Multimodal Object Detection via Probabilistic Ensembling

Yi-Ting Chen[*1], Jinghao Shi[*2], Zelin Ye[*2], Christoph Mertz[2],
Deva Ramanan[†2,3], Shu Kong[†2,4]

[1] University of Maryland, College Park
[2] Carnegie Mellon University
[3] Argo AI
[4] Texas A&M University

ytchen@umd.edu, {jinghaos, zeliny, cmertz}@andrew.cmu.edu,
deva@cs.cmu.edu, shu@tamu.edu
open-source code in Github

Abstract. Object detection with multimodal inputs can improve many safety-critical systems such as autonomous vehicles (AVs). Motivated by AVs that operate in both day and night, we study multimodal object detection with RGB and thermal cameras, since the latter provides much stronger object signatures under poor illumination. We explore strategies for fusing information from different modalities. Our key contribution is a probabilistic ensembling technique, **ProbEn**, a simple non-learned method that fuses together detections from multi-modalities. We derive ProbEn from Bayes' rule and first principles that assume conditional independence across modalities. Through probabilistic marginalization, ProbEn elegantly handles missing modalities when detectors do not fire on the same object. Importantly, ProbEn also notably improves multimodal detection even when the conditional independence assumption does not hold, e.g., fusing outputs from other fusion methods (both off-the-shelf and trained in-house). We validate ProbEn on two benchmarks containing both aligned (KAIST) and unaligned (FLIR) multimodal images, showing that ProbEn outperforms prior work by more than **13%** in relative performance!

Keywords: Object Detection · Multimodal Detection · Infrared · Thermal · Probabilistic Model · Ensembling · Multimodal Fusion · Uncertainty

1 Introduction

Object detection is a canonical computer vision problem that has been greatly advanced by the end-to-end training of deep neural detectors [45,23]. Such detectors are widely adopted in various safety-critical systems such as autonomous vehicles (AVs) [19,7]. Motivated by AVs that operate in both day and night,

---

[*]Equal contribution. The work was mostly done when authors were with CMU.
[†]Equal supervision.
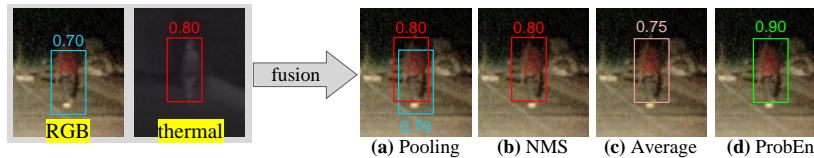
(a) Pooling   (b) NMS   (c) Average   (d) ProbEn

Fig. 1. Multimodal detection via ensembling single-modal detectors. **(a)** A naive approach is to pool detections from each modality, but this will result in multiple detections that overlap the same object. **(b)** To remedy this, one can apply non-maximal suppression (NMS) to suppress overlapping detections from different modalities, which always returns the higher (maximal) scoring detection. Though quite simple, NMS is an effective fusion strategy that has *not* been previously proposed as such. However, NMS fails to incorporate cues from the lower-scoring modality. **(c)** A natural strategy for doing so might average scores of overlapping detections (instead of suppressing the weaker ones) [33,36]. However, this must decrease the reported score compared to NMS. Intuitively, if two modalities agree on a candidate detection, one should *boost* its score. **(d)** To do so, we derive a simple probabilistic ensembling approach, **ProbEn**, to score fusion that increases the score for detections that have strong evidence from multiple modalities. We further extend ProbEn to box fusion in Section 3. Our *non-learned* ProbEn significantly outperforms prior work (Table 2&4).

we study multimodal object detection with RGB and thermal cameras, since the latter can provide much stronger object signatures under poor illumination [26,51,32,12,57,4].

Multimodal Data. There exists several challenges in multimodal detection. One is the lack of data. While there exists large repositories of annotated single-modal datasets (RGB) and pre-trained models, there exists much less annotated data of other modalities (thermal), and even less annotations of them paired together. One often-ignored aspect is the alignment of the modalities: aligning RGB and thermal images requires special purpose hardware, e.g., a beam-splitter [26] or a specialized rack [48] for spatial alignment, and a GPS clock synchronizer for temporal alignment [42]. Fusion on *un*aligned RGB-thermal inputs (cf. Fig. 4) remains relatively unexplored. For example, even annotating bounding boxes is cumbersome because separate annotations are required for each modality, increasing overall cost. As a result, many unaligned datasets annotate only one modality (e.g., FLIR [17]), further complicating multimodal learning.

Multimodal Fusion. The central question in multimodal detection is *how* to fuse information from different modalities. Previous work has explored strategies for fusion at various stages [9,51,32,56,57,4], which are often categorized into early-, mid- and late-fusion. Early-fusion constructs a four-channel RGB-thermal input [49], which is then processed by a (typical) deep network. In contrast, mid-fusion keeps RGB and thermal inputs in different streams and then merges their features downstream within the network (Fig. 2a) [49,36,30]. The vast majority of past work focuses on architectural design of where and how to merge. Our key contribution is the exploration of an extreme variant of *very*-late fusion of detectors trained on separate modalities (Fig. 2b) through *detector ensembling*. Though conceptually simple, ensembling can be effective because one can learn
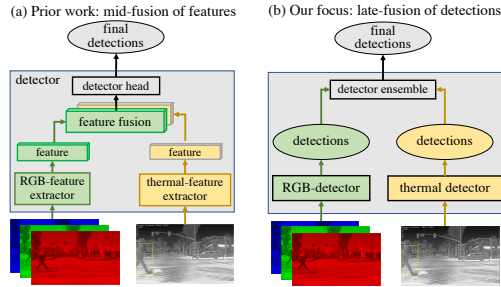
(a) Prior work: mid-fusion of features

(b) Our focus: late-fusion of detections

Fig. 2. High-level comparisons between mid- and late-fusion. **(a)** Past work primarily focuses on mid-fusion, e.g., concatenating features computed by single-modal feature extractors. **(b)** We focus on late-fusion via *detector ensemble* that fuses detections from independent detectors, e.g., two single-modal detectors trained with RGB and thermal images respectively.

from single-modal datasets that often dwarf the size of multimodal datasets. However, ensembling can be practically challenging because different detectors might not fire on the same object. For example, RGB-based detectors often fail to fire in nighttime conditions, implying one needs to deal with "missing" detections during fusion.

