Multimodal Object Detection via Probabilistic Ensembling *(Supplementary Material)*

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

Outline

The supplemental document provides additional studies about the proposed probabilistic ensembling technique (ProbEn). Below is a sketch of document and we refer the reader to each of these sections for details.

1 Probabilistic Fusion for Logits

In this section, we compare ProbEn to additional late fusion approaches in the literature that extends beyond detection. Because classic fusion approaches [5,7,3] often operate on logit scores that are input into a softmax (rather than operating on the output of a softmax), we re-examine ProbEn 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

*Equal contribution. The work was mostly done when authors were with CMU.
†Equal supervision.

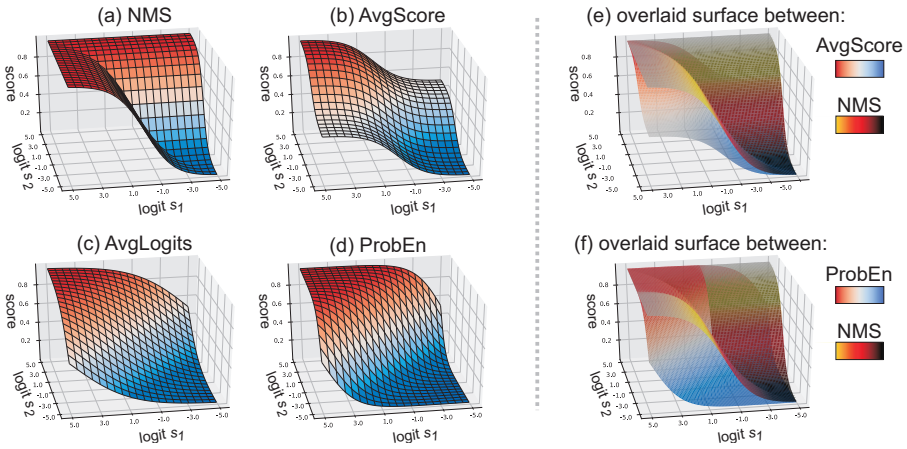

Fig. 1. Fusing logits from two single-modal, single-class detectors. Given a single class detector $k \in \{0, 1\}$, the single-modal class posterior for modality i depends on the relative logit $s_i = s_i[1] - s_i[0]$. We visualize the probability surface obtained by different fusion strategies that 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 1 in appendix). AvgLogits (c) and ProbEn (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 ProbEn and NMS (f), we find that ProbEn 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; ProbEn decreases the score, which helps reduce false positives as illustrated in Fig. 3.

its dependence on the underlying input modality x_i:

$$p(y = k|x_i) = \frac{e^{s_i[k]}}{\sum_j e^{s_i[j]}} \propto e^{s_i[k]} \tag{1}$$

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:

$$p(y = k|x_1, x_2) \propto \frac{p(y = k|x_1)p(y = k|x_2)}{p(y = k)} \propto \frac{e^{s_1[k]+s_2[k]}}{p(y = k)} \tag{2}$$

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.

Table 1. 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 underperform a simple NMS (MaxFusion). Eq.(2) derives that ProbEn 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 ProbEn at the cost of training on a carefully curated multimodal dataset. Our analysis shows that learned fusion can be seen as a generalization of ProbEn that no longer assumes conditionally-independant modalities (6).

Method	*Day*	*Night*	*All*
RGB	14.56	27.42	18.67
Thermal	24.59	7.76	18.99
Pooling	37.92	22.61	32.68
NMS (MaxFusion)	13.25	6.42	10.78
AvgScore	21.68	15.16	19.53
AvgLogits	18.78	11.70	16.28
LogRegFusion	10.70	6.11	9.08
ProbEn	10.21	5.45	8.62
ProbEn+bbox	9.93	5.41	8.50

Hence, *ProbEn corresponds to adding logits from each modality.* This suggests another practical implementation of ProbEn 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 [5]. This seminal work proposed an influential baseline for "fusing softmax scores" by averaging. However, practical implementations average logits [9,1] or sum logits [2], often omitting the final softmax [6] 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 (2)).

