Sensor Fusion for Object Detection based on Test-Time Augmentation

Alexandre Fonseca
Instituto Superior Técnico
Lisbon, Portugal
alexandre.andrade.fonseca@tecnico.ulisboa.pt

***Abstract*—In this work, we propose a method for late-stage sensor fusion of object detection predictions in road environments using RGB and thermal images, to leverage complementary information in unfavourable illumination conditions. We use Test Time Augmentation to obtain uncertainty estimations on predictions made by a neural network object detector on images from different modalities. We show that fusion results outperform base predictions on both accuracy of detection and quantification of classification uncertainty. We also briefly study the impact that different augmentations have on results. The proposed method is extendable to an arbitrary number of detectors and/or sensor modalities as long as data is strongly synchronized spatially and temporally.**

***Index Terms*—Sensor Fusion, Uncertainty Quantification, Test Time Augmentation**

I. Introduction

The task of object detection consists in identifying the position and type of objects in an image. Since the introduction of the AlexNet [1], many advances on the field of neural networks have been made. Some notable architectures, like YOLO [2], are capable of performing real-time object detection by achieving high inference quality at high frame rates.

Despite efforts to integrate such systems in real world applications, such as autonomous driving and medical image analysis, their use is still severely limited because predictions obtained from these models are not trustworthy [3].

Given the importance of reliable predictions in safety-critical applications, several techniques for estimating detector uncertainty have been developed, such as Monte Carlo Dropout, Deep Ensembles, Direct Modeling, and more recently, Test Time Augmentation (TTA). An in-depth review of those and other methods can be found in [4].

Fig. 1. Example of RGB/thermal image pair at night. A pedestrian was detected in the thermal image, but not in the RGB image.

TTA has two main advantages when compared to alternative methods for uncertainty estimation: it does not require retraining of the neural network, enabling the use of the detector as a black box, and it is extremely easy to implement and integrate onto existing systems. Through that method, many predictions of a same object can be clustered, and probabilistic distributions for both bounding box and classification score can be estimated, allowing uncertainty to be quantified.

For some applications like autonomous driving, the range of operation conditions is quite diverse due to different weather conditions (sun glare, fog), environments (urban, rural, highway) and road types. In some of these conditions, sensor data can be unreliable. For instance, RGB cameras are not reliable at night, as can be observed in Figure 1. As such, it is necessary to gather data from different types of sensors and fuse individual information to obtain a more trustworthy prediction, through a process called Sensor Fusion (SF).

With the goal of improving object detection in autonomous vehicles equipped with RGB and thermal cameras, this work offers the following contributions:

- We propose a late-stage fusion method for RGB and

thermal images based on TTA to characterize both individual and fused uncertainty. It is also extensible to an arbitrary number of detection models and does not require retraining.

- We propose to model classification uncertainty as a Dirichlet distribution to obtain better estimates compared to categorical distributions (softmax output) [5]. We show that such type of fusion produces more confident predictions than individual models when compared to score averaging, which increases uncertainty [6].

In Section II a literature review on current methods for sensor fusion and uncertainty estimation in deep learning models will be conducted. In Section III we will describe the proposed method, as well as its theoretical foundations. In Section IV, we detail the experimental setup, and report the results of experiments that expose the performance of the proposed method and its advantages over the state of the art. Finally, in Section V we present our conclusions, and suggest some directions for future work.

II. Related Work

A. Sensor Fusion

The process of sensor fusion can occur at three different levels [7]: low-level (or early-stage), mid-level (or feature-level) and high-level (or late-stage).

Low-level fusion: Low-level fusion focuses on fusing raw-data inputs. Farahnakian and Heikkonen [8] concatenate channels from RGB and thermal images to create a 4-channel multispectral image. The new image is then used used as input of a neural network object detector.

Mid-level fusion: Given the capability of deep learning models to represent data as low-dimension feature vectors, mid-level fusion is often referred to as feature-level fusion, and it is a popular approach to design end-to-end fusion architectures.

Nobis et al. [9] proposed the CRF-Net to fuse RGB images with radar detections in an end-to-end fashion. Fusion occurred in several intermediate layers so that the network could learn at which point fusion would be optimal. DenseFuse [10] makes use of auto-encoder technology to extract and reconstruct image features from RGB and thermal inputs.

