Multimodal Object Detection via Bayesian Fusion

Yi-Ting Chen\textsuperscript{1}, Jinghao Shi\textsuperscript{1}, Christoph Mertz, Shu Kong, Deva Ramanan\textsuperscript{*}

Carnegie Mellon University \textsuperscript{*}Argo AI

\{yitingc, jinghaos, cmertz, shuk, deva\}@andrew.cmu.edu

[Project Page] [Github]

Abstract

Object detection with multimodal inputs can improve many safety-critical perception 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 can provide much stronger object signatures under poor illumination. We explore strategies for fusing information from different modalities. Our key contribution is a non-learned late-fusion method that fuses together bounding box detections from different modalities via a simple probabilistic model derived from first principles. Our simple approach, which we call Bayesian Fusion, is readily derived from conditional independence assumptions across different modalities. We apply our approach to benchmarks containing both aligned (KAIST) and unaligned (FLIR) multimodal sensor data. Our Bayesian Fusion outperforms prior work by more than 13\% in relative performance.

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 [37, 19]. Such detectors are widely adopted in various safety-critical systems such as autonomous vehicles (AVs) [16, 3]. Motivated by AVs that operate in both day and night, we study multimodal object detection with RGB and thermal cameras, since the latter can provide much stronger object signatures under poor illumination [21, 43, 25, 8, 48, 1].

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), much 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 for spatial alignment [21] and a timer synchronizer for temporal alignment [34]. Fusion on unaligned RGB-thermal inputs, such as those shown in Fig. 3, 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 a single modality (e.g., FLIR [15]), further complicating multimodal learning.

First two authors contribute equally and are ordered alphabetically.
Multimodal fusion. Perhaps the central challenge in multimodal detection is how to fuse information from different modalities. Previous work has explored architectural strategies for fusing information at various stages [5, 43, 25, 47, 48, 1], which we loosely categorize into early-, mid- and late-fusion (Fig. 2). Early fusion constructs a four-channel RGB-thermal input image [40], which is then processed by a (typical) deep network. Mid-level fusion keeps RGB and thermal inputs in different streams that are merged together downstream within the network [40, 29, 24]. The vast majority of past work explores architectural choices of where and how to merge. We focus on the extreme scenario of very-late fusion of bounding box detections, which is conceptually and practically simple because it requires training and deploying independent, single-modal detectors. We demonstrate that this can offer tremendous performance advantages because one can learn from single modal datasets that often dwarf the size of multimodal datasets. However, very-late fusion of detectors trained on different modalities is difficult because detectors across modalities may not even fire on the same object, complicating fusion.

Bayesian Fusion. Our key contribution is a simple approach for very-late fusion of detectors derived from first principles given conditional independence assumptions [33]. Simply put, if modal signals are conditionally independent of each other given the true object label, one can probabilistically fuse together detections via Bayes rule. Importantly, this requires no learning, and so does not require any multimodal data for training. We call our approach Bayesian Fusion (Fig. 1). We demonstrate that Bayesian Fusion can be used to fuse correlated detectors that are technically not conditionally independent, fusing together outputs from other fusion methods. From this perspective, Bayesian Fusion can be seen as a general technique for ensembling detectors. We demonstrate significant improvements over prior art, both on aligned and unaligned multimodal benchmarks (KAIST [21] and FLIR [15]).

2. Related Work

Object Detection is a core problem in computer vision. State-of-the-art approaches [37, 35, 30] train deep detectors over large-scale datasets (e.g., COCO [27]). There are different types of deep detectors, e.g., anchor-based detection in pre-defined grids of spatial locations [30, 35, 36] and region proposal detection followed by classification [37]. In this work, we use Faster-RCNN detector [37] for exploring multimodal detection. Note that our explored approaches are general and applicable to other network architectures.

Multimodal Detection, specifically with RGB-thermal images, has attracted increasing attention. The KAIST pedestrian detection dataset [21] is one of the first benchmark for RGB-thermal detection, fostering growth of research in this area. Inspired by the successful RGB-based detectors [37, 35, 30], current multimodal detectors train deep models with various methods for fusing multimodal signals [5, 43, 25, 47, 48, 1]. Most of these multimodal detection methods work on well-aligned RGB-thermal images, but it is unclear how they perform on heavily unaligned modalities such as images in Fig. 3 taken from FLIR dataset [15]. We study multimodal detection under both aligned and unaligned RGB-thermal scenarios.

Multimodal fusion is the core problem in multimodal detection. We categorize different fusion methods into three types (Fig. 2): early-, mid- and late-fusion. We call methods early-fusion if they directly fuse RGB-thermal images, e.g., simply concatenating RGB-thermal as a four-channel inputs [40]. Mid-fusion operates on the features at some layers of a deep detector, oftentimes concatenating features.
computed by two backbones that process single-modal inputs respectively [40, 29, 24]. Late-fusion focuses on fusing detection results [17, 25, 48] computed by multiple (independent single-modal) detectors (Fig. 1). In this work, we explore all the three types of fusion methods with both aligned and unaligned RGB-thermal inputs.

3. Fusion Strategies for Multimodal Detection

We now present multimodal fusion for detection. We first note that single-modal detectors are viable methods for processing multimodal signals (Fig. 2a), so we include them as a baseline. We briefly introduce early- and mid-fusion methods which are widely studied in the literature, and then elaborate late-fusion.

**Early-fusion** simply concatenates RGB and thermal as a four-channel input (Fig. 2b). Despite its simplicity, prior work largely explore it on well-aligned RGB-thermal inputs (e.g., KAIST dataset [21]). It is unclear whether it still works on unaligned data (e.g., FLIR dataset [15]). We explore early-fusion on both aligned and unaligned RGB-thermal images, and conclude with a positive answer.

**Mid-fusion** fuses features inside a deep network (Fig. 2c). Comparing to early-fusion, we hypothesize that learned mid-fusion might better handle unaligned RGB-thermal inputs. Our rationale is that the large receptive field makes the network more resilient to misaligned modalities.

### 3.1. Late-Fusion

We now describe strategies for late-fusion of detectors from different modalities, beginning with a naive approach that motivates our final Bayesian Fusion (Fig. 1). Late fusion consists of score fusion and bounding box (bbox) fusion. We elaborate the latter at the end of this section.