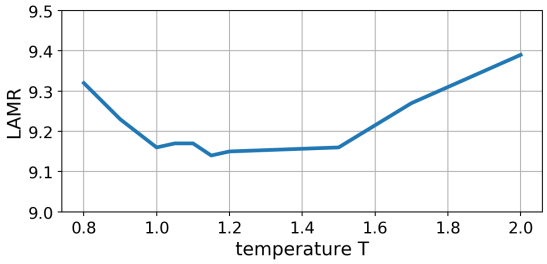

Fig. 2. LAMR as a function of a calibration temperature parameter T (designed to return more realistic probabilities) [4] on KAIST reasonable-test, We fuse detections from two single-modal detectors (RGB and thermal). Here, T=1 corresponds to ProbEn. Tuning the temperature T yields only marginally better 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.

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

$$s_{\text{AvgLogit}}[k] = .5(s_1[k] + s_2[k]) \tag{3}$$

$$s_{\text{Bayes}}[k] = s_1[k] + s_2[k] \tag{4}$$

It is easy to see that

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

Note that the relative ordering of the fused logits does *not* necessarily imply the same holds for the final posterior because 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. 1. Finally, Table 1 explicitly compares the performance of such fusion approaches with other diagnostic variants. We refer the reader to both captions for more analysis.

2 Score Calibration for Fusion

ProbEn assumes that detectors return true class posteriors. However, deep networks are notoriously over-confident in their predictions, even when wrong [4]. One popular calibration strategy is adding a temperature parameter T to the final softmax, typically to "soften" overconfident estimates [4]. This can be implemented by scaling logits by a temperature T:

$$s_i[k] \leftarrow s_i[k]/T, \qquad T > 0 \tag{5}$$

Table 2. Late-fusion methods on different underlying detectors measured by LAMR↓ on KAIST reasonable-test. This table is comparable to Table 1 in the main paper. A: RGB detector; B: Thermal detector; C: EarlyFusion detector; D: MidFusion detector. Clearly, ProbEn 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., $A+B$). Lastly, probabilistically fusing boxes (using v-avg) improves further over 8 / 9 fusion methods.

Method	*A+B*	*A+C*	*A+D*	*B+C*	*B+D*	*C+D*	*A+B+C*	*A+B+D*	*A+B+C+D*
Pooling	32.68	28.87	29.70	36.68	36.36	23.24	43.04	43.56	46.03
AvgScore	19.53	19.94	18.67	21.58	18.18	22.26	21.98	21.06	24.06
NMS	10.85	11.59	13.05	18.74	13.81	14.18	10.91	12.11	12.09
ProbEn	8.62	**9.63**	10.99	16.88	11.90	11.58	8.40	8.54	8.21
ProbEn + bbox	**8.50**	9.87	**10.30**	**16.87**	**11.20**	**11.32**	**8.55**	**7.66**	**7.45**

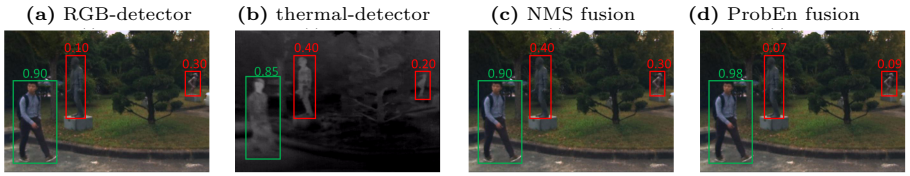

Fig. 3. ProbEn handles false positives by lowering scores. Fig. 1 (d) shows that ProbEn will *reduce* the fused score of overlapping detections with at least one low-scoring modality. This is an example from KAIST, where RGB- and thermal-detectors produce false-positive pedestrian detections for the statues. NMS fusion keeps the higher-scoring false-positive, while ProbEn lowers the fused score while keeping the higher score for the true-positive (that contain overlapping detections with consistently high scores).

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 2 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 ProbEn (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.

