Mixed Patch Visible-Infrared Modality Agnostic Object Detection

Heitor R. Medeiros* David Latortue*
Eric Granger Marco Pedersoli
Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA)
International Laboratory on Learning Systems (ILLS)
Dept. of Systems Engineering, ETS Montreal, Canada

Abstract

In real-world scenarios, using multiple modalities like visible (RGB) and infrared (IR) can greatly improve the performance of a predictive task such as object detection (OD). Multimodal learning is a common way to leverage these modalities, where multiple modality-specific encoders and a fusion module are used to improve performance. In this paper, we tackle a different way to employ RGB and IR modalities, where only one modality or the other is observed by a single shared vision encoder. This realistic setting requires a lower memory footprint and is more suitable for applications such as autonomous driving and surveillance, which commonly rely on RGB and IR data. However, when learning a single encoder on multiple modalities, one modality can dominate the other, producing uneven recognition results. This work investigates how to efficiently leverage RGB and IR modalities to train a common transformer-based OD vision encoder while countering the effects of modality imbalance. For this, we introduce a novel training technique to Mix Patches (MiPa) from the two modalities, in conjunction with a patch-wise modality agnostic module, for learning a common representation of both modalities. Our experiments show that MiPa can learn a representation to reach competitive results on traditional RGB/IR benchmarks while only requiring a single modality during inference. Our code is available at: `https://github.com/heitorrapela/MiPa`.

1. Introduction

In recent years, the reducing costs in data acquisition and labeling have proportioned the advancements in multimodality. Various fields are increasingly using this form of learning to enhance applications, such as surveillance [1, 6, 28], industrial monitoring [17, 21, 22], smart buildings [10, 13], self-driving cars [27, 29, 35], and robotics [14, 20, 32], due to their powerful ability to operate better in the presence of diverse environmental information [37]. For instance, the combination of visible (RGB) and infrared (IR) has been showing promising results regarding such applications due to the difference in light spectrum sensing by different sensors, which provide not only additional but also complementary information [40].

An unimodal learning (Figure 1a), utilizes data from a single modality, for instance, an object detector trained and used in production with RGB images. In multimodal learning (Figure 1b), the objective is to create a model able to incorporate information from multiple modalities, such as RGB and IR, from different sensors and requires paired modalities for both training and inference. Although this multimodal learning covers a wide range of applications, as aforementioned, we have identified an underserved scenario where one might want an RGB/IR modality agnostic model that is trained on both modalities but is subjected to only either one or another during inference (Figure 1c). One example of that is a surveillance system where a server model is running all the time, and this model can provide detections for different RGB or IR sensors to address the need to make accurate detection in every lighting condition during different pre-defined conditions.

Despite the strong interest and business value in multimodal systems, most publicly available datasets and powerful pre-trained models are built around one modality: RGB. Furthermore, the lack of IR data gives additional motives to build a detector upon an already pre-trained unimodal RGB detector. However, the current methods proposed in research to incorporate dual-modality information into a model require dedicated components associated with each modality, making them incompatible with such RGB detectors. These methods are mainly based on fusion. For instance, these techniques adopt different modalities by distributing the RGB/IR across a four-channel input (three RGB followed by one for IR), in the case of early fusion [39]., or merging both modalities later in the model architectures [4, 43, 45] for mid-stage fusion or ensembling

[*] Equal contribution. Contact: heitor.rapela-medeiros.1@ens.etsmtl.ca

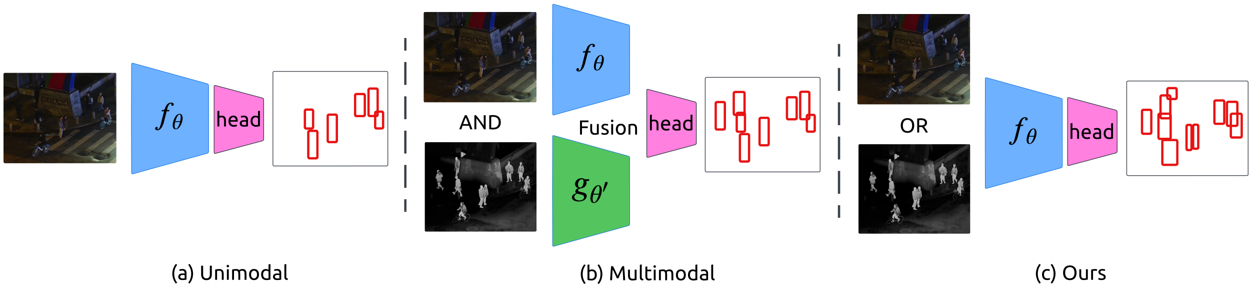

Figure 1. **Differences in inputs for different modality learning.** (a) *Unimodal* learning assumes that only one modality is used for both training and testing. (b) *Multimodal* learning requires multiple modalities and a special architecture to fuse them in order to improve performance. (c) *Ours* assumes that a model should be able to perform well for both modalities by using both for training but only one at a time for testing and with a shared vision encoder.

different unimodal modality detectors [8] for late-stage fusion. This constrains the model to utilize both modalities during inference, which significantly increases inference speed compared to an unimodal architecture.

Typically, one probable phenomenon that can occur during multimodal training is modality imbalance. This happens when the strongest modality is leveraged more than the others, leading to better overall performance while discarding contributions from the others. [9]. In this work, we provide a way to train a single shared vision encoder to be agnostic to its input RGB/IR modality yet still extract its knowledge during training to attain results almost as good on both modalities as if it was trained solely on each during testing/production. The naive solution for this type of task is to train a model with a dataset that blends both modalities.