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

**Non-Maximum Suppression (NMS).** The natural solution for dealing with overlapping detections is NMS, a crucial component in contemporary RGB detectors [9, 51, 20]. NMS finds bounding box predictions with high spatial overlap and remove the lower-scoring bounding box. 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 [2]. Surprisingly, though it has not been advocated as an approach to fusion of multi-modal detectors, we find it quite intuitive and effective. 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. Rather than selecting one modality at the global image level, 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 Score Fusion.** To actually fuse information across modalities, a straightforward strategy is to modify NMS to average the scores of detections from different modalities, rather than suppressing the weaker modality. Such an averaging has been proposed in prior work [29, 25]. Note that averaging scores will necessarily decrease the NMS score which reports the max of an overlapping set of detections (Fig. 1c). Our experiments suggest that averaging produces worse results than NMS or even
single modal detection. Intuitively, if two modalities agree that there exist strong detections, fusion should increase the overall confidence rather than decrease.

**Probabilistic Bayesian Fusion.** We now derive a probabilistic approach for fusing multimodal detections, consisting of fusion of classification scores and fusion of bounding boxes (Algorithm 1). We begin with the former. 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 $y$:

$$p(x_1, x_2 | y) = p(x_1 | y) p(x_2 | y)$$ (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. Note that such independence assumptions may be effective even when they hold only approximately [6, 33].

**Bayes rule.** We wish to infer labels given multimodal measurements:

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

$$\propto p(x_1, x_2 | y) p(y)$$ (3)

**Bayesian fusion.** By applying the conditional independence assumption from (1) to (3), we have:

$$p(y | x_1, x_2) \propto p(x_1 | y) p(x_2 | y) p(y)$$ (4)

$$\propto \frac{p(x_1 | y) p(y | x_2) p(y | x_2)}{p(y)}$$ (5)

$$\propto \frac{p(y | x_1) p(y | x_2)}{p(y)}$$ (6)

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

1. Train separate single-modal classifiers that predict the distributions over the label $y$ given 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 to sum-to-one (6).

**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 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 $x_i$:

$$p(y = k | x_i) = \frac{e^{s_i[k]}}{\sum_j e^{s_i[j]}} \propto e^{s_i[k]}$$ (7)

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. (6):

$$p(y = k | x_1, x_2) \propto \frac{p(y = k | x_1) p(y = k | x_2)}{p(y = k)}$$ (8)

$$\propto \frac{e^{s_1[k] + s_2[k]}}{p(y = k)}$$ (9)

Bayesian Fusion is thus equivalent to summing pre-logit scores, dividing by the class prior and normalizing via a softmax. Prior work in late-fusion for multimodal detection simply averages scores from predictions of separate detectors [29, 26]. Our derivation reveals that naive averaging may “double-count” the class prior; it is straightforward to show that double-counting is more severe when fusing across more modalities. The appendix explicitly computes such class priors from training examples, but empirically demonstrates that uniform priors (that naturally avoid double-counting) work surprisingly well even on imbalanced datasets like FLIR. We posit that its impact maybe more pronounced when fusing across many modalities.

**Missing modalities.** Importantly, summing versus averaging behaves profoundly differently when fusing across “missing” modalities. Intuitively, different single-modal detectors often do not fire on the same object. This means that in order to output a final set of detections above a confidence threshold (e.g., necessary for computing precision-recall / ROC metrics), one will need to compare scores from fused multi-modal detections with single modal detections. Bayesian Fusion elegantly deals with missing features/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 Bayesian Fusion to probabilistically fuse bounding box (bbox) coordinates of overlapping detections. To do so, we simply re-purpose the derivation from (6) for a continuous bbox label rather than a discrete one. Specifically, write $z$ 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 $p(z | x_i)$ that takes the form of a Gaussian with identity covariance, i.e., $p(z | x_i) = \mathcal{N}(\mu_i, I)$ where $\mu_i$ are box coordinates predicted from modality $i$. We also assume a
uniform prior on \( p(z) \), implying bbox coordinates can lie anywhere in the image plane. Doing so, we can write

\[
p(z|x_1, x_2) \propto p(z|x_1) p(z|x_2)
\]

\[
\propto \exp\left(-\frac{1}{2}\|z - \mu_1\|^2\right) \exp\left(-\frac{1}{2}\|z - \mu_2\|^2\right)
\]

\[
\propto \exp\left(-\frac{1}{2}(z^T z - 2\mu_1^T z + \mu_1^2 z + z^T z - 2\mu_2^T z)\right)
\]

\[
\propto \exp\left(-\|z - \mu\|^2\right), \quad \text{where} \quad \mu = 0.5(\mu_1 + \mu_2)
\]

The above shows that simply averaging bbox coordinates produces a fused bbox that doubles in precision (i.e., inverse covariance) [7]. In theory, one could modify detectors to return distributions over bounding box coordinates manifested as variance uncertainties [22], allowing one to compute a weighted average of bboxes by weighting each by its uncertainty. In the appendix, we explore a simple heuristic that approximates the certainty using the classification confidence, but saw only marginal improvement. As such, we follow the simple averaging approach outlined above.

4. Experiments

We explore different fusion strategies with extensive experiments on two datasets: KAIST [21] and FLIR [15] (cf. Fig. 3). We first introduce our implementation, then report the experimental results on each dataset (along with evaluation metrics) in separate subsections.

4.1. Implementation

We conduct experiments with PyTorch [32] and the Detectron2 toolbox [41] on a single GPU (Nvidia GTX 1080). Detectron2 has an excellent wrapper for Faster-RCNN, which is the detector architecture in our work. We use SGD optimizer with learning rate 5e-3. For data augmentation, we use random flip and random resizing. We pre-train a Faster-RCNN detector on COCO dataset [27]. As COCO only has RGB images, adapting the pre-trained detector to thermal inputs needs careful preprocessing of the thermal images (detailed below).

**Preprocessing.** 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 [41] where the mean values are computed over all the training images. Similarly, we also calculate the mean value (135.438) in the thermal training data. We find using a precise mean subtraction to process thermal images leads to better performance for when fine-tuning the pre-trained Faster-RCNN detector.