Probabilistic Ensembling (ProbEn). We derive our very-late fusion approach, ProbEn, from first principles: simply put, if single-modal signals are conditionally independent of each other given the true label, the optimal fusion strategy is given by Bayes rule [41]. ProbEn requires no learning, and so does not require any multimodal data for training. Importantly, ProbEn elegantly handles "missing" modalities via probabilistic marginalization. While ProbEn is derived assuming conditional independence, we empirically find that it can be used to fuse outputs that are not strictly independent, by fusing outputs from *other* fusion methods (both off-the-shelf and trained in-house). In this sense, ProbEn is a general technique for ensembling detectors. We achieve significant improvements over prior art, both on aligned and unaligned multimodal benchmarks.

Why ensemble? One may ask why detector ensembling should be regarded as an interesting contribution, given that ensembling is a well-studied approach [18,29,3,13] that is often viewed as an "engineering detail" for improving leaderboard performance [31,25,22]. Firstly, we show that the precise ensembling technique matters, and prior approaches proposed in the (single-modal) detection literature such as score-averaging [31,14] or max-voting [52], are not as effective as ProbEn, particularly when dealing with missing modalities. Secondly, to our knowledge, we are the first to propose detector ensembling as a fusion method for multimodal detection. Though quite simple, it is remarkably effective and should be considered a baseline for future research.

2   Related Work

Object Detection and Detector Ensembling. State-of-the-art detectors train deep neural networks on large-scale datasets such as COCO [34] and often focus on architectural design [37,43,44,45]. Crucially, most architectures generate overlapping detections which need to be post-processed with non-maximal suppression (NMS) [10,5,47]. Overlapping detections could also be generated by detectors tuned for different image crops and scales, which typically make use of ensembling techniques for post-processing their output [1,22,25]. Somewhat

surprisingly, although detector ensembling and NMS are widely studied in single-modal RGB detection, to the best of our knowledge, they have *not* been used to (very) late-fuse multimodal detections; we find them remarkably effective.

Multimodal Detection, particularly with RGB-thermal images, has attracted increasing attention. The KAIST pedestrian detection dataset [26] is one of the first benchmarks for RGB-thermal detection, fostering growth of research in this area. Inspired by the successful RGB-based detectors [45,43,37], current multimodal detectors train deep models with various methods for fusing multimodal signals [9,51,32,56,57,4,58,57,28]. Most of these multimodal detection methods work on aligned RGB-thermal images, but it is unclear how they perform on heavily unaligned modalities such as images in Fig. 4 taken from FLIR dataset [17]. We study multimodal detection under both aligned and unaligned RGB-thermal scenarios. **Multimodal fusion** is the central question in multimodal detection. Compared to early-fusion that simply concatenates RGB and thermal inputs, mid-fusion of single-modal features performs better [49]. Therefore, most multimodal methods study how to fuse features and focus on designing new network architectures [49,36,30]. Because RGB-thermal pairs might not be aligned, some methods train an RGB-thermal translation network to synthesize aligned pairs, but this requires annotations in each modality [12,38,27]. Interestingly, few works explore learning from unaligned data that are annotated only in single modality; we show that mid-fusion architectures can still learn in this setting by acting as an implicit alignment network. Finally, few fusion architectures explore (very) late fusion of single-modal detections via detector ensembling. Most that do simply take heuristic (weighted) averages of confidence scores [20,32,57]. In contrast, we introduce probabilistic ensembling (ProbEn) for late-fusion, which significantly outperforms prior methods on both aligned and unaligned RGB-thermal data.

3   Fusion Strategies for Multimodal Detection

We now present multimodal fusion strategies for detection. We first point out that **single-modal** detectors are viable methods for processing multimodal signals, and so include them as a baseline. We also include fusion baselines for **early-fusion**, which concatenates RGB and thermal as a four-channel input, and **mid-fusion**, which concatenates single-modal features inside a network (Fig. 2). As a preview of results, we find that mid-fusion is generally the most effective baseline (Table 1). Surprisingly, this holds even for unaligned data that is annotated with a single modality (Fig. 4), indicating that mid-fusion can perform some implicit alignment (Table 3).

We describe strategies for late-fusing detectors from different modalities, or detector ensembling. We begin with a naive approach (Fig. 1). Late-fusion needs to fuse scores and boxes; we discuss the latter at the end of this section.

Naive Pooling. The possibly simplest strategy is to naively pool detections from multiple modalities together. This will probably result in multiple detections overlapping the same ground-truth object (Fig. 1a).

Algorithm 1 Multimodal Fusion by NMS or ProbEn

1: Input: class priors π_k for $k \in \{1, \ldots, K\}$; the flag of fusion method (`NMS` or `ProbEn`); set $\mathcal{D}$: detections from multiple modalities. Each detection $d = (\mathbf{y}, \mathbf{z}, m) \in \mathcal{D}$ contains classification posteriors $\mathbf{y}$, box coordinates $\mathbf{z}$ and modality tag m.
2: Initialize set of fused detections $\mathcal{F} = \{\}$
3: **while** $\mathcal{D} \neq \emptyset$ **do**
4:      Find detection $d \in \mathcal{D}$ with largest posterior
5:      Find all detections in $\mathcal{D}$ that overlap d (e.g., > 0.5 IoU), denoted as $\mathcal{T} \subseteq \mathcal{D}$
6:      **if** `NMS` **then**
7:          $d' \leftarrow d$
8:      **else if** `ProbEn` **then**
9:          Find highest scoring detection in $\mathcal{T}$ of each modality, denoted as $\mathcal{S} \subseteq \mathcal{T}$
10:         Compute d' from $\mathcal{S}$ by fusing scores $\mathbf{y}$ with Eq. (4) and boxes $\mathbf{z}$ with Eq. (8)
11:     **end if**
12:     $\mathcal{F} \leftarrow \mathcal{F} + \{d'\}$, $\mathcal{D} \leftarrow \mathcal{D} - \mathcal{T}$
13: **end while**
14: **return** set $\mathcal{F}$ of fused detections

Non-Maximum Supression (NMS). The natural solution for dealing with overlapping detections is NMS, a crucial component in contemporary RGB detectors [14,60,24]. NMS finds bounding box predictions with high spatial overlap and remove the lower-scoring bounding boxes. This can be implemented in a sequential fashion via sorting of predictions by confidence, as depicted by Algorithm 1, or in a parallel fashion amenable to GPU computation [6]. While NMS has been used to ensemble single-modal detectors [47], it has (surprisingly) *not* been advocated for fusion of *multi*-modal detectors. We find it be shockingly effective, outperforming the majority of past work on established benchmarks (Fig. 2). Specifically, when two detections from two different modalities overlap (e.g., IoU>0.5), NMS simply keeps the higher-score detection and suppresses the other (Fig. 1b). This allows each modality to "shine" where effective – thermal detections tend to score high (and so will be selected) when RGB detections perform poorly due to poor illumination conditions. That said, rather than selecting one modality at the global image level (e.g., day-time vs. night time), NMS selects one modality at the local bounding box level. However, in some sense, NMS fails to "fuse" information from multiple modalities together, since each of the final detections are supported by only one modality.