Recently advances on patch-based transformers, such as ViT [11], and Multi-Modal Masked Autoencoders [2] have steered us towards exploring patch-based architectures to build a powerful and yet simple training technique to create a RGB/IR modality agnostic vision encoder for object detection. Such approaches have been promising for multimodal learning, which allows an efficient combination of different information [2, 18]. Our work investigates how to use RGB and IR modalities efficiently by using a patch-based transformer encoder. Thus, Mi(xed) Pa(tch) does not introduce any inference overhead during the testing phase while exploring an effective way to use the two modalities during the training. To accomplish such a task, we introduce a stochastic complementary patch mixing method, allowing the detector to explore each modality without having to rely on both of them simultaneously. This is possible by effectively sampling the optimal ratio of patches for each modality, which is then mixed using our technique. Subsequently, we enhance the training by suppressing the modality imbalances by proposing a modality-agnostic training technique, making the modalities indistinguishable from each other, a module inspired by Gradient Reversal Layer (GRL) [16] but with a novel design for patch based architectures. This approach is designed to allow low-cost inference in production while removing all requirements to know beforehand which modality the detector is going to be used with. Hence, in applications that run a detector all day, we can know beforehand that any of the modalities, RGB or IR, whenever they are being used, are going to perform optimally for the same shared vision encoder.

Our work provides empirical results alongside a theoretical explanation based on information theory describing the benefits of using MiPa with transformer-based backbones. Additionally, we study the ability of our MiPa to also be used as a regularization method for the more robust modality to boost the overall performance of the detector and we show that we can achieve competitive results on two traditional RGB/IR benchmarks: LLVIP and FLIR.

Our main contributions can be summarized as follows:

(1) We introduce MiPa, a novel mix patches RGB/IR modality agnostic training method for transformer-based object detectors, which learns how effectively sample the RGB and IR patches for best compressing the information of both modalities in a single encoder, without additional inference overhead.

(2) We propose a novel patch-wise modality agnostic module, which is inspired by the gradient reversal layer (GRL) for modality adaptation and is responsible for making the RGB/IR modalities invariant by the detector.

(3) We empirically demonstrate that the proposed method can also be used to improve the overall performance of detection when utilized as regularization for the strongest modality and achieve competitive results when compared with multimodal fusion methods, with less information during inference. Furthermore, MiPa can simply be applied to different transformer-based detectors, such as DINO [46] and Deformable DETR [48].

2. Related Work

2.1. Patch-Based Vision Encoding

With the integration of Transformers in the vision field, researchers have started to deconstruct images into patches to allow the modeling of long-range relationships between patches [11]. This powerful approach yielded great results and quickly became the norm amongst the top-performing models, ranking well on popular benchmarks such as ImageNet-1k [34]. Multiple variants of the vision transformer have been proposed in recent years, for instance, ViT [11], DeiT [38], Swin [25], and VOLO [42]. Alongside the new way of utilizing input images came a novel pretraining method for vision encoding: Masked Autoencoders [19] (MAE). Indeed, this technique, which is simple to understand and easy to implement, consists of using a classifier as an encoder in an autoencoder architecture to generate images by only using a small fraction of the patches as input. This unsupervised method has proven to be very useful in terms of improving results for downstream tasks. Furthermore, a similar idea has also been influential in the world of multi-modality models by building a multimodal MAE with one encoder and multiple decoders to reconstruct all the different modalities [2]. Recently, advances towards using Swin Transformer as a backbone of DINO [46], an object detector descendant of the DETR [5], were responsible for reaching competitive results in detection benchmarks, such as in COCO dataset [24].

2.2. Multimodal Visible-Infrared Object Detectors

Regarding object detection, the primary methods of exploiting pairs of modalities, even when unaligned, are multimodal techniques; mainly fusion [3]. Fusion is a technique where the advantage of multiple modalities is taken to better optimize one training objective by combining them to develop a multimodal representation [30]. Fusion can be achieved at different stages, i.e., *early-stage fusion*, which concatenates the modalities across the channels, *mid-stage fusion*, where modalities are processed through dedicated decoders then merged e.g., Channel Switching and Spatial Attention (CSSA) [4], Halfway Fusion [43], RS-Det [47], CrossFormer [23] or Guided Attentive Feature Fusion (GAFF) [45], and finally *late-stage fusion*, where typically modalities are processed independently through different models and combined at the end using ensembling [8], e.g. ProbEn [8]. The limitations of multimodal learning are that they require a custom architecture to handle each modality and are constrained to use both modalities during inference. A cross-modal with shared encoder vision models, however, are not affected by these limitations as the different modalities are only used during training and share the same encoder. This type of architecture unlocks the ability for detectors to have a higher degree of freedom for inference without compromising real-time applications.

2.3. Modality Imbalance

A potential obstacle to an RGB/IR modality-agnostic network is the phenomenon of modality imbalance. Given a dataset with multi-modal inputs, modality imbalance occurs when a model becomes more biased towards the contribution of one modality [9] than the others. To counter that, some methods have been proposed for classification, for instance, gradient modulation [31], Gradient-Blending [41], and Knowledge Distillation from the well-trained unimodal model [12]. In gradient modulation, Peng et al. proposed a mechanism to control the adaptive optimization of each modality by monitoring their contributions to the learning objective. In gradient blending, Wang et al. identified that multi-modal learning can overfit due to the increased capacity of the networks and proposed a mechanism to blend the gradients effectively [41]. Du et al. [12] show that training multi-modal models on joint training can suffer from learning inferior representations for each modality because of the imbalance of the modalities and the implicit bias of the common objectives in the fusion strategy. An effective approach to help on the modality imbalance in a shared encoder consists of using a Gradient Reversal Layer (GRL) [15], which was introduced for domain adaptation to reduce a network's reliance on a specific domain. GRL was exhaustively applied in object detection to create a shared domain; for instance, in the work of Chen et al. [7], the GRL is used to adapt Faster R-CNN to distribution shifts in illumination or object appearance. The core idea of GRL involves training a classifier to identify the class of a data example during training. During backpropagation, the gradients are reversed to train the network to deceive the classifier.