**Stage-wise Training.** We fine-tune the pre-trained Faster-RCNN to train our single-modal detectors and the early-fusion detectors (Fig. 2). 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. 2c). The late-fusion methods operate on the classification scores computed by the detectors. Note that all the late-fusion methods presented hereby are non-learned. We have also experimented with learning-based late-fusion methods (e.g., learning to fuse logits or calibrate scores), but find these methods to only perform marginally better than our non-learned Bayesian Fusion. Therefore, we focus on the non-learned late fusion methods in the main paper and study learning-based ones in the appendix.

**4.2. Multimodal Pedestrian Detection on KAIST**

**Dataset.** KAIST Multispectral Pedestrian Benchmark [21] is a popular multimodal dataset that exclusively benchmarks on pedestrian detection. RGB and thermal images are well-aligned using a beam-splitter, and have resolutions of 640x480 and 320x256, respectively. We resize thermal images to 640x480 during training (along with random-resizing augmentation). 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). As the original KAIST dataset contains noisy annotations, the literature introduces cleaned version of the train/test sets: a sanitized train-set (7,601 examples) [25] and a cleaned test-set (2,252 examples) [28]. Because of this, we find the literature has two evaluation protocols, 1) reporting on the old official test-set [21] and 2) reporting on the new cleaned test-set [28]. We report with both protocols. Moreover, we also follow the literature that evaluates under the “reasonable setting” [21], i.e., ignoring annotated persons that are occluded (tagged by KAIST) or too small (<55 pixels).

**Metric.** Log-Average Miss Rate (LAMR) is widely used in the pedestrian detection community [10]. This is also the
Tables 1 and 2: Ablation study on KAIST new test-set under the “reasonable” setting, measured by percent LAMR. Please see text for a detailed discussion, but overall, we find our proposed BayesFusion approach to outperform all other variants, including end-to-end learned approaches such as Early and MidFusion. Fig. 5 shows the corresponding MR-FPPI curves.

### Table 1: Ablation study on KAIST new test-set under the “reasonable” setting

| Method               | Day   | Night | All   |
|----------------------|-------|-------|-------|
| RGB                  | 15.16 | 32.47 | 21.01 |
| Thermal              | 25.65 | 8.34  | 20.95 |
| EarlyFusion          | 25.17 | 8.16  | 19.87 |
| MidFusion            | 17.59 | 11.29 | 15.45 |
| Pooling              | 37.92 | 22.61 | 32.68 |
| NMS                  | 14.47 | 7.57  | 11.84 |
| AvgScore             | 23.29 | 16.30 | 21.02 |
| BayesFusion          | 11.01 | 6.08  | 9.16  |
| BayesFusion+bbox     | 10.49 | 5.60  | 8.65  |
| BayesFusion++bbox    | 9.59  | 5.34  | 8.02  |

### Table 2: Quantitative comparison on KAIST measured by LAMR, in percentage, on the two KAIST test-sets (old and new).

| Testset | Method         | Day   | Night | All   |
|---------|----------------|-------|-------|-------|
| old     | KAIST baseline | 64.17 | 63.99 | 64.76 |
|         | CMT-CNN [33]   | 47.30 | 54.78 | 49.55 |
|         | FeatureFusion [40] | 46.15 | 37.00 | 43.80 |
|         | BayesFusion++ (ours) | 32.90 | 20.89 | 29.35 |
| new     | HalfwayFusion [29] | 36.84 | 35.49 | 36.99 |
|         | RPN+BDT [24]   | 30.51 | 27.62 | 29.83 |
|         | TC-DET [1]     | 34.81 | 10.31 | 27.11 |
|         | IATDNN [17]    | 27.29 | 24.41 | 26.37 |
|         | IAF R-CNN [26] | 21.85 | 18.96 | 20.95 |
|         | CIAN [47]      | 14.77 | 11.13 | 14.12 |
|         | MSDS-RCNN [25] | 12.22 | 7.82  | 10.89 |
|         | AR-CNN [48]    | 9.94  | 8.38  | 9.34  |
|         | BayesFusion++ (ours) | 9.59  | 5.34  | 8.02  |

**Qualitative Results.** We qualitatively compare our best-performing model (BayesFusion++bboxFusion) and our mid-fusion method in Fig. 4. Visually, our Bayesian Fusion detects all persons, while the mid-fusion method has multiple false negatives, i.e., mis-detections.

### 4.2.1 Ablation Study on KAIST

We ablate all explored fusion methods on the new KAIST test-set in Table 1 and Fig. 5. Single modal detectors tend to work well in different environments, with RGB detectors working on well-lit day images while thermal works 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 ground-truth object, which are counted as false positives. Interestingly, simple NMS 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 (AvgScore) 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. BayesFusion accomplishes this via probabilistic integration of information from the RGB and Thermal single-modal detectors. We also find that it can be further improved by averaging bounding boxes of overlapping detections. Lastly, we find BayesFusion++ that fuses detections from more models (i.e., RGB, thermal and MidFusion), performs the best.

### 4.2.2 Quantitative Comparison on KAIST

We now compare our best-performing Bayesian Fusion model (i.e., BayesFusion++bboxFusion in Table 1) to the prior methods.

**Compared Methods.** Recall that there are two KAIST test-sets, the old (official) KAIST test-set and the new one [28]). On the old, [40] reports the best (till now) performance using a mid-fusion method (FeatureFusion). Over the new, two top-performing methods are MSDS-RCNN [25] and AR-CNN [48]. MSDS-RCNN leverages more supervision (through additional annotations on segmentation masks) through a learnable late-fusion method [25]. AR-CNN mainly focuses on weakly-unaligned RGB-thermal pairs, yet achieves the state-of-the-art on the new test-set. We refer readers to more compared methods listed in Table 2.

**Results.** Table 2 compares our Bayesian Fusion method with the prior work. Our method clearly outperforms all prior methods by a large margin. On the old test-set, Bayesian Fusion dramatically reduces detection error from 43.80 LAMR [40] to 29.35 LAMR. On the new test-set, our method also performs much better than the prior work [25], reducing error by 14% (from 9.34 to 8.02 LAMR). We reiterate that our Bayesian Fusion is a non-learned, late-fusion method that simply fuses detections from single-modal de-
4.3. Multimodal Object Detection on FLIR

Dataset. FLIR ADAS dataset [15] is a recent RGB-thermal dataset consisting 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 RGB and thermal images to resolution 640x512. FLIR dataset consists of 10,228 RGB-thermal image pairs in total, split into a train-set (8,862 images) and a validation set (1,306 images). Note that the RGB and thermal images in FLIR are strongly unaligned (cf. Fig. 3 and 6). As a result, FLIR only annotates thermal images. In literature [4, 23, 46, 31, 8], FLIR evaluates on three classes which have imbalanced examples: 28,151 persons, 46,692 cars, and 4,457 bicycles. Moreover, there are a few thermal images that do not have corresponding RGB images. To explore fusion methods, we follow [46] that removes these unpaired images, resulting in the final validation set of 1,258 images for benchmarking.

