang, and Yan Zhu. Improving rgb-infrared object detection by reducing cross-modality redundancy. *Remote Sensing*, 14(9):2020, 2022. 2
- [46] Xiangyang Xu, Yuncheng Li, Gangshan Wu, and Jiebo Luo. Multi-modal deep feature learning for rgb-d object detection. *Pattern Recognition*, 72:300–313, 2017. 2
- [47] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via modality interaction. *Advances in Neural Information Processing Systems*, 35:1992–2005, 2022. 2
- [48] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10899–10909, 2023. 3, 5
- [49] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, pages 106–122. Springer, 2022. 2
- [50] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International*

- Conference on Image Processing (ICIP)*, pages 276–280. IEEE, 2020. [5](#)
- [51] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022. [2](#), [3](#)
 - [52] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2023. [2](#)
 - [53] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803, 2022. [3](#)
 - [54] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. [2](#), [5](#)
 - [55] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. [2](#), [3](#), [5](#)
 - [56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. [3](#)
 - [57] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023. [1](#)