In this work, we adapt this technique to address modality imbalance learning. Unlike typical cases where data belongs to a single domain/modality, a single training example of MiPa consists of a mosaic of the two modalities: RGB and IR. Therefore, our classifier is trained to predict a modality map instead. In our work, we tackle the imbalance with an adjustable balancing sampling, which learns how to effectively sample the RGB and IR patches during training, and a patch-based GRL module responsible for encoding in the same vision encoder the information of both modalities while improving detection performance.

3. Proposed Method

While the naive way to create a multimodal vision encoder for an OD is to blend both modalities during training, we empirically show, in Section 4, that this approach leads to an imbalanced performance across modalities. In this section, we present our proposed solution.

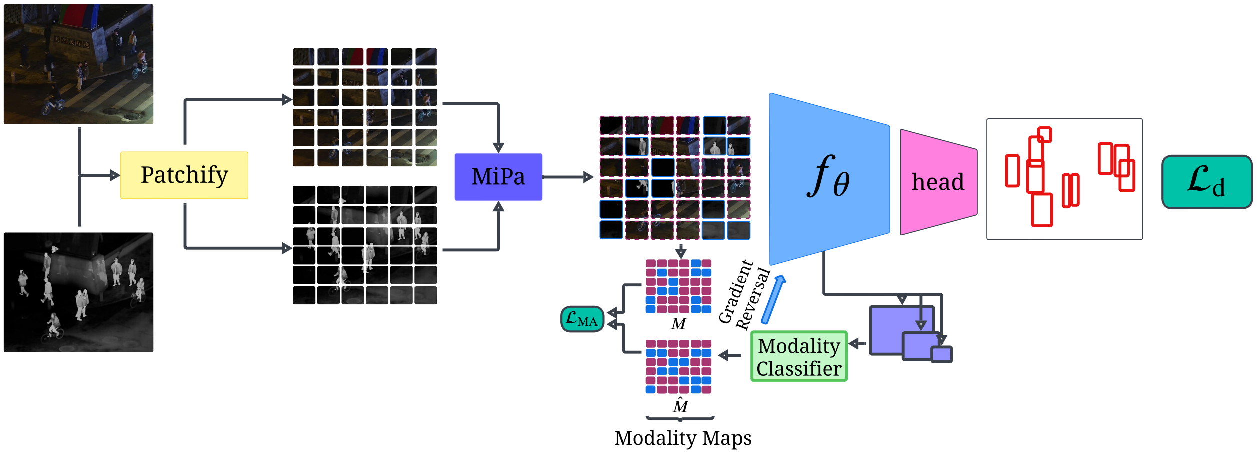

Figure 2. **Mixed Patches (MiPa) with Modality Agnostic (MA) module.** In yellow is the patchify function. In purple is the MiPa module, followed by the feature extractor (encoder). In green is the modality classifier, and in pink is the detection head.

3.1. Preliminary definitions

Let us consider a set of training samples $\mathcal{D} = \{(x_i, B_i)\}$ where $x_i \in \mathbb{R}^{W\times H\times C}$ is the image i with spatial resolution $W \times H$ and C channels. Here, a set of bounding boxes is represented by $B_i = \{b_0, b_1, ..., b_N\}$ with $b = (c_x, c_y, w, h)$ being c_x and c_y coordinates of the center of the bounding box with size $w \times h$. During the training process of a neural network-based detector, we aim to learn a parameterized function $f_\theta : \mathbb{R}^{W\times H\times C} \to \mathcal{B}$, being $\mathcal{B}$ the family of sets B_i and θ the parameters vector. For such, the optimization is guided by a loss function, which is a combination of a regression $\mathcal{L}_r$ and a classification $\mathcal{L}_c$ term, i.e., l_2 loss and binary cross-entropy, respectively. The following Equation (1) defines a general loss function ($\mathcal{L}_d$) for object detection:

$$\mathcal{L}_d(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(x,B)\in\mathcal{D}} \mathcal{L}_c(f_\theta(x), B) + \lambda\mathcal{L}_\text{r}(f_\theta(x), B). \quad (1)$$

3.2. Mixed Patches (MiPa)

The MiPa training method is a training technique that leverages the patch input channel from transformer-based feature extractors to build a powerful common representation between RGB/IR modalities for the unique vision encoder, which can be used in different transformer-based detectors. In short, it consists of a single encoder that receives sampling complementary patches from each modality and rearranges the input into a sort of mosaic image as shown in Figure 2. Such a mechanism forces the model to see both modalities for each inference without being forced to have parameters specialized on a specific one. Depending on how the nature of the patches are sampled, the technique can act as a way to gather the union of information between both modalities or as a regularization for the strongest modality (the easiest modality that tends to drive the learning process). Throughout this paper, we will reference the sampling ratio of the patches as ρ. There are several ways to pick the sampling ratio ρ; the naive way of selecting ρ is to use a fixed ratio during the training of 50%. Then, we can randomly generate a ρ value for each inference. If we have an intuition of which modality needs to be sampled more, we can manually move ρ during the training with a certain curriculum. Finally, we can let the model learn the optimal ratio by itself. In this work, we have explored all these variations to see which one is the most suitable for MiPa.