Interestingly, when we ensemble an off-the-self multimodal detector GAFF [8], our Thermal and RGB detectors (trained in-house), we find score calibration is particularly important. Importantly, we find that calibration requires not only a temperature variable but also a shift variable on the logits of GAFF. We conjecture that this is because GAFF is trained in a very different way; we do not know how GAFF is trained as there is not a publicly available codebase. Fig. 4 depicts the miss-rate as a function of the two variables, temperature T and shift b. Clearly, the shift variable b makes a significant impact on the fusion results.

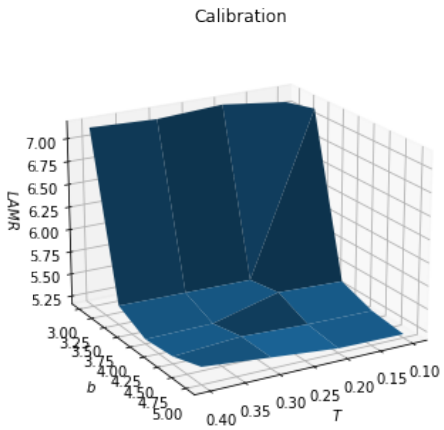

Fig. 4. LAMR as a function of calibration temperature parameter T and shift parameter b [4] on the KAIST validation set. We fuse single-modal detectors (RGB and thermal trained in-house) and an off-the-shelf detector GAFF [8]. Clearly, both the temperature T and shift b greatly affect the final detection performance.

Table 3. Late-fusion methods on different underlying detectors on FLIR dataset, measured by percent AP↑ in percentage. A: thermal detector; B: EarlyFusion detector; C: MidFusion detector. Our ProbEn method consistently outperforms other late-fusion methods. By fusing all the underlying detectors, ProbEn performs the best. Lastly, probabilistically fusing boxes (using v-avg) improves further for 3 / 4 fusion methods.

Method	*A+B*	*A+C*	*B+C*	*A+B+C*
Pooling	54.04	61.48	63.38	53.66
AvgScore	81.65	81.47	82.43	82.65
NMS	81.75	82.34	82.43	83.14
ProbEn	**82.05**	82.26	82.67	83.27
ProbEn + bbox	81.93	**82.85**	**83.04**	**83.76**

3 Further Study of Weighted Score Fusion

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

$$s_{\text{Learned}}[k] = w_1[k]s_1[k] + w_2[k]s_2[k] \tag{6}$$

One can view ProbEn, AvgLogits, and Temperature Scaling as special cases of the above. ProbEn 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

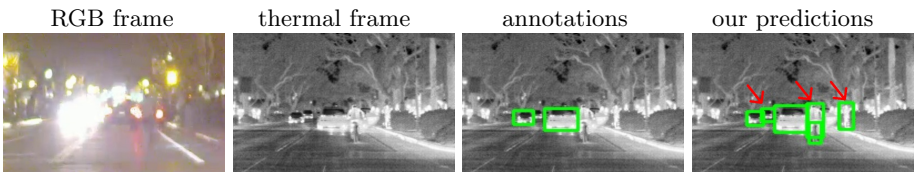

Fig. 5. We zoom in a frame from Fig. 8 to visualize more clearly that the ground-truth anntoations can even miss `bicycles` and `persons` as shown in the third image. In contrast, our ProbEn model can detect these miss-labeled objects (cf. red arrows). This shows the issues in the FLIR dataset.

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 1), consistent with prior art on late fusion [5]. 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 ProbEn (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. 8 and 5.

4 Further Study of Class Prior in ProbEn

In the main paper, we assume uniform class priors when using ProbEn. Now we test ProbEn 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. 6. Clearly, ProbEn works better with uniform priors than the computed the class piriors.

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. 7. We can see tuning specific class priors leads to marginal improvements compared to using uniform prior.