Average Fusion. To actually fuse multimodal information, a straightforward strategy is to modify NMS to average confidence scores of overlapping detections from different modalities, rather than suppressing the weaker modality. Such an averaging has been proposed in prior work [52,36,32]. However, averaging scores will necessarily *decrease* the NMS score which reports the max of an overlapping set of detections (Fig. 1c). Our experiments demonstrate that averaging produces worse results than NMS and single-modal detectors. Intuitively, if two modalities agree that there exist a detection, fusion should *increase* the overall confidence rather than decrease.

Probabilistic Ensembling (ProbEn). We derive our probabilistic approach for late-fusion of detections by starting with how to fuse detection scores

(Algorithm 1). Assume we have an object with label y (e.g., a "person") and measured signals from two modalities: x_1 (RGB) and x_2 (thermal). We write out our formulation for two modalities, but the extension to multiple (evaluated in our experiments) is straightforward. Crucially, we assume measurements are conditionally independent given the object label y:

$$p(x_1, x_2|y) = p(x_1|y)p(x_2|y) \tag{1}$$

This can also be written as $p(x_1|y) = p(x_1|x_2, y)$, which may be easier to intuit. Given the person label y, predict its RGB appearance x_1; if this prediction would not change the given knowledge of the thermal signal x_2, then conditional independence holds. We wish to infer labels given multimodal measurements:

$$p(y|x_1, x_2) = \frac{p(x_1, x_2|y)p(y)}{p(x_1, x_2)} \propto p(x_1, x_2|y)p(y) \tag{2}$$

By applying the conditional independence assumption from (1) to (2), we have:

$$p(y|x_1, x_2) \propto \ p(x_1|y)p(x_2|y)p(y) \propto \frac{p(x_1|y)p(y)p(x_2|y)p(y)}{p(y)} \tag{3}$$

$$\propto \frac{p(y|x_1)p(y|x_2)}{p(y)} \tag{4}$$

The above suggests a simple approach to fusion that is provably optimal when single-modal features are conditionally-independent of the true object label:

1. Train independent single-modal classifiers that predict the distributions over the label y given each individual feature modality $p(y|x_1)$ and $p(y|x_2)$.
2. Produce a final score by multiplying the two distributions, dividing by the class prior distribution, and normalizing the final result (4) to sum-to-one.

To obtain the class prior $p(y)$, we can simply normalize the counts of per-class examples. Extending ProbEn (4) to M modalities is simple:

$$p(y|\{x_i\}_{i=1}^M) \propto \frac{\Pi_{i=1}^M p(y|x_i)}{p(y)^{M-1}}. \tag{5}$$

Independence assumptions. ProbEn is optimal given the independence assumption from (1). Even when such independence assumptions do not hold in practice, the resulting models may still be effective [11] (i.e., just as assumptions of Gaussianity can still be useful even if strictly untrue [29,41]). Interestingly, many fusion methods including NMS and averaging make the same underlying assumption, as discussed in [29]. In fact, [29] points out that Average Fusion (which averages class posteriors) makes an even stronger assumption: posteriors do not deviate dramatically from class priors. This is likely not true, as corroborated by the poor performance of averaging in our experiments (despite its apparent widespread use [52,36,32]).

Relationship to prior work. To compare to prior fusion approaches that tend to operate on logit scores, we rewrite the single-modal softmax posterior for
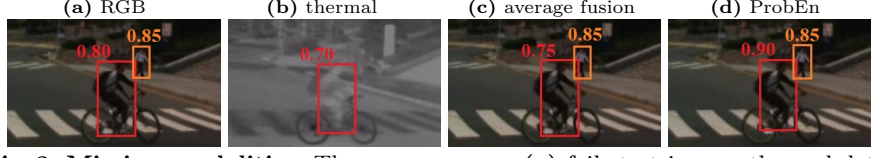
(a) RGB   **(b)** thermal   **(c)** average fusion   **(d)** ProbEn

Fig. 3. Missing modalities. The orange-person **(a)** fails to trigger a thermal detection **(b)**, resulting in a single-modal RGB detection (0.85 confidence). To generate an output set of detections (for downstream metrics such as average precision), this detection must be compared to the fused multimodal detection of the red-person (RGB: 0.80, thermal: 0.70). **(c)** averaging confidences for the red-person lowers their score (0.75) below the orange-person, which is unintuitive because additional detections should boost confidence. **(d)** ProbEn increases the red-person fused score to 0.90, allowing for proper comparisons to single-modal detections.

class-k given modality i in terms of single-modal logit score $s_i[k]$. For notational simplicity, we suppress its dependence on the underlying input modality x_i: $p(y{=}k|x_i) = \frac{\exp(s_i[k])}{\sum_j \exp(s_i[j])} \propto \exp(s_i[k])$, where we exploit the fact that the partition function in the denominator is not a function of the class label k. We now plug the above into Eq. (5):

$$p(y{=}k|\{x_i\}_{i=1}^M) \propto \frac{\Pi_{i=1}^M p(y{=}k|x_i)}{p(y{=}k)^{M-1}} \propto \frac{\exp(\sum_{i=1}^M s_i[k])}{p(y{=}k)^{M-1}} \tag{6}$$

ProbEn is thus equivalent to *summing logits*, dividing by the class prior and normalizing via a softmax. Our derivation (6) reveals that summing logits without the division may over-count class priors, where the over-counting grows with the number of modalities M. The supplement shows that dividing by class posteriors $p(y)$ marginally helps. In practice, we empirically find that assuming uniform priors works surprisingly well, even on imbalanced datasets. This is the default for our experiments, unless otherwise noted.