Theoretical explanation behind the MiPa approach. Here, we detail our theoretical understanding of the MiPa method. We refer to Table 1 for all definitions. The variable $\mathcal{X}$ can be thought of as a scene where you would see individuals walking in the street, for instance, and the functions f and g are camera lenses capturing the information of the scene via IR and RGB, respectively. The goal of MiPa ($\mathcal{M}$) is to enhance learning efficiency by merging information from both modalities, eliminating redundancy, and filtering out noise, all in a **single inference**. *Thus, say we have:*

$$f(\mathcal{X}) = P + \eta_f; g(\mathcal{X}) = Q + \eta_g, \quad (2)$$

where Equation (2) represents the visualization of the scene, which is composed of the information captured by the sensor (P or Q), P is information captured from the sensor of one modality and Q for the other modality, and some noise (η). Then the application of MiPa ($\mathcal{M}$) can be summarized as the following Equation (3):

$$\mathcal{M}(f(\mathcal{X}), g(\mathcal{X})) = \begin{cases} f(\mathcal{X}_i), & i \in m \\ g(\mathcal{X}_i), & i \in l, \end{cases} \quad (3)$$

where $f(\mathcal{X}_i)$ represents the mapping of the patch $\mathcal{X}$ with id i using f (IR lens) and $g(\mathcal{X}_i)$ using RGB lens. Then, the combination of the individual patches of each modality is given by Equation (4):

$$\mathcal{M} = (P_0+P_1+Q_2+...+Q_{n-1}+P_n)+(m{\cdot}\eta_f+l{\cdot}\eta_g). \quad (4)$$

As RGB and IR patches do not encode the same information in the same patch visualization $\mathcal{X}_i$, the additional information of one modality improves, for instance, IR on the night, the other one. Also, this variation in the sense of information for both modalities is responsible for regularizing the training when the patches are mixed. The following Equation (5) represents the approximation of the real mutual information I by $\mathcal{M}$ using Equation (4) and approximating the noise from the scene to be similar for both sensors:

$$\mathcal{M} = I_a + \eta. \quad (5)$$

This approximation means that the encoded information on MiPa represents the total scene composed by both sensors, which are compressed on the vision encoder while removing the redundancy information and noise by the training process.

3.3. Patch-Wise Modality Agnostic Training

As previously mentioned, modality imbalances can potentially cause the model to rely mostly on one modality. Since the objective of this work is to preserve the original architecture of the model for inference, we opted for an approach where the backbone would be responsible for mediating the modalities. To do so, we designed an adaptation of the GRL technique [15] called ***patch-wise modality agnostic*** (MA) module. The key idea is to prevent the detector from relying too much on the strongest modality, the easiest modality driven by the learning process, by making the features from each modality indistinguishable, therefore sharing the same encoding. Considering that the input has a different modality for each patch, a modality that we pick during the patch mixing process, we build what we call a *modality map*, denoted as M, that specifies which modality each patch belongs to for each inference during training. Then, we use a modality classifier to predict the modality map of the features coming from the backbone. Finally, we compute the loss between the target and outputted modality maps and back-propagate the opposite gradients to the backbone encoder. To reduce the noise coming from the classifier at the beginning of the training, we slowly increase the weight (λ) of the gradients propagated to the backbone as the training goes on. We use the Binary Cross-Entropy (BCE) to compute the loss between the predicted and target modality maps, as described by the following Equation (6):

Table 1. Definition of the random variables and information measures used to explain MiPa.

General	
Input scene in patches	$\mathcal{X}_n$
Number of patches	$n \in \mathbb{N}$
Patch id	$i \in \mathbb{N}$
Random variables (RVs)	
Patch ratio	$\rho \sim U(0,1)$
Patch channel f	$m \sim \binom{n \cdot \rho}{p}$, p = $\frac{1}{2}$
Patch channel g	$l \sim n$ - m
Functions	
MiPa	$\mathcal{M}$
Self-Attention	SA
Modality channels	f, g
Information measures	
Entropy of V	$\mathcal{H}(V) \coloneqq \mathbb{E}_{p_V}\left[-\log p_V(V)\right]$
Information of X	Q, P where Q =$\mathcal{H}(q)$, P = $\mathcal{H}(p)$
Noise modality channels	η
Mutual information between P and Q	$I(Q,P)$ = Q + P - $Q \cap P$
Approximation of mutual information between P and Q	$I_a \approx I$

$$\mathcal{L}_{\text{MA}} = \frac{1}{n}\sum_{i=1}^{n} -M\log(\hat{M}) - (1-M)\log(1-\hat{M}), \quad (6)$$

where M is the modality map generated from ρ. The aforementioned approach for the full training pipeline can be seen in Figure 2. We use the following Equation (7) to increment the factor λ.

$$\lambda = \frac{2}{1+exp(-\gamma s)} - 1, \quad (7)$$

where s is the speed to which λ increases based on training epoch and γ is a hyperparameter to adjust this speed. The modality classifier can be used at any stage of the backbone; we have found empirically that using it on the features from the stage 1 works well. Finally, MiPa loss ($\mathcal{L}_{\text{MiPa}}$) can be defined as the following Equation (8):

$$\mathcal{L}_{\text{MiPa}} = \mathcal{L}_d + \lambda\mathcal{L}_{\text{MA}}. \quad (8)$$

4. Results and Discussion

4.1. Experimental Methodology

(a) Datasets: During our experiments, we explored two different RGB/IR benchmarking datasets: LLVIP and FLIR. **LLVIP:** The LLVIP dataset is a surveillance dataset composed of $12,025$ RGB/IR pairs of images for training and $3,463$ pairs for testing. The original resolution is 1280 by 1024 pixels but was resized to 640 by 512 to accelerate the training. The sole annotated class of this dataset is pedestrians. **FLIR ALIGNED:** For the FLIR dataset, we used the sanitized and aligned paired sets provided by Zhang et al. [44], which has $4,129$ aligned pairs for training and $1,013$ pairs for testing. The FLIR images are taken from the perspective of a camera in front of a car, and the resolution is 640 by 512. It contains annotations of bicycles, dogs, cars, and people. It has been found that for the case of FLIR, the "dog" objects are inadequate for training [4], but since our objective is to evaluate if our method can make a detector modality agnostic and not beat any prior benchmark, we have decided to keep it during our evaluations.