5 A Detailed Derivation of Probabilistic Box Fusion

In the main paper, we present a probabilistic method to fuse multiple bounding boxes. Below is a detailed derivation. 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 bbox coordinates can lie anywhere in the image plane. Doing so, we derive probabilistic box fusion:

$$
\begin{aligned}
p(\mathbf{z}|x_1, x_2) &\propto\ p(\mathbf{z}|x_1)p(\mathbf{z}|x_2) \\
&\propto\ \exp\big(\frac{\|\mathbf{z}-\boldsymbol{\mu}_1\|^2}{-2\sigma_1^2}\big)\exp\big(\frac{\|\mathbf{z}-\boldsymbol{\mu}_2\|^2}{-2\sigma_2^2}\big) \\
&\propto\ \exp\big(\frac{\mathbf{z}^T\mathbf{z}-2\boldsymbol{\mu}_1^T\mathbf{z}+\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1}{-2\sigma_1^2}\big)\exp\big(\frac{\mathbf{z}^T\mathbf{z}-2\boldsymbol{\mu}_2^T\mathbf{z}+\boldsymbol{\mu}_2^T\boldsymbol{\mu}_2}{-2\sigma_2^2}\big) \\
&\propto\ \exp\big(\frac{\mathbf{z}^T\mathbf{z}-2\boldsymbol{\mu}_1^T\mathbf{z}+\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1}{-2\sigma_1^2}+\frac{\mathbf{z}^T\mathbf{z}-2\boldsymbol{\mu}_2^T\mathbf{z}+\boldsymbol{\mu}_2^T\boldsymbol{\mu}_2}{-2\sigma_2^2}\big) \\
&\propto\ \exp\Big(\frac{(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2})\mathbf{z}^T\mathbf{z}-(\frac{2\boldsymbol{\mu}_1^T}{\sigma_1^2}+\frac{2\boldsymbol{\mu}_2^T}{\sigma_2^2})\mathbf{z}+\frac{\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1}{\sigma_1^2}+\frac{\boldsymbol{\mu}_2^T\boldsymbol{\mu}_2}{\sigma_2^2}}{-2}\Big) \\
&\propto\ \exp\Big(\frac{\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}}{-2}*(\mathbf{z}^T\mathbf{z}-2\frac{\frac{\boldsymbol{\mu}_1^T}{\sigma_1^2}+\frac{\boldsymbol{\mu}_2^T}{\sigma_2^2}}{\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}}*\mathbf{z})\Big) \\
&\propto\ \exp(\frac{||\mathbf{z}-\boldsymbol{\mu}||^2}{-2(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2})}), \quad \text{where} \quad \boldsymbol{\mu}=\frac{(\boldsymbol{\mu}_1/\sigma_1^2+\boldsymbol{\mu}_2/\sigma_2^2)}{(1/\sigma_1^2+1/\sigma_2^2)}
\end{aligned}
$$

6 A Study of Fusing More Models

We study late fusion methods on more combinations of underlying detectors. Table 2 and 3 list results on KAIST and FLIR datasets, respectively. Importantly, ProbEn consistently performs the best on each of combinations. Interestingly, applying ProbEn method to detectors that are not independent to each other (e.g., Thermal and MidFusion) can achieve better performance. Admittedly, the improvements may not be statistically significant and overfitting may be an issue. This can not be resolved or studied further using contemporary datasets which are relatively small. Therefore, we solicit a larger-scale dataset to benchmark multimodal detection in the commmunity.

7 Qualitative Results and Video Demo

We attach a demo video named `video_demo.mp4` on a testing video (captured at night) provided by the FLIR dataset. In the demo video, we compare the

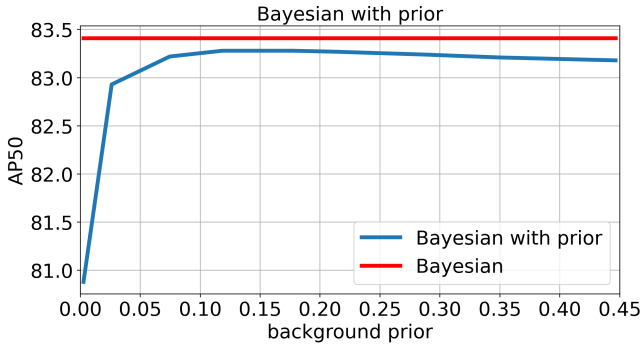

Fig. 6. A study of ProbEn with class priors as class frequencies in the training set. We use FLIR dataset for this study as it has 3 imbalanced classes. We fuse three models (Thermal, Early and Mid) as used in the main paper. As there is a background class, we vary the background class and proportionally change the class priors. Clearly, ProbEn 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.