    Missing modalities. Importantly, summing and averaging behave profoundly differently when fusing across "missing" modalities (Fig. 3). Intuitively, different single-modal detectors often do not fire on the same object. This means that to output a final set of detections above a confidence threshold (e.g., necessary for computing precision-recall metrics), one will need to compare scores from fused multi-modal detections with single modal detections, as illustrated in Fig. 3. ProbEn elegantly deals with missing modalities because *probabilistically-normalized* multi-modal posteriors $p(y|x_1, x_2)$ can be directly compared with single-modal posteriors $p(y|x_1)$.

    Bounding Box Fusion. Thus far, we have focused on fusion of class posteriors. We now extend ProbEn to probabilistically fuse bounding box (bbox) coordinates of overlapping detections. We repurpose the derivation from (4) for a continuous bbox label rather than a discrete one. Specifically, we write $\mathbf{z}$ for the continuous random variable defining the bounding box (parameterized by its centroid, width, and height) associated with a given detection. We assume single-modal detections provide a posterior $p(\mathbf{z}|x_i)$ that takes the form of a Gaussian with a single variance σ_i^2, i.e., $p(\mathbf{z}|x_i) = \mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2\mathbf{I})$ where $\boldsymbol{\mu}_i$ are box coordinates predicted from modality i. We also assume a uniform prior on $p(\mathbf{z})$, implying
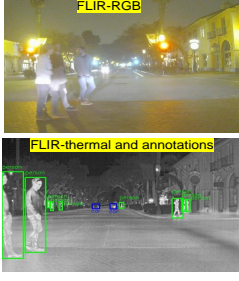
Fig. 4. RGB and thermal images are unaligned both spatially and temporally in FLIR [17], which annotates only thermal images. As a result, prior methods relies on thermal and drop the RGB modality. We find mid-fusion, taking both RGB and thermal as input, notably improves detection accuracy. When late-fusing detections computed by the mid-fusion and thermal-only detectors, our ProbEn yields much better performance (Table 3 and 4).

bbox coordinates can lie anywhere in the image plane. Doing so, we can write

$$p(\mathbf{z}|x_1, x_2) \propto p(\mathbf{z}|x_1)p(\mathbf{z}|x_2) \propto \exp\big(\frac{\|\mathbf{z} - \boldsymbol{\mu}_1\|^2}{-2\sigma_1^2}\big) \exp\big(\frac{\|\mathbf{z} - \boldsymbol{\mu}_2\|^2}{-2\sigma_2^2}\big) \qquad (7)$$

$$\propto \exp\big(\frac{\|\mathbf{z} - \boldsymbol{\mu}\|^2}{-2(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2})}\big), \quad \text{where} \quad \boldsymbol{\mu} = \frac{\frac{\boldsymbol{\mu}_1}{\sigma_1^2} + \frac{\boldsymbol{\mu}_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} \qquad (8)$$

We refer the reader to the supplement for a detailed derivation. Eq. (8) suggests a simple way to probabilistically fuse box coordinates: compute a weighted average of box coordinates, where weights are given by the inverse covariance. We explore three methods for setting σ_i^2. The first method "avg" fixes σ_i^2=1, amounting to simply averaging bounding box coodinates. The second "s-avg" approximates $\sigma_i^2 \approx \frac{1}{p(y=k|x_i)}$, implying that more confident detections should have a higher weight when fusing box coordinates. This performs marginally better than simply averaging. The third "v-avg" train the detector to predict regression *variance*/uncertainty using the Gaussian negative log likelihood (GNLL) loss [39] alongside the box regression loss. Interestingly, incorporating GNLL not only produces better variance/uncertainty estimate helpful for fusion but also improves detection performance of the trained detectors (details in supplement).

4 Experiments

We validate different fusion methods on two datasets: KAIST [26] which is released under the Simplified BSD License, and FLIR [17] (Fig. 4), which allows for non-commercial educational and research purposes. Because the two datasets contain personally identifiable information such as faces and license plates, we assure that we (1) use them only for research, and (2) will release our code and models to the public without redistributing the data. We first describe implementation details and then report the experimental results on each dataset (alongside their evaluation metrics) in separate subsections.

4.1 Implementation

We conduct experiments with PyTorch [40] on a single GPU (Nvidia GTX 2080). We train our detectors (based on Faster-RCNN) with Detectron2 [50], using
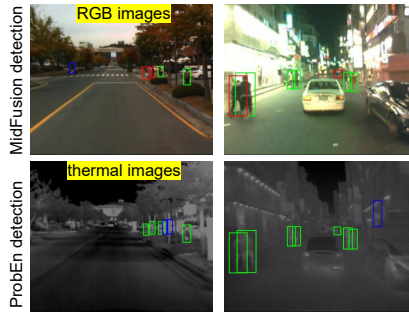
Fig. 5. Detections overlaid on two KAIST testing examples in columns. **Top**: detections by our mid-fusion model. **Bottom**: detections by our ProbEn by fusing detections of thermal-only and mid-fusion models. Green, red and blue boxes stand for true positives, false negative (miss-detection) and false positives. Visually, ProbEn performs much better than the mid-fusion model, which is already comparable to the prior work as shown in Table 1 and 2.

SGD and learning rate 5e-3. For data augmentation, we adopt random flipping and resizing. We pre-train our detector on COCO dataset [34]. As COCO has only RGB images, fine-tuning the pre-trained detector on thermal inputs needs careful pre-processing of thermal images (detailed below).

Pre-processing. All RGB and thermal images have intensity in [0, 255]. In training an RGB-based detector, RGB input images are commonly processed using the mean subtraction [50] where the mean values are computed over all the training images. Similarly, we calculate the mean value (135.438) in the thermal training data. We find using a precise mean subtraction to process thermal images yields better performance when fine-tuning the pre-trained detector.

Stage-wise Training. We fine-tune the pre-trained detector to train single-modal detectors and the early-fusion detectors. To train a mid-fusion detector, we truncate the *already-trained* single-modal detectors, concatenate features add a new detection head and train the whole model (Fig. 2a). The late-fusion methods fuse detections from (single-modal) detectors. Note that all the late-fusion methods are *non-learned*. We also experimented with learning-based late-fusion methods (e.g., learning to fuse logits) but find them to be only marginally better than ProbEn (9.08 vs. 9.16 in LAMR using argmax box fusion). Therefore, we focus on the non-learned late fusion methods in the main paper and study learning-based ones in the supplement.