To analyze performance on day/night scenes, we manually annotate them for this analysis. Clearly, incorporating RGB by our learning-based fusion methods notably improves performance on both day and night scenes. We explore late-fusion with detection outputs from our three models: Thermal, Early and Mid. We find all AvgScore, NMS and BayesFusion lead to better performance than the learning-based MidFusion model. Especially, BayesFusion performs the best; using bounding box fusion (bbox) improves further.

| Method     | Day     | Night   | All      |
|------------|---------|---------|----------|
| Thermal    | 75.35   | 82.90   | 78.05    |
| EarlyFusion| 77.37   | 79.56   | 78.65    |
| MidFusion  | 79.37   | 81.64   | 80.22    |
| Pooling    | 57.02   | 59.47   | 57.97    |
| NMS        | 82.22   | 83.94   | 83.08    |
| AvgScore   | 81.49   | 85.02   | 82.97    |
| BayesFusion| 82.51   | 85.40   | 83.81    |
| BayesFusion+bbox | **83.04** | **86.40** | **84.35** |

Metric. Average Precision (AP) is a common metric in object detection [12, 38]. 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 the literature of multimodal detection that benchmarks on the FLIR dataset [8, 31, 46, 23, 4], we define a true positive as a detection that overlaps a ground-truth with IoU>0.5. Note that AP used in the multimodal detection literature is different from mAP [27], which computes the area under the interpolated precision-recall curve.

4.3.1 Ablation study on FLIR

We compare our fusion methods in Table 3, along with qualitative results in Fig. 6. We break down analysis according to our annotated day/night tags. Compared to the single-modal detector (Thermal), our learning-based methods early-fusion (EarlyFusion) and mid-fusion (MidFusion)
lead to improved performance, while the latter works better. As hypothesized, MidFusion outperforms EarlyFusion, justifying that end-to-end learning of fusing features better handles mis-alignment between RGB and thermal images. When applying late-fusion methods (AvgScore, NMS and BayesFusion) to detections of Thermal, EarlyFusion and MidFusion models, we improve detection performance further. To note, in fusion of multiple detectors, AvgScore has a flavor of ensemble learning which averages results from multiple models. But our Bayesian Fusion method shows better results, implying that it should be potentially a better choice in fusing results of ensemble learning.

### 4.3.2 Quantitative Comparison on FLIR

**Compared Methods.** We compare our fusion methods with prior work: ThermalDet [4], BU [23], CFR [46], ODSC [31] and MMTOD [8]. As FLIR does not have aligned RGB-thermal images and only annotates thermal images, many methods adopt a domain adaptation idea that adapts a pre-trained detector (over large-scale RGB images like COCO [27]) to thermal input. For example, MMTOD [8] and ODSC [31] adopt the image-to-image-translation technique [49, 45] to generate RGB images from thermal images, hypothesizing that this helps train a better multimodal detector by finetuning a detector that is pretrained over large-scale RGB images. BU [23] operates such a translation/adaptation on features that generates thermal features to be similar to RGB features. ThermalDet [4] exclusively exploits thermal images and ignores RGB images; it proposes to combine features from multiple layers for the final detection. Surprisingly, to the best of our knowledge, there is no prior work that directly trains on the heavily unaligned RGB-thermal image pairs (like in FLIR) for multimodal detection. In our exploration, we find directly training an end-to-end model with simple early-fusion and mid-fusion strategies leads to much better performance than prior work.

As shown in Table 4, all of our methods outperform the previous approaches. Our simple single-modal detector (trained only on thermal images) achieves slightly better performance than ThermalDet [4], which also exclusively trains on thermal images, probably owing to better augmentation techniques and 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 better. This confirms our hypothesis that fusing features better handles mis-alignment of RGB-thermal images than the early-fusion method. Our Bayesian Fusion method (with bounding box fusion) performs the best, significantly outperforming all the compared methods! Notably, our fusion methods achieve a huge boost in “bicycle”. We conjecture that bicycles do not emit heat to make them stand out in thermal, but are more visible in

| Method          | Bicycle | Person | Car   | All   |
|-----------------|---------|--------|-------|-------|
| MMTOD-CG [8]    | 50.26   | 63.31  | 70.63 | 61.40 |
| MMTOD-UNIT [8]  | 49.43   | 64.47  | 70.72 | 61.54 |
| ODSC [31]       | 55.53   | 71.01  | 82.33 | 69.62 |
| CFR3 [46]       | 57.77   | 74.49  | 84.91 | 72.39 |
| BU(AT,T) [23]   | 56.10   | 76.10  | 87.00 | 73.10 |
| BU(LT,T) [23]   | 57.40   | 75.60  | 86.50 | 73.20 |
| ThermalDet [4]  | 60.04   | 78.24  | 85.52 | 74.60 |
| Thermal (ours)  | 62.70   | 84.01  | 87.44 | 78.05 |
| EarlyFusion (ours) | 62.60 | 84.15  | 88.20 | 78.65 |
| MidFusion (ours) | 69.58   | 84.17  | 86.91 | 80.22 |
| BayesFusion (ours) | **74.61** | **88.22** | **90.21** | **84.35** |

Table 4: Quantitative comparison on FLIR measured by AP↑ in percentage with IoU>0.5. Following the literature, we evaluate on the three categories annotated by FLIR. Perhaps surprisingly, end-to-end training on thermal images already outperforms all the prior methods, presumably because of better augmentations and a better pre-trained model (Faster-RCNN). Moreover, our fusion methods perform even better. Lastly, our Bayesian Fusion method performs the best. These results are comparable to Table 3.