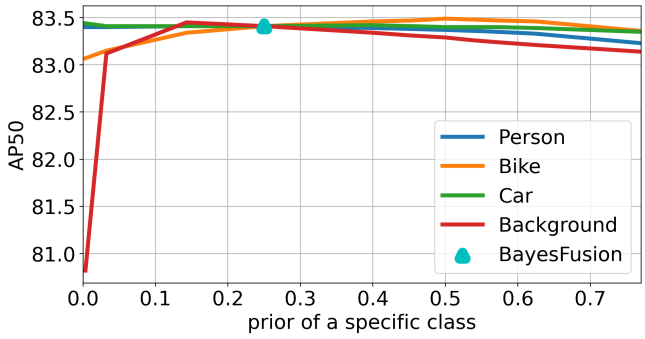

Fig. 7. A study of tuning a single class prior while keeping others the same. Motivated by the superior performance of ProbEn with uniform priors, we tune each of the class prior by fixing others the same. We study this on the FLIR dataset by fusing three models (Thermal, Early and Mid). We can see that tuning specific classes only marginally improves detection performance.

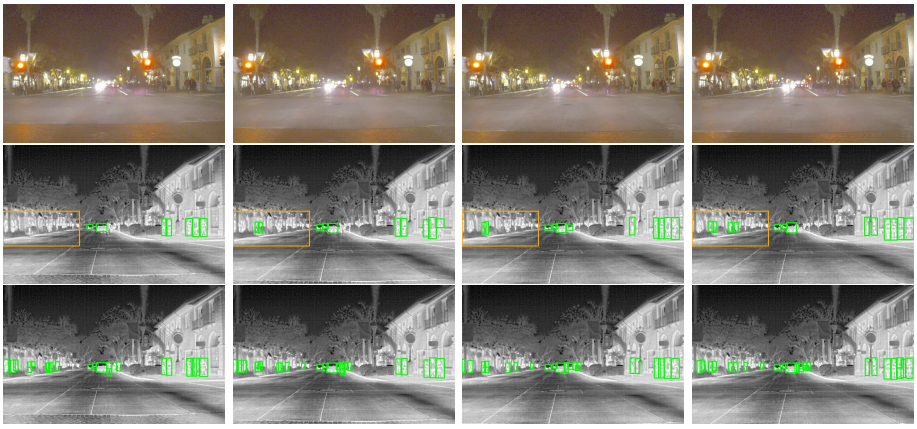

Fig. 8. We demosntrate inconsistent annotations in FLIR dataset with four consecutive frames in the validation set. **top-row** lists four RGB frames for reference. **mid-row** displays thermal images and the ground-truth annotations. Looking at the annotations in the orange rectangle, we can see that the annotations are not consistent across frames. This is an critical issue that prevent learning-based late fusion from improving further on the FLIR dataset. **Bottom-row** displays the detection results by ProbEn 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. 5 for a zoom-in visualization.

detection results by the Thermal model and ProbEn that fuses results of three models (Thermal + Eary + 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. 9).

Lastly, we provide more qualitative results in Figure 10 and 11 for KAIST and FLIR, respectively. Visually, we can see our ProbEn method performs better than the compared methods.

References

1. Dong, W.: https://github.com/wushidonguc/two-stream-action-recognition-keras/blob/master/fuse_validate_model.py#L61. commit 0a3e722 3
2. Feichtenhofer, C.: https://github.com/feichtenhofer/twostreamfusion/blob/master/cnn_ucf101_fusion.m#L206. commit 3e313c4 3
3. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: CVPR (2016) 1
4. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. arXiv:1706.04599 (2017) 4, 6
5. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NeurIPS. pp. 568–576 (2014) 1, 3, 7

Fig. 9. We attach a demo video in our Github repository 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 ProbEn 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, ProbEn successfully detects all the cars and produces tight bounding boxes. We refer the reader to the video demo for convincing visualization.