Post-processing. When ensembling two detectors, we find it crucial to calibrate scores particularly when we we fuse detections from our in-house models and off-the-shelf models released by others. We adopt the simple temperature scaling for score calibration [21]. Please refer to the supplement for details.

4.2 Multimodal Pedestrian Detection on KAIST

Dataset. The KAIST dataset is a popular multimodal benchmark for pedestrian detection [26]. In KAIST, RGB and thermal images are aligned with a beam-splitter, and have resolutions of 640x480 and 320x256, respectively. We resize thermal images to 640x480 during training. KAIST also provides day/night tags for breakdown analysis. The original KAIST dataset contains 95,328 RGB-thermal image pairs, which are split into a training set (50,172) and a testing set (45,156). Because the original KAIST dataset contains noisy annotations, the literature introduces cleaned version of the train/test sets: a sanitized train-set

baselines		Day	Night	All
RGB		14.56	27.42	18.67
Thermal		24.59	7.76	18.99
EarlyFusion		26.30	6.61	19.36
MidFusion		17.55	9.30	14.48
Pooling		37.92	22.61	32.68
score-fusion	*box-fusion*	*Day*	*Night*	*All*
max	argmax	13.25	6.42	10.78
max	avg	13.25	6.65	10.89
max	s-avg	13.35	6.65	10.96
max	v-avg	13.19	6.65	10.79
avg	argmax	21.68	15.16	19.53
avg	avg	21.59	15.46	19.47
avg	s-avg	21.67	15.46	19.55
avg	v-avg	21.51	15.46	19.42
ProbEn	argmax	10.21	5.45	8.62
ProbEn	avg	10.14	5.41	8.58
ProbEn	s-avg	10.27	5.41	8.67
ProbEn	v-avg	9.93	5.41	8.50
$ProbEn_3$	argmax	13.67	6.31	11.00
$ProbEn_3$	avg	9.07	4.89	7.68
$ProbEn_3$	s-avg	**9.07**	**4.89**	7.68
$ProbEn_3$	v-avg	**9.07**	**4.89**	**7.66**

Table 1. Ablation study on KAIST (LAMR↓ in %). The upper panel shows that (1) RGB-only and Thermal-only detectors perform notably better than each other on *Day* and *Night* respectively, and (2) MidFusion strikes a balance and performs better overall. In the lower panel, we focus on the very-late fusion of RGB and Thermal. We ablate methods for *score fusion* (max as in NMS, avg and ProbEn), and *box fusion* (argmax as in NMS, ProbEn that uses avg, s-avg or v-avg). Somewhat surprisingly, "max + argmax", or NMS, performs quite well on both *Day* and *Night*; average score fusion performs poorly because it double counts class prior. As for box fusion, using the learned variance / uncertainty by v-avg performs better than the heuristic methods (avg and s-avg). Our ProbEn performs significantly better and $ProbEn_3$ is the best by fusing three models: RGB, Thermal, and MidFusion.

(7,601 examples) [32] and a cleaned test-set (2,252 examples) [35]. We also follow the literature [26] to evaluate under the "reasonable setting" for evaluation by ignoring annotated persons that are occluded (tagged by KAIST) or too small (<55 pixels). We follow this literature for fair comparison with recent methods.

Metric. We measure detection performance with the Log-Average Miss Rate (LAMR), which is a standard metric in pedestrian detection [15] and KAIST [26]. LAMR is computed by averaging the miss rate (false negative rate) at nine false positives per image (FPPI) rates evenly spaced in log-space from the range 10^{-2} to 10^0 [26]. It does not evaluate the detections that match to ignored ground-truth [15,26]. A true positive is a detection that matches a ground-truth object with IoU>0.5 [26]; false positives are detections that do not match any ground-truth; false negatives are miss-detections.

Ablation Study on KAIST Table 1 shows ablation studies on KAIST. Single modal detectors tend to work well in different environments, with RGB detectors working on well-lit day images while Thermal working well on nighttime images. EarlyFusion reduces the miss rate by a modest amount, while MidFusion is more effective. Naive strategies for late fusion (such as pooling together detections from different modalities) are quite poor because they generate many repeated detections on the same object, which are counted as false positives. Interestingly, simple NMS that has max score fusion and argmax box fusion, is quite effective at removing overlapping detections from different modalities, already outperforming Early and MidFusion. Instead of suppressing the weaker modality, one might average the scores of overlapping detections but this is quite ineffective because it always decreases the score from NMS. Intuitively, one should increase the score when different modalities agree on a detection.

Method	Day	Night	All
HalfwayFusion [36]	36.84	35.49	36.99
RPN+BDT [30]	30.51	27.62	29.83
TC-DET [4]	34.81	10.31	27.11
IATDNN [20]	27.29	24.41	26.37
IAF R-CNN [33]	21.85	18.96	20.95
SyNet [2]	22.64	15.80	20.19
CIAN [56]	14.77	11.13	14.12
MSDS-RCNN [32]	12.22	7.82	10.89
AR-CNN [57]	9.94	8.38	9.34
MBNet [58]	8.28	7.86	8.13
MLPD [28]	7.95	6.95	7.58
GAFF [55]	8.35	3.46	6.48
MaxFusion (NMS)	13.25	6.42	10.78
ProbEn	9.93	5.41	8.50
ProbEn$_3$	9.07	4.89	7.66
ProbEn$_3$ w/ MLPD	7.81	5.02	6.76
ProbEn$_3$ w/ GAFF	**6.04**	**3.59**	**5.14**

Table 2. Benchmarking on KAIST measured by % LAMR↓. We report numbers from the respective papers. Results are comparable to Table 1. *Simple probabilistic ensembling of independently-trained detectors (ProbEn) outperforms $\frac{9}{12}$ methods on the leaderboard. Infact, even NMS (MaxFusion) outperforms $\frac{8}{12}$ methods, indicating the under-appreciated effectiveness of detector-ensembling as a multimodal fusion technique.* Performance further increases when adding a MidFusion detector to the probabilistic ensemble (ProbEn$_3$). Replacing our in-house MidFusion with off-the-shelf mid-fusion detectors MLPD [28] and GAFF [55] significantly boosts the state-of-art from 6.48 to 5.14! This shows ProbEn remains effective even when fusing models for which conditional independence does not hold.