Figure 6: Qualitative multimodal detection results on FLIR images. We show three examples (in columns) with RGB (top) thermal images (middle and bottom). We overlay the ground-truth annotations on the RGB, highlighting that RGB and thermal images are strongly unaligned. To avoid clutter, we do not mark class labels for the bounding boxes. On the thermal images, we show qualitative results from our thermal-only (mid-row) and best-performing BayesFusion (with bounding box fusion) model (bottom-row). Green, red and blue boxes stand for true positives, false negative (mis-detected persons) and false positives. Particularly from the third column, thermal-only model has many false negatives (or mis-detections), which are “bicycles”. Understandably, thermal images will not deliver strong signatures for bicycles, but RGB images do. This explains why our fusion model performs better in detecting bicycles.
RGB, fusing which leads to better detection of bicycles.

5. Conclusion

In this work, we explore different fusion strategies for multimodal detection under both aligned and unaligned RGB-thermal images. We demonstrate that non-learned Bayesian Fusion, when integrated in a NMS post-processing loop, significantly outperforms prior learning-based approaches. Key reasons for its strong performance are that (a) it can take advantage of highly-tuned single modal detectors trained on large-scale single-modal datasets and (b) 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 (though not as effectively). Bayesian Fusion yields more than 13% relative improvement over prior work on two benchmark datasets.

Acknowledgement

This work was supported by the CMU Argo AI Center for Autonomous Vehicle Research. Authors acknowledge helpful discussions and valuable feedbacks from Jessica Lee, Peiyun Hu, Jianren Wang, David Held, Kangle Deng, and Michel Laverne.

References

[1] Marco Bertini and Alberto del Bimbo. Task-conditioned domain adaptation for pedestrian detection in thermal imagery. In ECCV, 2020. 1, 2, 3
[2] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In ICCV, 2019. 3
[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giangarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11621–11631, 2020. 1
[4] Yu Cao, Tong Zhou, Xinhua Zhu, and Yan Su. Every feature counts: An improved one-stage detector in thermal imagery. In IEEE International Conference on Computer and Communications (ICCC), 2019. 7, 8
[5] Hangil Choi, Seungryong Kim, Kihong Park, and Kwanghoon Sohn. Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks. In International Conference on Pattern Recognition (ICPR), 2016. 2
[6] A Philip Dawid. Conditional independence in statistical theory. Journal of the Royal Statistical Society: Series B (Methodological), 41(1):1–15, 1979. 4
[7] Morris H DeGroot. Optimal statistical decisions, volume 82. John Wiley & Sons, 2005. 5, 14
[8] Chaitanya Deveguptapu, Ninad Akolekar, Manuj M Sharma, and Vineeth N Balasubramanian. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 0–0, 2019. 1, 7, 8
[9] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 304–311. IEEE, 2009. 3
[10] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. IEEE transactions on pattern analysis and machine intelligence, 34(4):743–761, 2011. 5, 6
[11] Wushi Dong. commit 0a3e722. https://github.com/wushidonguc/two-stream-action-recognition-keras/blob/master/fuse_validate_model.py. 11
[12] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. International journal of computer vision, 111(1):98–136, 2015. 7
[13] Christoph Feichtenhofer. commit 3e313e4. https://github.com/fcichtenhofer/twostreamfusion/blob/master/cnn_uci101_fusion.m. 11
[14] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016. 11
[15] FLIR. Pifir thermal dataset for algorithm training. https://www.flir.io/om/adas/adas-dataset-form, 2018. 1, 2, 3, 5, 7
[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In Conference on Computer Vision and Pattern Recognition (CVPR), 2012. 1
[17] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. Information Fusion, 50:148–157, 2019. 3, 6
[18] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. arXiv preprint arXiv:1706.04599, 2017. 12, 13
[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In ICCV, 2017. 1
[20] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. Learning non-maximum suppression. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4507–4515, 2017. 3
[21] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1037–1045, 2015. 1, 2, 3, 5, 6
[22] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? arXiv preprint arXiv:1703.04977, 2017. 5, 14
[23] My Kiew, Andrew D Bagdanov, and Marco Bertini. Bottom-up and layer-wise domain adaptation for pedestrian detection in thermal images. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2020. 7, 8

[24] Daniel Konig, Michael Adam, Christian Jarvers, Georg Lauer, Heiko Neumann, and Michael Teutsch. Fully convolutional region proposal networks for multispectral person detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 49–56, 2017. 2, 3, 6

[25] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Multispectral pedestrian detection via simultaneous detection and segmentation. *arXiv preprint arXiv:1808.04818*, 2018. 1, 2, 3, 5, 6

[26] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019. 1, 4, 6

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 5, 7, 8

[28] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris Metaxas. Improved annotations of test set of kaist, 2018. 5, 6

[29] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N Metaxas. Multispectral deep neural networks for pedestrian detection. *BMVC*, 2016. 1, 2, 3, 4, 6

[30] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2

[31] Farzeen Munir, Shoaib Azam, Muhammad Asim Rafique, Ahmad Muqeem Sheri, and Moongu Jeon. Thermal object detection using domain adaptation through style consistency. *arXiv preprint arXiv:2006.00821*, 2020. 7, 8

[32] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[33] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014. 2, 4

[34] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5. Kobe, Japan, 2009. 1

[35] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[36] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017. 2

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1, 2

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 7

[39] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, pages 568–576, 2014. 11, 13

[40] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *Proceedings of European Symposium on Artificial Neural Networks*, 2016. 2, 3, 6

[41] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2: https://github.com/facebookresearch/detectron2, 2019. 5

[42] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *Proceedings of the ACM international conference on Multimedia*, 2015. 11

[43] Dan Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, and Nicu Sebe. Learning cross-modal deep representations for robust pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5363–5371, 2017. 1, 2, 6

[44] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015. 11

[45] Hang Zhang and Kristin Dana. Multi-style generative network for real-time transfer. *arXiv preprint arXiv:1703.06953*, 2017. 8

[46] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *IEEE International Conference on Image Processing (ICIP)*, 2020. 7, 8

[47] Lu Zhang, Zhiyong Liu, Shifeng Zhang, Xu Yang, Hong Qiao, Kaizhu Huang, and Amir Hussain. Cross-modality interactive attention network for multispectral pedestrian detection. *Information Fusion*, 50:20–29, 2019. 2, 6

[48] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *ICCV*, 2019. 1, 2, 3, 6

[49] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 8

[50] Yi Zhu. commit 32b6354. https://github.com/bryanzhu/two-stream-pytorch/blob/master/scripts/eval_ucf101.pytorch/ temporal_demo.py, 2011

[51] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014. 3
Appendix

The appendix provides additional studies. First, we analyze Bayesian Fusion by comparing with other commonly used late fusion methods such as average score fusion or logit fusion. Then, we study learning-based late fusion versus the non-learned Bayesian Fusion, using class priors computed by class frequency versus uniform prior, weighted box fusion versus average box fusion, and fusing more detectors. Lastly, we show more qualitative results.