6. Wu, Z., Wang, X., Jiang, Y.G., Ye, H., Xue, X.: Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In: Proceedings of the ACM international conference on Multimedia (2015) 3
7. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: CVPR (2015) 1
8. Zhang, H., Fromont, E., Lefèvre, S., Avignon, B.: Guided attentive feature fusion for multispectral pedestrian detection. In: WACV (2021) 5, 6
9. Zhu, Y.: https://github.com/bryanyzhu/two-stream-pytorch/blob/master/scripts/eval_ucf101_pytorch/temporal_demo.py#L73. commit 32b6354 3

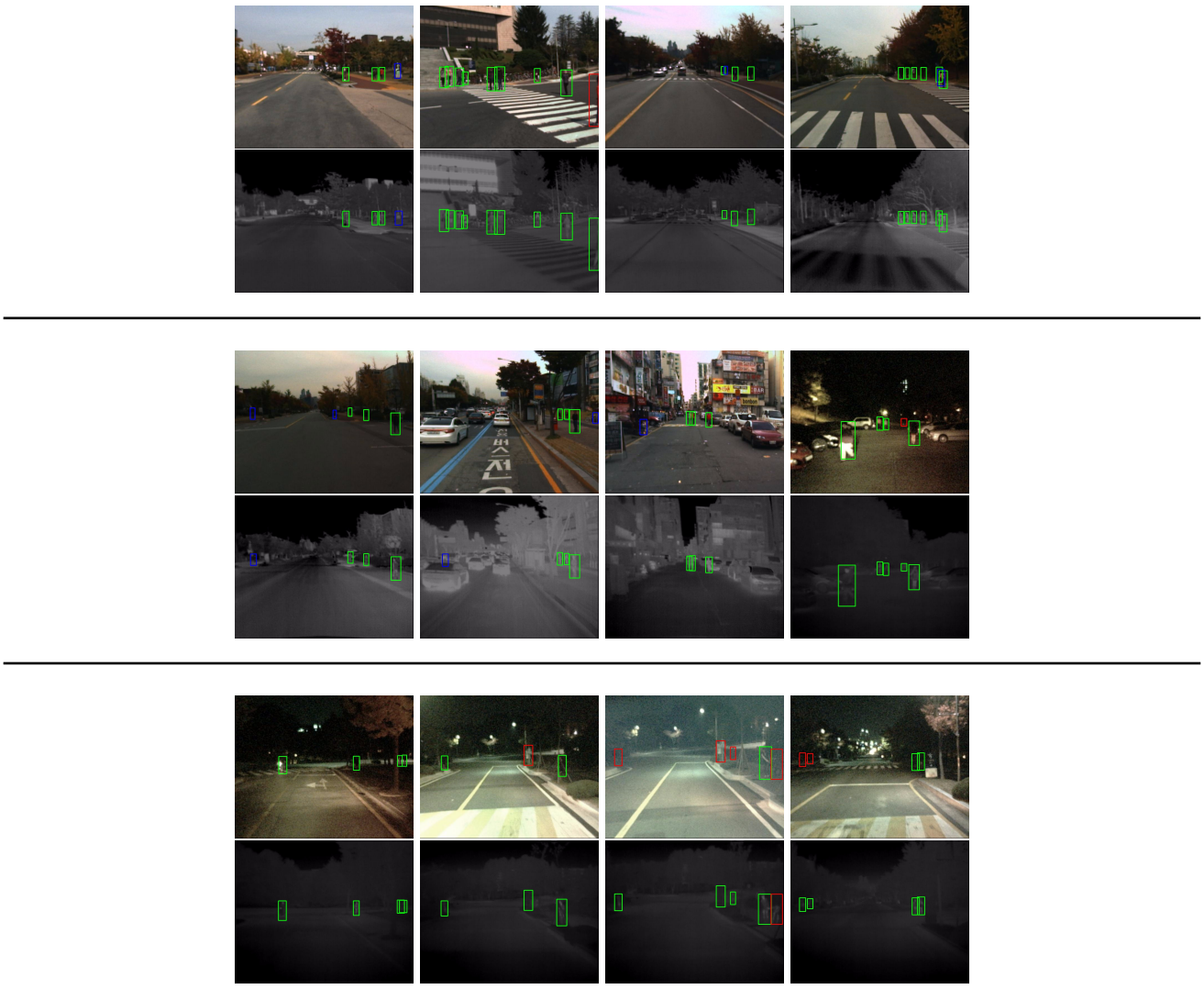

Fig. 10. Qualitative results on more testing examples in KAIST dataset. We place RGB-thermal images in pairs: in each macro row, we show RGB images in the upper row and thermal images in lower row. Over RGB images, we overlay