(b) Implementation Details: All detectors were trained on an A100 NVIDIA GPU and were implemented using PyTorch. We use AdamW [26] as an optimizer with a learning rate of $1e^{-4}$, a batch size of 6, and for a total of 12 epochs for the case of the DINO [46] OD. For Swin, we start with the pre-trained weights from ImageNet [34]. The models are evaluated in terms of performance AP_{50}, and we additionally reported the AP_{75} and AP in the supplementary material. The evaluation is also performed in terms of RGB performance, IR performance, and our target metric, the average of both, because our setup requires a model that is equally good on both modalities during test time. In this work, we replicate the 1-channel IR to have 3-channel input for further use with 3-channel RGB data.

(c) Baseline Methods: In the course of this work, we considered different baselines to compare to our proposed method (MiPa). Firstly, we measure the performance of the detector trained on one modality, unimodal setup, to gain a reference of the expected detection coming from each modality. Secondly, we evaluate the naive solution of simply using a dataset comprised of both modalities during training (multimodal setting), which we call *Both*. To account for the modality imbalances and further increase the fairness of our comparisons, we balanced the datasets with 25%, 50%, and 75% of one modality and the rest of the other. All models were evaluated separately on RGB and IR. Additionally, the mean of the modalities, which represents how well the model is balanced for the two desired modalities, is calculated.

Table 2. Comparison of different ratio ρ sampling methods on LLVIP. Using DINO with Swin backbone.

Model	Dataset: LLVIP (AP_{50} ↑)		
	RGB	IR	Average
Fixed [ρ=0.25]	78.9	**98.2**	88.55
Fixed [ρ=0.50]	73.0	97.6	85.30
Fixed [ρ=0.75]	77.4	97.5	87.45
Curriculum (ρ=0.25 / 4 epochs)	76.6	97.8	87.20
Curriculum (ρ=0.25 / 8 epochs)	80.1	97.8	88.95
Variable	**88.5**	97.5	**93.00**

4.2. Towards the optimal ρ

Since the way of selecting the ideal ρ was unclear, we designed different experimental settings to study the influence of ρ on learning the best way to balance the amount of RGB/IR information during the training. Let us start with a few definitions:

- Fixed ρ. In this setting, we selected a fixed proportion of RGB/IR samples, such as 0%, 25%, 50%, 75% and 100%, in which 0% correspond to no IR images in the training batch, and 100% correspond to only IR in the batch.

- Curriculum ρ. For this strategy, we analyzed which modalities were easy to learn; in this case, it was IR. Then, during the initial epochs over the training, the model focuses on the easier-to-learn modality (IR modality tends to drive the learning process when a balanced jointly dataset is given), providing between 0% to 25% of ratio for IR, which means that the model for the initial epochs is going to see more RGB data, which is harder. Then, over the rest of the training epochs, it samples from the uniform distribution such as variable ρ.

- Variable ρ. In the variable ρ, the ratio of mixed patches per batch is drawn from a uniform distribution. For each batch, a different ρ is redrawn.

We tested all the different configurations of ρ on LLVIP (see Table 2). For this experiment, we have made two findings. First, using an I_a following a uniform distribution gives us a better approximation of the range of information from $IR \cup RGB$ as the results from the variable give us a better balance between both modalities. Second, using less of the weaker modality (hard to learn) strengthens the learning of the strongest one (easier to learn modality), as it can be seen in Sec. 4.4 (Table 5), that we were actually able to beat the state-of-the-art by sampling 25% of RGB images and 75% of IR.

4.3. Patch-wise Modality Agnostic Training

The subsequent ablation shows the efficacy of the patch-wise modality agnostic method towards obtaining a single model capable of dealing with both modalities while keeping the performance stable, see Table 3. Additionally, we studied the sensibility of the model performances influenced by different γ hyperparameters (see Table 4), seen in Equation (7), which tunes the speed that the λ factor increases at each step the weight of gradients propagated to the encoder. We empirically demonstrate that the optimal γ varies between datasets and detectors due to the number of epochs required for each one, whereas if the model requires more training epochs, the γ should be higher. Additionally, MiPa was designed for computational efficiency during test time, so it does not increase the computational cost when the model is deployed in real-world scenarios.