6. Bayesian Fusion for Logits

In this section, we compare Bayesian Fusion to additional late fusion approaches in the literature that extends beyond detection. Because classic fusion approaches [39, 44, 14] often operate on logit scores that are input into a softmax (rather than operating on the output of a softmax), we re-examine Bayesian Fusion in terms of logit scores.

Let us rewrite the single-modal softmax posterior for class $k$ given modality $i$ in terms of single-modal logit scores $s_i[k]$. For notational simplicity, we suppress its dependence on the underlying input modality $x_i$:

$$p(y = k|x_i) = \frac{e^{s_i[k]}}{\sum_j e^{s_j[k]}} \propto e^{s_i[k]} \quad (14)$$

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. 6 from the main paper:

$$p(y = k|x_1, x_2) \propto \frac{p(y = k|x_1)p(y = k|x_2)}{p(y = k)} \propto \frac{e^{s_1[k]+s_2[k]}}{p(y = k)} \quad (15)$$

If we assume a uniform prior over classes, Bayesian posteriors are proportional to $e^{s[k]}$ where $s[k] = s_1[k] + s_2[k]$ are the summed per-modality logits. **Hence Bayesian Fusion corresponds to adding logits from each modality.** This suggests another practical implementation of Bayesian Fusion that may improve numerical stability: given single-modal detections with cached logit scores, sum logit scores on overlapping detections before pushing them through a softmax.

**Summing vs. averaging logits.** Let us now revisit prior approaches to logit-based fusion in detail. Late fusion was popularized by video classification networks that made use of two-stream architectures [39]. This seminal work proposed an influential baseline for “fusing softmax scores” by averaging. However, practical implementations average logits [50, 11] or sum logits [13], often omitting the final softmax [42] because one can obtain a class prediction by simple maximization of the fused logits. In the classification setting, the distinction between summing versus averaging does not matter because both produce the same argmax label prediction. **But the distinction does matter in detection, which requires ranking and comparison of scores for non-maximal suppression (NMS) and global thresholding.** Intuitively, summing allows detections to become more confident as more modalities agree, while averaging does not. Most crucially, summing logits allows one to optimally compare detections with missing modalities, which is frequently needed in NMS whenever all modalities fail to fire on a given object. Here, optimality holds in the Bayesian sense whenever modalities are conditionally independent (as derived in (17)).

**Fusion from logits.** We can succinctly compare various fusion approaches from the logit perspective with the following:

$$s_{AvgLogit}[k] = .5(s_1[k] + s_2[k]) \quad (18)$$

$$s_{Bayes}[k] = s_1[k] + s_2[k] \quad (19)$$

It is easy to see that

$$s_{AvgLogit}[k] \leq s_{Bayes}[k]$$

Table 5: **Additional late fusion baselines** measured by LAMR↓ on KAIST reasonable-test. Numbers are identical to Table 1 from the main paper with an additional row for logit averaging (AvgLogits), which outperforms class-posterior averaging (AvgScore). However, both methods still underperform a simple NMS. Eq.(17) derives that BayesFusion is equivalent to summing logits instead of averaging. Intuitively, summing allows fusion to become more confident as more modalities agree, while averaging does not. Even more importantly, this small modification allows one to properly compare detections with missing modalities, which is frequently needed in NMS whenever all modalities fail to fire on a given object. Finally, we also explore a learned late fusion baseline that learns to combine logits with logistic regression (LogRegFusion), which provides a marginal improvement over Bayesian Fusion at the cost of training on a carefully curated multimodal dataset. Our analysis shows that learned fusion can be seen as a generalization of Bayesian Fusion that no longer assumes conditionally independent modalities [21].

| Method        | Day | Night | All  |
|---------------|-----|-------|------|
| RGB           | 15.16 | 32.47 | 21.01 |
| Thermal       | 25.65 | 8.34  | 20.95 |
| Pooling       | 37.92 | 22.61 | 32.68 |
| NMS           | 14.47 | 7.57  | 11.84 |
| AvgScore      | 23.29 | 16.30 | 21.02 |
| AvgLogits     | 18.78 | 11.70 | 16.28 |
| LogRegFusion  | 10.70 | 6.11  | 9.08  |
| BayesFusion   | 11.01 | 6.08  | 9.16  |
| BayesFusion+bbox | 10.49 | 5.60  | 8.65  |
cause the other class logits are needed to compute the softmax partition function. One particularly simple case to analyze is a single-class detector $k \in \{0, 1\}$, as is true for the KAIST benchmark (that evaluates only pedestrians). Here we can analytically compute posteriors by looking at the relative logit score $s_i = s_i[1] - s_i[0]$ for modality $i$ (by relying on the well-known fact that a 2-class softmax function reduces to a sigmoid function of the relative input scores). We visualize the fused probability as a function of the relative per-modality logits $s_1$ and $s_2$ in Fig. 7. Finally, Table 5 explicitly compares the performance of such fusion approaches with other diagnostic variants. We refer the reader to both captions for more analysis.

7. Score Calibration before Fusion

Bayesian Fusion assumes that detectors return true class posteriors. However, deep networks are notoriously overconfident in their predictions, even when wrong [18]. One popular calibration strategy is adding a temperature parameter $T$ to the final softmax, typically to “soften” overconfident predictions. Figure 8: LAMR as a function of a calibration temperature parameter $T$ (designed to return more realistic probabilities) [18] on KAIST reasonable-test. We fuse detections from two single-modal detectors (RGB and thermal). Here, $T=1$ corresponds to Bayesian Fusion. Tuning the temperature $T$ yields only marginally better performance. We conjecture that the scores from the two single-modal detectors are already comparable, presumably because both of them are trained with the same loss function, annotation labels, and network architecture.