the detection results from our MidFusion model; on the thermal images, we show results from our best-performing ProbEn model. Green, red and blue boxes stand for true positives, false negative (miss-detected persons) and false positives.

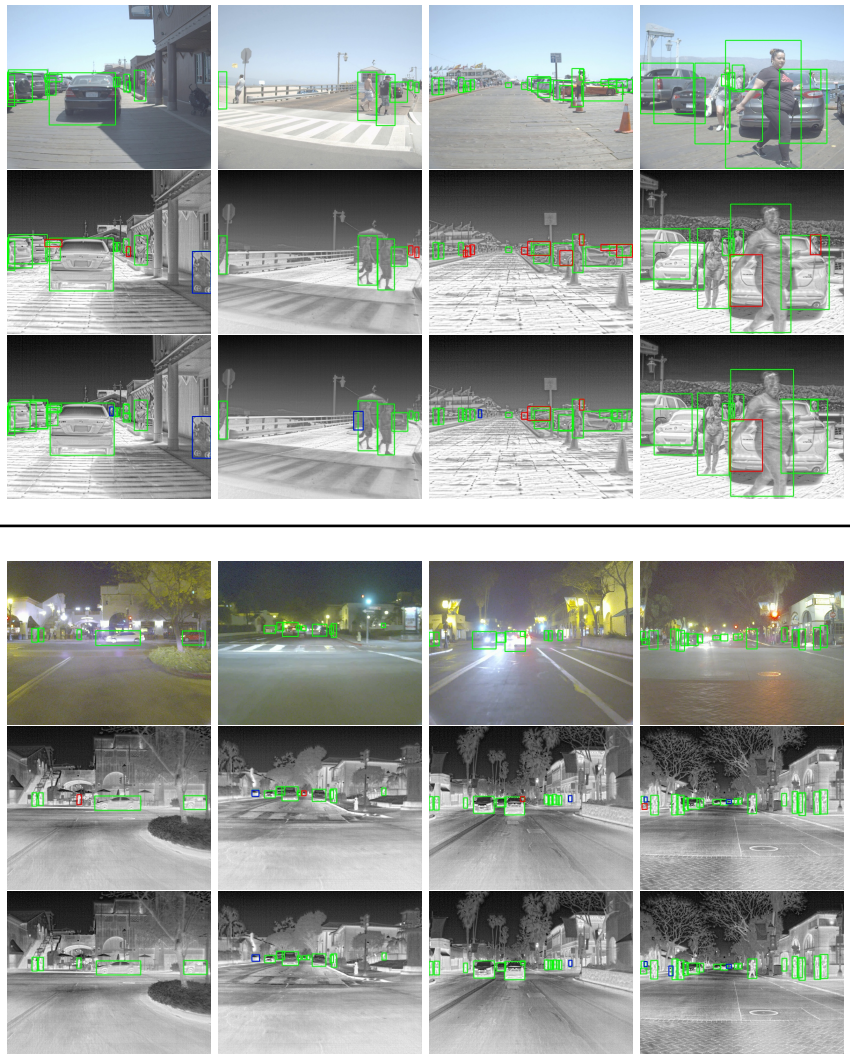

Fig. 11. Qualitative results on more testing examples in FLIR dataset. We place RGB-thermal images in triplet: in each macro row (divided by the black line), we show RGB images in the upper row and thermal images in two lower rows. Over RGB images, we overlay ground-truth annotations, highlighting that 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 ProbEn (with bounding box fusion) model (bottom-row). Green, red and blue boxes stand for true positives, false negative (mis-detected persons) and false positives.