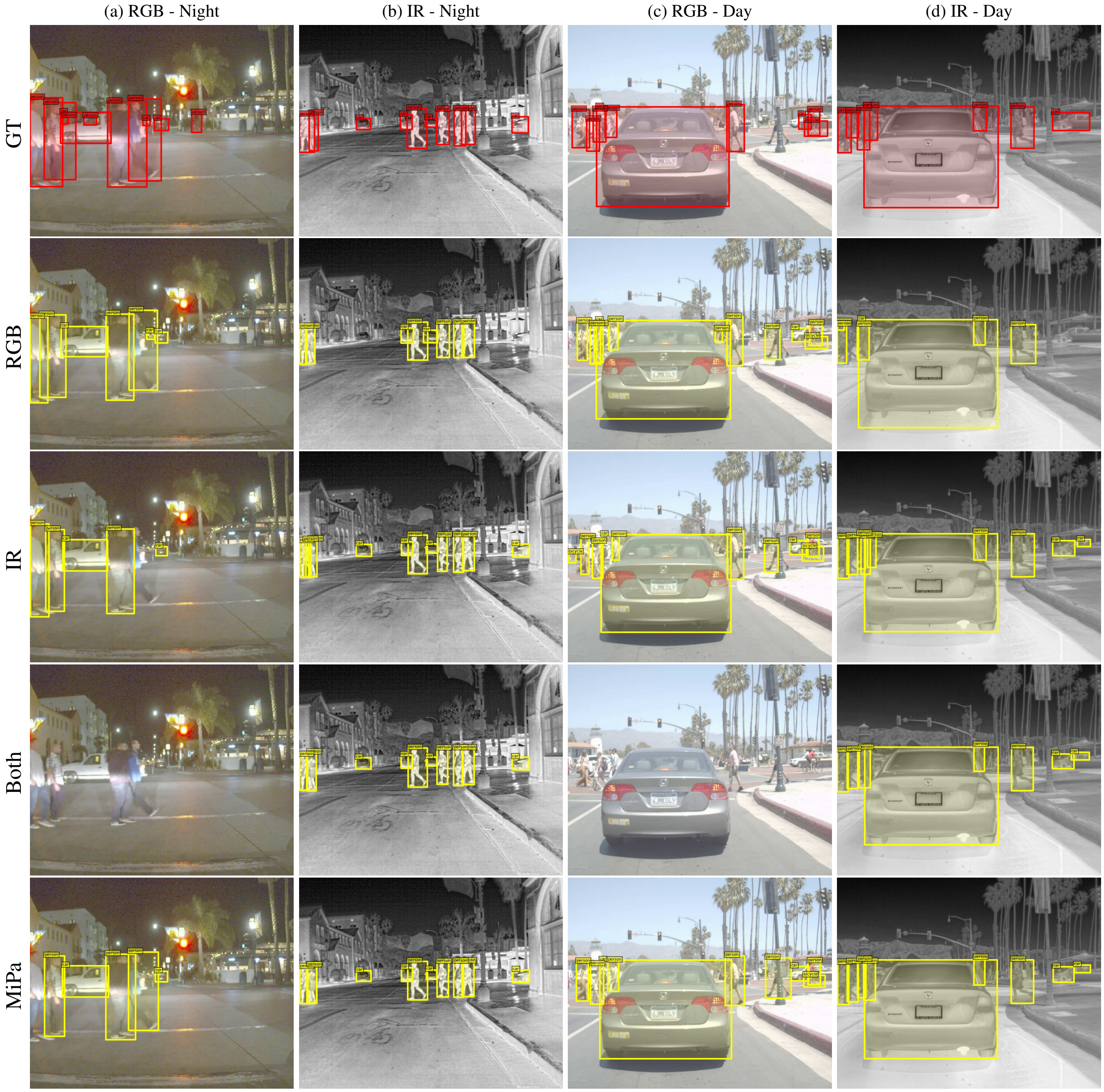

Figure 3. **Detection over different methods for two different daytimes:** Night and Day and two different modalities: RGB and IR. Detectors trained on *RGB* work better in the daytime. Detectors trained on *IR* work better at nighttime. Detectors trained on *Both* modalities in a naive way cannot work only on the dominant modality. Our *MiPa* manages to work well in all conditions.

4.4. Comparison with RGB/IR Competitors

In this section, we compare our approach in terms of detection performance with other strong methods in the literature that use RGB/IR modalities. Table 5 shows that MiPa is a competitive method under RGB/IR benchmarks. For instance, on FLIR, MiPa has 81.3 AP_{50}, while CSSA [4] has 79.2, ProbEn [8] has 75.5, GAFF [45] 74.6 and Halfway Fusion [44] 71.5, RSDet [47] 81.1 and CrossFormer [23] 79.3. Furthermore, we report competitive results on LLVIP, which can be seen in the table as the people detection performance over different methods inclusively; for competitors, both modalities are used during training and inference, which is not our case (as we just use the IR modality for in-

Table 3. Comparison of detection performance over different baselines and MiPa for different models on Swin backbone for DINO and Deformable DETR. The evaluation is done for RGB, IR, and the average of the modalities.