Figure 7: Fusing logits from two single-modal, single-class detectors. Given a single class detector $k \in \{0, 1\}$, the single-modal class posterior for modality $i$ depends on the relative logit $s_i = s_i[1] - s_i[0]$. We visualize the probability surface obtained by different fusion strategies that operate on logit scores $s_1$ and $s_2$ (associated with two overlapping detections). We first point out that simply returning the maximum score, corresponding to non-maximal suppression (NMS), is a surprisingly effective late fusion strategy that already outperforms much prior work (see Table 1 from main paper and Table 5 from appendix). AvgLogits (c) and BayesFusion (d) have similar score landscapes, but differ in a scaling parameter. Our empirical results show that this scaling parameter has a large effect in multimodal detection, because one needs to compare multi-modal detections with single-modal detections with “missing data modalities”. By overlaying the score landscapes of NMS and AvgScore (e), one can see that AvgScore is always less than NMS. Similarly, by overlaying the score landscapes of BayesFusion and NMS (f), we find that BayesFusion returns (1) a higher probability than NMS when both modalities have large logits (e.g., $s_1=4$ and $s_2=4$) but (2) a lower probability than NMS when logit scores disagree (e.g., $s_1 = 3$ and $s_2 = -3$, corresponding to $p(y = 1|x_1) = 0.95$ and $p(y = 1|x_2) = 0.05$). In the latter case, NMS outputs an over-confident score 0.95.
Bayesian Fusion consistently outperforms all other late-fusion methods. Interestingly, fusing detections from non-independent detectors (e.g., \(A+B+D\)) achieves better performance than independent detectors (e.g., \(A+B\)).

| Method            | \(A+B\) | \(A+C\) | \(A+D\) | \(B+C\) | \(B+D\) | \(C+D\) | \(A+B+C\) | \(A+B+D\) |
|-------------------|---------|---------|---------|---------|---------|---------|-----------|-----------|
| Pooling           | 32.68   | 28.87   | 29.70   | 36.68   | 36.36   | 23.24   | 43.04     | 43.56     | 46.03     |
| AvgScore          | 21.02   | 19.37   | 18.94   | 22.01   | 18.80   | 20.91   | 22.69     | 21.83     | 23.92     |
| NMS               | 11.84   | 10.59   | 13.51   | 19.16   | 14.26   | 14.62   | 11.66     | 12.46     | 12.44     |
| BayesFusion       | 9.16    | 8.63    | 12.00   | 16.99   | 12.83   | 12.92   | 8.79      | 9.65      | 9.41      |
| BayesFusion+bbox  | 8.65    | 9.18    | 10.79   | 16.43   | 11.86   | 12.28   | 8.33      | 8.02      | 8.11      |

Table 6: Late-fusion methods on different underlying detectors on FLIR dataset, measured by percent AP\(\downarrow\) in percentage. A: thermal detector; B: EarlyFusion detector; C: MidFusion detector. Our Bayesian Fusion method consistently outperforms other late-fusion methods. By fusing all the underlying detectors, Bayesian Fusion performs the best.

Table 7: Late-fusion methods on different underlying detectors on FLIR dataset, measured by percent AP\(\downarrow\) in percentage. A: RGB detector; B: Thermal detector; C: EarlyFusion detector; D: MidFusion detector. Clearly, BayesFusion+bbox performs the best.

| Method            | \(A+B\) | \(A+C\) | \(B+C\) | \(A+B+C\) |
|-------------------|---------|---------|---------|-----------|
| Pooling           | 68.1    | 71.5    | 62.9    | 59.5      |
| AvgScore          | 80.1    | 80.7    | 82.3    | 82.8      |
| NMS               | 80.8    | 81.4    | 82.7    | 83.0      |
| BayesFusion       | 80.9    | 81.7    | 82.8    | 83.8      |
| BayesFusion+bbox  | 81.7    | 82.2    | 83.4    | 83.9      |

In the two-modality detection setting, because monotonic transformations of probability scores will not affect ranks (and hence not effect LAMR or AP), one can show that we need only calibrate one of two modalities. In practice, we calibrate thermal detector scores so as to better match scores from the RGB detector. Figure 8 plots LAMR as a function of a single scalar temperature \(T\) used to scale thermal detections. Tuning \(T\) yields only a marginal improvement over standard Bayesian Fusion (i.e., when \(T = 1\)). We conjecture that the two single-modal detectors are trained with the same annotation and network architecture, making their output scores comparable to each other already.

8. Further Study of Weighted Score Fusion

All late fusion approaches discussed thus far do not require training on multimodal data. Because prior work on late fusion has also explored learned variants, we also consider (learned) linear combinations of single-modal logits:

\[
s_{\text{Learned}}[k] = w_1[k]s_1[k] + w_2[k]s_2[k] \quad (21)
\]

One can view Bayesian Fusion, AvgLogits, and Temperature Scaling as special cases of the above. Bayesian Fusion and AvgLogits use predefined weights that do not require learning and so are easy to implement. Temperature scaling requires single-modal validation data to tune each temperature parameter, but does not require multimodal learning. This can be advantageous in settings where modalities do not align (e.g., FLIR) or where there exists larger collections of single-modal training data (e.g., COCO training data for RGB detectors). Truly joint learning of weights requires multimodal training data, but joint learning may better deal with correlated modalities by downweighting the contribution of modalities that are highly correlated (and don’t provide independant sources of information). We experimented with joint learning of the weights with logistic regression. To do so, we assembled training examples of overlapping single-modal detections (and cached logit scores) encountered during NMS, assigning a binary target label (corresponding to true vs false positive detection). After training on such data, we observe a small improvement over non-learned fusion (Table 5), consistent with prior art on late fusion [39]. We also tested learning-based late fusion methods on the FLIR dataset. We further tested learning class priors. However, these methods do not yield better performance than the simple non-learned Bayesian Fusion (both achieve 82.91 AP). The reason is that FLIR annotations are inconsistent across frames, making it hard for learning-based late fusion methods to shine, as explained in Fig. 9 and 10.