ProbEn accomplishes this by probabilistic integration of information from the RGB and Thermal single-modal detectors. Moreover, it can be further improved by probabilisitcally fusing coordinates of overlapping boxes. Lastly, ProbEn$_3$ that ensembles three models (RGB, thermal and MidFusion), performs the best.

Qualitative Results are displayed in Fig. 5. Visually, ProbEn detects all persons, while the MidFusion model has multiple false negatives / miss-detections.

Quantitative Comparison on KAIST Compared Methods. Among many prior methods, we particularly compare against four recent ones: AR-CNN [57], MBNet [58], MLPD [28], and GAFF [55]. AR-CNN focuses on weakly-unaligned RGB-thermal pairs and explores multiple heuristic methods for fusing features, scores and boxes. MBNet addresses modality imbalance w.r.t illumination and features to improve detection; both MLPD and GAFF are mid-fusion methods that design sophisticated network architectures; MLPD adopts aggressive data augmentation techniques and GAFF extensively exploits attentive modules to fuse multimodal features. Table 2 lists more methods.

Results. Table 2 compares ProbEn against the prior work. ProbEn+ that ensembles three models trained in-house (RGB, Thermal, and MidFusion) achieves competitive performance (7.95 LAMR) against the prior art. When replacing our MidFusion detector with off-the-shelf mid-fusion detectors [28,55], ProbEn++ significantly outperforms all the existing methods, boosting the performance from the prior art 6.48 to 5.14! This clearly shows that ProbEn works quite well when the conditional independence assumption does not hold, i.e., fusing outputs from other fusion methods (both off-the-shelf and trained in-house). As ProbEn performs better than past work as a non-learned solution, we argue that it should serve as a new baseline for future research on multimodal detection.

baselines		Day	Night	All
Thermal		75.35	82.90	79.24
EarlyFusion		77.37	79.56	78.80
MidFusion		79.37	81.64	80.53
Pooling		52.57	55.15	53.66
score-fusion	box-fusion	Day	Night	All
max	argmax	81.91	84.42	83.14
max	avg	81.84	84.62	83.21
max	s-avg	81.85	84.48	83.19
max	v-avg	81.80	85.07	83.31
avg	argmax	81.34	84.69	82.65
avg	avg	81.26	84.81	82.91
avg	s-avg	81.26	84.72	82.89
avg	v-avg	81.26	85.39	83.03
$ProbEn_3$	argmax	82.19	84.73	83.27
$ProbEn_3$	avg	82.19	84.91	83.63
$ProbEn_3$	s-avg	82.20	84.84	83.61
$ProbEn_3$	v-avg	**82.21**	**85.56**	**83.76**

Table 3. Ablation study on FLIR day/night scenes (AP↑ in percentage with IoU>0.5). Compared to thermal-only detector, incorporating RGB by EarlyFusion and MidFusion notably improves performance. Late-fusion (lower panel) ensembles three detectors: Thermal, EarlyFusion and MidFusion. All the explored late-fusion methods lead to better performance than MidFusion. In particular, ProbEn performs the best. Moreover, similar to the results on KAIST, using predicted uncertainty to fuse boxes (v-avg) performs better than the other two heuristic box fusion methods, avg that naively averages box coordinates and s-avg that uses classification scores to weighted average box coordinates.

4.3   Multimodal Object Detection on FLIR

Dataset. The FLIR dataset [17] consists of RGB images (captured by a FLIR BlackFly RGB camera with 1280x1024 resolution) and thermal images (acquired by a FLIR Tau2 thermal camera 640x512 resolution). We resize all images to resolution 640x512. FLIR has $10,228$ *unaligned* RGB-thermal image pairs and annotates only for thermal (Fig. 4). Image pairs are split into train-set ($8,862$ images) and a validation set ($1,366$ images). FLIR evaluates on three classes which have imbalanced examples [8,27,54,38,12]: $28,151$ persons, $46,692$ cars, and $4,457$ bicycles. Following [54], we remove 108 thermal images in the val-set that do not have the RGB counterparts. For breakdown analysis w.r.t day/night scenes, we manually tag the validation images with "day" (768) and "night" (490). We will release our annotations to the public.

Misaligned modalities. Because FLIR's RGB and thermal images are heavily unaligned, it labels only thermal images and does not have RGB annotations. We can still train Early and MidFusion models using multimodal inputs and the thermal annotations. These detectors might learn to internally align the unaligned modalities to predict bounding boxes according to the thermal annotations. Because we do not have an RGB-only detector, our ProbEn ensembles EarlyFusion, MidFusion, and thermal-only detectors.

Metric. We measure performance using Average Precision (AP) [16,46]. Precision is computed over testing images within a single class, with true positives that overlap ground-truth bounding boxes (e.g., IoU>0.5). Computing the average precision (AP) across all classes measures the performance in multi-class object detection. Following [12,38,54,27,8], we define a true positive as a detection that overlaps a ground-truth with IoU>0.5. Note that AP used in the the multimodal detection literature is different from mAP [34], which averages over different AP's computed with different IoU thresholds.
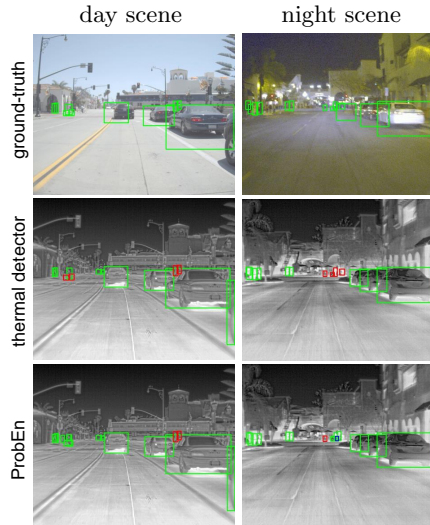
day scene   night scene

ground-truth / thermal detector / ProbEn

Fig. 6. Detections overlaid on two FLIR testing images (in columns) with RGB (top) and thermal images (middle and bottom). To avoid clutter, we do not mark class labels for the bounding boxes. Ground-truth annotations are shown on the RGB, emphaszing that RGB and thermal images are strongly unaligned. On the thermal images, we compare thermal-only (mid-row) and our ProbEn (bottom-row) models. Green, red and blue boxes stand for true positives, false negative (mis-detected persons) and false positives. In particular, in the second column, the thermal-only model has many false negatives (or miss-detections), which are "bicycles". Understandably, thermal cameras will not capture bicycles because they do not emit heat. In contrast, RGB capture bicycle signatures better than thermal. This explains why our fusion performs better on bicycles.