Detector	Model	Dataset: LLVIP (AP$_{50}$ ↑)		
		RGB	IR	Average
DINO	RGB	90.87 ± 0.84	94.23 ± 0.57	92.55
	IR	66.87 ± 0.90	96.87 ± 0.12	81.87
	Both [$\rho = 0.25$]	79.73 ± 1.03	97.40 ± 0.22	88.57
	Both [$\rho = 0.50$]	82.40 ± 1.50	96.50 ± 0.29	89.45
	Both [$\rho = 0.75$]	81.23 ± 2.89	97.07 ± 0.25	89.15
	MiPa (Ours)	88.70 ± 0.45	96.97 ± 0.26	**92.83**
	MiPa + MA (Ours)	89.10 ± 0.28	96.83 ± 0.09	**92.90**
Def.DETR	RGB	80.00 ± 1.50	90.03 ± 00.87	85.02
	IR	56.10 ± 2.50	94.20 ± 00.08	75.15
	Both [$\rho = 0.25$]	51.20 ± 3.47	83.73 ± 16.57	67.47
	Both [$\rho = 0.50$]	53.57 ± 4.17	83.87 ± 16.17	68.72
	Both [$\rho = 0.75$]	53.53 ± 4.55	82.33 ± 18.48	67.93
	MiPa (Ours)	78.60 ± 0.42	95.20 ± 0.16	**86.90**
	MiPa + MA (Ours)	79.02 ± 0.21	95.36 ± 0.25	**87.19**
Detector	**Model**	**Dataset: FLIR (AP$_{50}$ ↑)**		
		RGB	**IR**	**Average**
DINO	RGB	66.07 ± 0.98	56.60 ± 0.80	61.33
	IR	56.47 ± 0.79	70.40 ± 0.38	63.43
	Both [$\rho = 0.25$]	56.53 ± 0.76	67.57 ± 1.73	62.05
	Both [$\rho = 0.50$]	60.50 ± 0.66	68.93 ± 0.60	64.72
	Both [$\rho = 0.75$]	58.53 ± 0.92	70.43 ± 0.65	64.48
	MiPa (Ours)	63.53 ± 1.94	69.50 ± 1.84	**66.52**
	MiPa + MA (Ours)	64.80 ± 2.30	70.43 ± 0.53	**67.62**
Def.DETR	RGB	49.33 ± 1.39	43.77 ± 00.56	46.55
	IR	39.17 ± 1.48	59.20 ± 00.29	49.18
	Both [$\rho = 0.25$]	35.73 ± 4.95	43.00 ± 13.54	39.37
	Both [$\rho = 0.50$]	33.93 ± 5.15	43.33 ± 14.14	38.63
	Both [$\rho = 0.75$]	32.90 ± 3.54	44.13 ± 14.85	38.52
	MiPa (Ours)	48.00 ± 0.57	54.97 ± 00.90	**51.48**
	MiPa + MA (Ours)	48.27 ± 1.76	55.80 ± 00.22	**52.03**

ference, in Table 5). For example, in LLVIP, MiPa reached 98.8 AP$_{50}$, and the second best was CFT with 97.5.

5. Conclusion

In this work, we have introduced a novel training method leveraging a patch-based strategy using a single vision encoder for OD to consolidate the mutual information between different modalities. This method, named MiPa, has enabled two different object detectors, DINO [46] and Deformable DETR [48], to achieve *modality invariance* on LLVIP and FLIR datasets without having to make any specific changes for each modality, for example, additional encoding parameters for each modality, to their architecture or increase the testing inference time. Additionally, our method outperformed competitors on both datasets. Furthermore, we provide a definition from information theory regarding the knowledge captured by the MiPa method.

Table 4. MiPa ablation on γ and comparison with different baselines for DINO Swin. The evaluation is done for RGB, IR, and the average of the modalities in terms of AP$_{50}$ performance.

Modality	Dataset: LLVIP (AP$_{50}$ ↑)		
	RGB	IR	Average
RGB	90.87 ± 0.84	94.23 ± 0.57	92.55
IR	66.87 ± 0.90	96.87 ± 0.12	81.87
Both [$\rho = 0.25$]	79.73 ± 1.03	97.40 ± 0.22	88.57
Both [$\rho = 0.50$]	82.40 ± 1.50	96.50 ± 0.29	89.45
Both [$\rho = 0.75$]	81.23 ± 2.89	97.07 ± 0.25	89.15
MiPa	88.70 ± 0.45	96.97 ± 0.26	92.83
MiPa [$\gamma = 0.05$]	89.20 ± 0.43	96.57 ± 0.39	92.88
MiPa [$\gamma = 0.10$]	89.43 ± 0.25	96.57 ± 0.31	**93.00**
MiPa [$\gamma = 0.15$]	89.10 ± 0.28	96.83 ± 0.09	92.97
Modality	**Dataset: FLIR (AP$_{50}$ ↑)**		
	RGB	**IR**	**Average**
RGB	66.07 ± 0.98	56.60 ± 0.80	61.33
IR	56.47 ± 0.79	70.40 ± 0.38	63.43
Both [$\rho = 0.25$]	56.53 ± 0.76	67.57 ± 1.73	62.05
Both [$\rho = 0.50$]	60.50 ± 0.66	68.93 ± 0.60	64.72
Both [$\rho = 0.75$]	58.53 ± 0.92	70.43 ± 0.65	64.48
MiPa	63.53 ± 1.94	69.50 ± 1.84	66.52
MiPa [$\gamma = 0.05$]	64.80 ± 2.30	70.43 ± 0.53	**67.62**
MiPa [$\gamma = 0.10$]	64.03 ± 2.11	69.63 ± 1.45	66.83
MiPa [$\gamma = 0.15$]	64.27 ± 0.47	69.93 ± 1.02	67.10

Table 5. Comparison with different multimodal works on RGB/IR benchmarks.

Method	Dataset					
	FLIR			LLVIP		
	AP$_{50}$	AP$_{75}$	AP	AP$_{50}$	AP$_{75}$	AP
Halfway F. [44]	71.5	31.1	35.8	91.4	60.1	55.1
GAFF [45]	74.6	31.3	37.4	94.0	60.2	55.8
ProbEn [8]	75.5	31.8	37.9	93.4	50.2	51.5
CSSA [4]	79.2	37.4	41.3	94.3	66.6	59.2
CFT [33]	78.7	35.5	40.2	97.5	72.9	63.6
DIVFusion [36]	-	-	-	89.8	-	52.0
RSDet [47]	81.1	-	41.4	95.8	-	61.3
CrossFormer [23]	79.3	38.5	42.1	97.4	75.4	65.1
MiPa (Ours)	**81.3**	**41.8**	**44.8**	**98.2**	**78.1**	**66.5**