9. Further Study of Class Prior in BayesFusion

In the main paper, we assume uniform class priors when using Bayesian Fusion. Now we test Bayesian Fusion with computed class priors. For consistent experiments as done in the main paper, we use FLIR dataset and fuse three models (Thermal, Early, and Mid). Recall that FLIR has imbalanced classes: person (21,744), bicycle (3,806), and car (39,372). First, we count the number of annotated objects of each of the three class, and assign the fourth background class with a dummy number. Then, we normalize them to be sum-to-one as class priors. We vary the background prior and evaluate the final detection performance measured by AP at IoU>0.5, as shown in Fig. 11. Clearly, Bayesian Fusion works better with uniform priors than the
We demonstrate inconsistent annotations in FLIR dataset with four consecutive frames in the validation set. **Top-row** lists four RGB frames for reference. **Mid-row** displays thermal images and the ground-truth annotations. Looking at the annotations in the orange rectangle, we can see that the annotations are not consistent across frames. This is a critical issue that prevent learning-based late fusion from improving further on the FLIR dataset. **Bottom-row** displays the detection results by Bayesian Fusion of the three models (Thermal, Early, and Mid). Interestingly, the predictions look more reasonable in detecting pedestrians within the orange rectangles. In this sense, predictions is “better” than annotations, intuitively explaining why learning based late fusion does not improve performance further. Please also refer to Fig. 10 for a zoom-in visualization.

![Figure 9](image)

**Figure 9:** We demonstrate inconsistent annotations in FLIR dataset with four consecutive frames in the validation set. **Top-row** lists four RGB frames for reference. **Mid-row** displays thermal images and the ground-truth annotations. **Bottom-row** displays the detection results by Bayesian Fusion of the three models (Thermal, Early, and Mid). Interestingly, the predictions look more reasonable in detecting pedestrians within the orange rectangles. In this sense, predictions is “better” than annotations, intuitively explaining why learning based late fusion does not improve performance further. Please also refer to Fig. 10 for a zoom-in visualization.

![Figure 10](image)

**Figure 10:** We zoom in a frame from Fig. 9 to visualize more clearly that the ground-truth annotations can even miss bicycles and persons as shown in the third image. In contrast, our Bayesian Fusion model can detect these miss-labeled objects (cf. red arrows). This shows the issues in the FLIR dataset.

computed the class priors.

Furthermore, we ablate which class is more important by manually assigning a prior. Concretely, we vary one class prior by fixing all the others the same. We plot the performance vs. the per-class prior in Fig. 12. We can see tuning specific class priors leads to marginal improvements compared to using uniform prior.

**10. Further Study of Box Fusion**

In the main paper, to fuse multiple bounding boxes, we simply average their coordinates. By assuming identity covariance (Eq. 13 in the paper), this simple average fusion produces a fused bbox that doubles in precision (i.e., inverse covariance) [7]. In theory, one could modify detectors to return distributions over bounding box coordinates manifested as variance uncertainties [22], allowing one to compute a weighted average of bboxes by weighting each by its certainty. Hereby, rather than learning such an uncertainty-aware regression model, we explore a simple heuristic that approximates the certainty using the classification confidence, but saw only marginal improvement as shown in Table 8.

**11. Fusing More Models**

We study late fusion methods on more combinations of underlying detectors. Table 6 and 7 list results on KAIST and FLIR datasets, respectively. Importantly, BayesFusion consistently performs the best on each of combinations. Interestingly, applying Bayesian Fusion method to detectors that are not independent to each other (e.g., Thermal and
Figure 11: A study of Bayesian Fusion with class priors as class frequencies in the training set. We use FLIR dataset for this study as it has 3 imbalanced classes. We fuse three models (Thermal, Early and Mid) as used in the main paper. As there is a background class, we vary the background class and proportionally change the class priors. Clearly, Bayesian Fusion with uniform class priors performs better than using the computed priors. Tuning the background prior does not notably affect the final detection performance once this prior is set to be larger than 0.1.

Figure 12: A study of tuning a single class prior while keeping others the same. Motivated by the superior performance of Bayesian Fusion with uniform priors, we tune each of the class prior by fixing others the same. We study this on the FLIR dataset by fusing three models (Thermal, Early and Mid). We can see that tuning specific classes only marginally improves detection performance.

Table 8: Study of weighted box fusion. We explore a simple heuristic that approximates the box regression certainty using the classification confidence. We compare this weighted average box fusion with the simple average box fusion across models and datasets. Clearly, the weighted average fusion yields only marginal improvement over the simple average box fusion.

| model          | average | weighted average |
|----------------|---------|------------------|
| KAIST (LAMR↓)  | 8.65    | 8.71             |
| RGB + T        | 8.02    | 7.95             |
| RGB + T + Mid  | 81.67   | 81.71            |
| FLIR (AP↑)     | 83.94   | 83.94            |

Figure 13: We attach a demo video video_demo.mp4 as a part of the appendix. The demo video is generated based on a testing video (captured at night) provided by the FLIR dataset. Hereby we display two video frames for a same scene that compare detections by a thermal-only single-modal detector and the BayesFusion method that fuses three detectors (Thermal, Early-fusion and Mid-fusion). We can see Thermal detector mis-detects a car and produces larger bounding box for the rightmost car (right frame), in contrast, BayesFusion successfully detects all the cars and produces tight bounding boxes. We refer the reader to the video demo for convincing visualization.

12. Qualitative Results

We attach a demo video (https://youtu.be/vRJTlpGvTs) on a testing video (captured at night) provided by the FLIR dataset. In the demo video, we compare the detection results by the Thermal model and BayesFusion that fuses results of three models (Thermal + Early + Mid). Recall that the FLIR dataset does not align RGB and thermal frames, and annotates only thermal frames. Therefore, we only provide RGB frames as reference (cf. Fig. 13).

Lastly, we provide more qualitative results in Figure 14 and 15 for KAIST and FLIR, respectively. Visually, we can see our Bayesian Fusion method performs better than the compared methods.
Figure 14: Qualitative results on more testing examples in KAIST dataset. We place RGB-thermal images in pairs: in each macro row, we show RGB images in the upper row and thermal images in lower row. Over RGB images, we overlay the detection results from our MidFusion model; on the thermal images, we show results from our best-performing Bayesian Fusion model. Green, red and blue boxes stand for true positives, false negative (miss-detected persons) and false positives.
Figure 15: Qualitative results on more testing examples in FLIR dataset. We place RGB-thermal images in triplet: in each macro row (divided by the black line), we show RGB images in the upper row and thermal images in two lower rows. Over RGB images, we overlay ground-truth annotations, highlighting that RGB and thermal images are strongly unaligned. To avoid clutter, we do not mark class labels for the bounding boxes. On the thermal images, we show detection results from our thermal-only (mid-row) and best-performing Bayesian Fusion (with bounding box fusion) model (bottom-row). Green, red and blue boxes stand for true positives, false negative (mis-detected persons) and false positives.