Ablation study on FLIR We compare our fusion methods in Table 3, along with qualitative results in Fig. 6. We analyze results using our day/night tags. Compared to the single-modal detector (Thermal), our learning-based early-fusion (EarlyFusion) and mid-fusion (MidFusion) produce better performance. MidFusion outperforms EarlyFusion, implying that end-to-end learning of fusing features better handles mis-alignment between RGB and thermal images. By applying late-fusion methods to detections of Thermal, EarlyFusion and MidFusion detectors, we boost detection performance. Note that typical ensembling methods in the single-modal (RGB) detection literature [52,36,32] often use max / average score fusion, and argmax / average box fusion, which are outperformed by our ProbEn. This suggests that ProbEn should be potentially a better ensembling method for object detection.

Quantitative Comparison on FLIR Compared Methods. We compare against prior methods including ThermalDet [8], BU [27], ODSC [38], MM-TOD [12], CFR [54], and GAFF [55]. As FLIR does not have aligned RGB-thermal images and only annotates thermal images, many methods exploit domain adaptation that adapts a pre-trained RGB detector to thermal input. For example, MMTOD [12] and ODSC [38] adopt the image-to-image-translation technique [59,53] to generate RGB from thermal, hypothesizing that this helps train a better multimodal detector by finetuning a detector that is pre-trained over large-scale RGB images. BU [27] operates such a translation/adaptation on features that generates thermal features to be similar to RGB features. ThermalDet [8] exclusively exploits thermal images and ignores RGB images; it proposes to combine features from multiple layers for the final detection.

Method	Bicycle	Person	Car	All
MMTOD-CG [12]	50.26	63.31	70.63	61.40
MMTOD-UNIT [12]	49.43	64.47	70.72	61.54
ODSC [38]	55.53	71.01	82.33	69.62
CFR3 [54]	55.77	74.49	84.91	72.39
BU(AT,T) [27]	56.10	76.10	87.00	73.10
BU(LT,T) [27]	57.40	75.60	86.50	73.20
GAFF [55]	—	—	—	72.90
ThermalDet [8]	60.04	78.24	85.52	74.60
Thermal	62.63	84.04	87.11	79.24
EarlyFusion	63.43	85.27	87.69	78.80
MidFusion	69.80	84.16	87.63	80.53
ProbEn$_3$	**73.49**	**87.65**	**90.14**	**83.76**

Table 4. Benchmarking on FLIR measured by AP↑ in percentage with IoU>0.5 with breakdown on the three categories. Perhaps surprisingly, end-to-end training on thermal already outperforms all the prior methods, presumably because of using a better pre-trained model (Faster-RCNN). Importantly, our ProbEn increases AP from prior art 74.6% to 84.4%! These results are comparable to Table 3.

GAFF [55] trains on RGB-thermal image with a sophisticated attention module that fuse single-modal features. Perhaps because the complexity of the attention module, GAFF is limited to using small network backbones (ResNet18 and VGG16). Somewhat surprisingly, to the best of our knowledge, there is no prior work that trained early-fusion or mid-fusion deep networks (Fig. 2a) on the heavily unaligned RGB-thermal image pairs (like in FLIR) for multimodal detection. We find directly training them performs much better than prior work (Table 4).

Results. Table 4 shows that all our methods outperform the prior art. Our single-modal detector (trained on thermal images) achieves slightly better performance than ThermalDet [8], which also exclusively trains on thermal images. This is probably because we use a better pre-trained Faster-RCNN model provided by the excellent Detectron2 toolbox. Surprisingly, our simpler EarlyFusion and MidFusion models achieve big boosts over the thermal-only model (Thermal), while MidFusion performs much better. This confirms our hypothesis that fusing features better handles mis-alignment of RGB-thermal images than the early-fusion method. Our ProbEn performs the best, significantly better than all compared methods! Notably, our fusion methods boost "bicycle" detection. We conjecture that bicycles do not emit heat to deliver strong signatures in thermal, but are more visible in RGB; fusing them greatly improves bicycle detection.

5   Discussion and Conclusions

We explore different fusion strategies for multimodal detection under both aligned and unaligned RGB-thermal images. We show that non-learned probabilistic fusion, ProbEn, significantly outperforms prior approaches. Key reasons for its strong performance are that (1) it can take advantage of highly-tuned single-modal detectors trained on large-scale single-modal datasets, and (2) it can deal with missing detections from particular modalities, a common occurrence when fusing together detections. One by-product of our diagnostic analysis is the remarkable performance of NMS as a fusion technique, precisely because it exploits the same key insights. Our ProbEn yields >**13%** relative improvement over prior work, both on aligned and unaligned multimodal benchmarks.

References

1. Akiba, T., Kerola, T., Niitani, Y., Ogawa, T., Sano, S., Suzuki, S.: Pfdet: 2nd place solution to open images challenge 2018 object detection track. arXiv:1809.00778 (2018) 3
2. Albaba, B.M., Ozer, S.: Synet: An ensemble network for object detection in uav images. In: International Conference on Pattern Recognition (ICPR) (2021) 11
3. Bauer, E., Kohavi, R.: An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine learning **36**(1), 105–139 (1999) 3
4. Bertini, M., del Bimbo, A.: Task-conditioned domain adaptation for pedestrian detection in thermal imagery. In: ECCV (2020) 2, 4, 11
5. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-nms–improving object detection with one line of code. In: ICCV (2017) 3
6. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: Yolact: Real-time instance segmentation. In: ICCV (2019) 5
7. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020) 1
8. Cao, Y., Zhou, T., Zhu, X., Su, Y.: Every feature counts: An improved one-stage detector in thermal imagery. In: IEEE International Conference on Computer and Communications (ICCC) (2019) 12, 13, 14
9. Choi, H., Kim, S., Park, K., Sohn, K.: Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks. In: International Conference on Pattern Recognition (ICPR) (2016) 2, 4
10. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005) 3
11. Dawid, A.P.: Conditional independence in statistical theory. Journal of the Royal Statistical Society: Series B (Methodological) **41**(1), 1–15 (1979) 6
12. Devaguptapu, C., Akolekar, N., M Sharma, M., N Balasubramanian, V.: Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In: CVPR Workshops (2019) 2, 4, 12, 13, 14
13. Dietterich, T.G.: Ensemble methods in machine learning. In: International workshop on multiple classifier systems. pp. 1–15. Springer (2000) 3
14. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: CVPR (2009) 3, 5
15. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. PAMI **34**(4), 743–761 (2011) 10
16. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015) 12
17. FLIR: Flir thermal dataset for algorithm training. https://www.flir.in/oem/adas/adas-dataset-form (2018) 2, 4, 8, 12
18. Freund, Y., Schapire, R.E., et al.: Experiments with a new boosting algorithm. In: icml. vol. 96, pp. 148–156. Citeseer (1996) 3
19. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012) 1
20. Guan, D., Cao, Y., Yang, J., Cao, Y., Yang, M.Y.: Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. Information Fusion **50**, 148–157 (2019) 4, 11

21. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. arXiv:1706.04599 (2017) 9
22. Guo, R., Cui, C., Du, Y., Meng, X., Wang, X., Liu, J., Zhu, J., Feng, Y., Han, S.: 2nd place solution in google ai open images object detection track 2019. arXiv:1911.07171 (2019) 3
23. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017) 1
24. Hosang, J., Benenson, R., Schiele, B.: Learning non-maximum suppression. In: CVPR (2017) 5
25. Huang, Z., Chen, Z., Li, Q., Zhang, H., Wang, N.: 1st place solutions of waymo open dataset challenge 2020–2d object detection track. arXiv:2008.01365 (2020) 3
26. Hwang, S., Park, J., Kim, N., Choi, Y., So Kweon, I.: Multispectral pedestrian detection: Benchmark dataset and baseline. In: CVPR (2015) 2, 4, 8, 9, 10
27. Kiew, M.Y., Bagdanov, A.D., Bertini, M.: Bottom-up and layer-wise domain adaptation for pedestrian detection in thermal images. ACM Transactions on Multimedia Computing Communications and Applications (2020) 4, 12, 13, 14
28. Kim, J., Kim, H., Kim, T., Kim, N., Choi, Y.: Mlpd: Multi-label pedestrian detector in multispectral domain. IEEE Robotics and Automation Letters 6(4), 7846–7853 (2021) 4, 11
29. Kittler, J., Hatef, M., Duin, R.P., Matas, J.: On combining classifiers. IEEE PAMI 20(3), 226–239 (1998) 3, 6
30. Konig, D., Adam, M., Jarvers, C., Layher, G., Neumann, H., Teutsch, M.: Fully convolutional region proposal networks for multispectral person detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 49–56 (2017) 2, 4, 11
31. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25, 1097–1105 (2012) 3
32. Li, C., Song, D., Tong, R., Tang, M.: Multispectral pedestrian detection via simultaneous detection and segmentation. arXiv:1808.04818 (2018) 2, 4, 5, 6, 10, 11, 13
33. Li, C., Song, D., Tong, R., Tang, M.: Illumination-aware faster r-cnn for robust multispectral pedestrian detection. Pattern Recognition 85, 161–171 (2019) 2, 11
34. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 3, 9, 12
35. Liu, J., Zhang, S., Wang, S., Metaxas, D.: Improved annotations of test set of kaist (2018) 10
36. Liu, J., Zhang, S., Wang, S., Metaxas, D.N.: Multispectral deep neural networks for pedestrian detection. BMVC (2016) 2, 4, 5, 6, 11, 13
37. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV. pp. 21–37. Springer (2016) 3, 4
38. Munir, F., Azam, S., Rafique, M.A., Sheri, A.M., Jeon, M.: Thermal object detection using domain adaptation through style consistency. arXiv:2006.00821 (2020) 4, 12, 13, 14
39. Nix, D.A., Weigend, A.S.: Estimating the mean and variance of the target probability distribution. In: Proceedings of 1994 ieee international conference on neural networks (ICNN'94). vol. 1, pp. 55–60. IEEE (1994) 8
40. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017) 8

41. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Elsevier (2014) 3, 6
42. Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A.Y.: Ros: an open-source robot operating system. In: ICRA workshop on open source software. vol. 3, p. 5. Kobe, Japan (2009) 2
43. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR (2016) 3, 4
44. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: CVPR (2017) 3
45. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS (2015) 1, 3, 4
46. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015) 12
47. Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing **107**, 104117 (2021) 3, 5
48. Valverde, F.R., Hurtado, J.V., Valada, A.: There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In: CVPR (2021) 2
49. Wagner, J., Fischer, V., Herman, M., Behnke, S.: Multispectral pedestrian detection using deep fusion convolutional neural networks. In: Proceedings of European Symposium on Artificial Neural Networks (2016) 2, 4
50. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019) 8, 9
51. Xu, D., Ouyang, W., Ricci, E., Wang, X., Sebe, N.: Learning cross-modal deep representations for robust pedestrian detection. In: CVPR (2017) 2, 4
52. Xu, P., Davoine, F., Denoeux, T.: Evidential combination of pedestrian detectors. In: British Machine Vision Conference. pp. 1–14 (2014) 3, 5, 6, 13
53. Zhang, H., Dana, K.: Multi-style generative network for real-time transfer. arXiv:1703.06953 (2017) 13
54. Zhang, H., Fromont, E., Lefèvre, S., Avignon, B.: Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In: IEEE International Conference on Image Processing (ICIP) (2020) 12, 13, 14
55. Zhang, H., Fromont, E., Lefèvre, S., Avignon, B.: Guided attentive feature fusion for multispectral pedestrian detection. In: WACV (2021) 11, 13, 14
56. Zhang, L., Liu, Z., Zhang, S., Yang, X., Qiao, H., Huang, K., Hussain, A.: Cross-modality interactive attention network for multispectral pedestrian detection. Information Fusion **50**, 20–29 (2019) 2, 4, 11
57. Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z., Liu, Z.: Weakly aligned cross-modal learning for multispectral pedestrian detection. In: ICCV (2019) 2, 4, 11
58. Zhou, K., Chen, L., Cao, X.: Improving multispectral pedestrian detection by addressing modality imbalance problems. In: ECCV (2020) 4, 11
59. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017) 13
60. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: ECCV. pp. 391–405. Springer (2014) 5