Acknowledgments

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

References

[1] Mahdi Alehdaghi, Arthur Josi, Rafael MO Cruz, and Eric Granger. Visible-infrared person re-identification using privileged intermediate information. In *European Conference on Computer Vision*, pages 720–737. Springer, 2022. 1

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders, 2022. 2, 3

[3] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38:2939 – 2970, 2021. 3

[4] Yue Cao, Junchi Bin, Jozsef Hamari, Erik Blasch, and Zheng Liu. Multimodal object detection by channel switching and spatial attention. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 403–411, 2023. 1, 3, 6, 7, 8

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[6] Jianguo Chen, Kenli Li, Qingying Deng, Keqin Li, and S Yu Philip. Distributed deep learning model for intelligent video surveillance systems with edge computing. *IEEE Transactions on Industrial Informatics*, 2019. 1

[7] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 3

[8] Yi-Ting Chen, Jinghao Shi, Zelin Ye, Christoph Mertz, Deva Ramanan, and Shu Kong. Multimodal object detection via probabilistic ensembling, 2022. 2, 3, 7, 8

[9] Arindam Das, Sudip Das, Ganesh Sistu, Jonathan Horgan, Ujjwal Bhattacharya, Edward Jones, Martin Glavin, and Ciarán Eising. Revisiting modality imbalance in multimodal pedestrian detection, 2023. 2, 3

[10] Thisun Dayarathna, Thamidu Muthukumarana, Yasiru Rathnayaka, Simon Denman, Chathura de Silva, Akila Pemasiri, and David Ahmedt-Aristizabal. Privacy-preserving in-bed pose monitoring: A fusion and reconstruction study. *Expert Systems with Applications*, 213:119139, 2023. 1

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 2, 3

[12] Chenzhuang Du, Tingle Li, Yichen Liu, Zixin Wen, Tianyu Hua, Yue Wang, and Hang Zhao. Improving multi-modal learning with uni-modal teachers, 2021. 3

[13] 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. 1

[14] 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. 1

[15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation, 2015. 3, 5

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 2

[17] Cinmayii A Garillos-Manliguez and John Y Chiang. Multimodal deep learning and visible-light and hyperspectral imaging for fruit maturity estimation. *Sensors*, 21(4):1288, 2021. 1

[18] Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurmans, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022. 2

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3

[20] Eugenio Ivorra, Mario Ortega, Mariano Alcañiz, and Nicolás Garcia-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. 1

[21] Jyoti Kini, Sarah Fleischer, Ishan Dave, and Mubarak Shah. Egocentric rgb+ depth action recognition in industry-like settings. *arXiv preprint arXiv:2309.13962*, 2023. 1

[22] Xiangyin Kong and Zhiqiang Ge. Deep learning of latent variable models for industrial process monitoring. *IEEE Transactions on Industrial Informatics*, 18(10):6778–6788, 2021. 1

[23] Seungik Lee, Jaehyeong Park, and Jinsun Park. Crossformer: Cross-guided attention for multi-modal object detection. *Pattern Recognition Letters*, 179:144–150, 2024. 3, 7, 8

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 3

[25] 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, 2021. 3

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[27] Heitor Rapela Medeiros, Masih Aminbeidokhti, Fidel Guerrero Pena, David Latortue, Eric Granger, and Marco Pedersoli. Modality translation for object detection adaptation without forgetting prior knowledge. *arXiv preprint arXiv:2404.01492*, 2024. 1

[28] 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. 1

[29] 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. 1

[30] Maciej Pawłowski, Anna Wróblewska, and Sylwia Sysko-Romańczuk. Effective techniques for multimodal data fusion: A comparative analysis. *Sensors*, 23(5), 2023. 3

[31] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8228–8237, 2022. 3

[32] Harry A Pierson and Michael S Gashler. Deep learning in robotics: a review of recent research. *Advanced Robotics*, 31(16):821–835, 2017. 1

[33] Fang Qingyun, Han Dapeng, and Wang Zhaokui. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*, 2021. 8

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. 3, 6

[35] Jack Stilgoe. Machine learning, social learning and the governance of self-driving cars. *Social studies of science*, 48(1):25–56, 2018. 1

[36] Linfeng Tang, Xinyu Xiang, Hao Zhang, Meiqi Gong, and Jiayi Ma. Divfusion: Darkness-free infrared and visible image fusion. *Information Fusion*, 91:477–493, 2023. 8

[37] Qin Tang, Jing Liang, and Fangqi Zhu. A comparative review on multi-modal sensors fusion based on deep learning. *Signal Processing*, page 109165, 2023. 1

[38] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *CoRR*, abs/2012.12877, 2020. 3

[39] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke. Multispectral pedestrian detection using deep fusion convolutional neural networks. 04 2016. 1

[40] 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. 1

[41] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12692–12702, 2020. 3

[42] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. VOLO: vision outlooker for visual recognition. *CoRR*, abs/2106.13112, 2021. 3

[43] Heng Zhang, Elisa Fromont, Sébastien Lefevre, 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, 2020. 1, 3

[44] 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, 7, 8

[45] Heng ZHANG, Elisa FROMONT, Sébastien LEFEVRE, and Bruno AVIGNON. Guided attentive feature fusion for multispectral pedestrian detection. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 72–80, 2021. 1, 3, 7, 8

[46] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 2, 3, 6, 8

[47] Tianyi Zhao, Maoxun Yuan, and Xingxing Wei. Removal and selection: Improving rgb-infrared object detection via coarse-to-fine fusion. *arXiv preprint arXiv:2401.10731*, 2024. 3, 7, 8

[48] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2, 8