High-level fusion: Most high-level fusion methods employ so-called traditional methods. Those are typically based on probabilistic and statistical models like the Kalman Filter (KF) and Joint Probabilistic Data Association (JPDA) [11]. Shahian et al. [12] use the Extended KF to fuse LiDAR data with RGB detections. While that approach was only tested for single-target tracking, it can be extended for the multi-object case.

Haghbayan et al. [13] use JPDA to fuse region proposals from RGB and thermal cameras, as well as radar and LiDAR on maritime environment. Fused proposals are then fed to a CNN for classification.

ProbEn [6] models object location and classification obtained from a detector as Gaussian and categorical distributions, respectively. Through direct application of Bayes' theorem, fused distributions are computed to produce better results in detection metrics. However, some details remain unexplored. First, a clear way to estimate individual sensor uncertainty is not proposed. Second, estimated distributions are not evaluated on any metric. Finally, most proposed methods for bounding box fusion are based on heuristics, rather than being mathematically formulated.

B. Uncertainty Quantification in Deep Learning Models

In deep learning models, uncertainty is often categorized as epistemic and aleatoric [14]. The former refers to uncertainty manifested by the model when making predictions on out-of-distribution data, i.e. patterns not learned during training. The latter expresses uncertainty intrinsic to the data itself.

Over the last few years, several methods for estimating both types of uncertainty in deep neural networks have been developed. Some of these include Bayesian Neural Networks (BNNs), Direct Modeling (DM), and more recently, TTA.

BNNs: Those networks quantify uncertainty by estimating a posterior distribution over the model's weights. It is necessary to use approximation techniques, since this problem is, in general, mathematically intractable. One of such methods is Monte Carlo Dropout (MCD). BayesOD [15] uses MCD to obtain dense anchor predictions from an object detector, and estimates uncertainty by performing Bayesian Inference.

DM: Direct Modeling estimates uncertainty at the output layers of a neural network. Contrary to BNNs which marginalize the distribution over the networks parameters, DM uses point estimates of the weights to predict the parameters of an assumed distribution. Gaussian YOLOv3

[16] employs DM by outputting bounding box mean and variance estimations.

TTA: TTA applies data augmentation at test-time in order to obtain a larger set of predictions.

Wang et al. [17] use TTA to estimate uncertainty in pixel-segmentation of magnetic resonance images. They conclude that TTA provides higher quality estimates than MCD by providing less overconfident misclassifications.

TTA has not yet received much attention as a technique for uncertainty estimation in the domain of autonomous vehicles [18] being mostly used to boost detection metrics.

III. Methodology

A. Proposed Method

A simple overview of the proposed method is shown in Figure 2. First, several augmentations of the input image are fed to the object detector. Predictions are accumulated and clustered by Intersection over Union (IoU) thresholding. The output of the detector are bounding boxes and their respective classification vectors after Non-Maximum Supression (NMS). Parameters of Gaussian and Dirichlet distributions are computed to fit, respectively, the bounding boxes and classification vectors of a cluster.

Repeating this process for all detection models results in several clusters, some of which correspond to a same object in different sensors. Clusters from different models are matched by IoU, considering the sample mean as a cluster's bounding box coordinates, and statistics are fused to generate a final result.

We note that the method can be extended to an arbitrary number of detections models or sensors. For instance, if an object detector fuses RGB images with radar or LiDAR scans, its predictions could also be used with this method.

Since data from different sensors does not require necessarily different models, and simultaneously, the same data can be evaluated on different models, we will use the terms "model" and "sensor" interchangeably in this section.

B. Image Augmentations

When using TTA, augmentations must be chosen purposefully. Generally speaking, augmentations should not alter the perceived label of a classification target nor bounding box locations. As such, we opt to use intensity level transformations for both types of images, in lieu of typical geometric augmentations. Specifically, we use linear brightness and contrast transformations, gamma correction and Gaussian blurring. For our experiments and model parameter selection, we used a fixed number of augmentations with fixed parameter values [18]: (i) brightness with factors 0.7 and 1.4 (ii) linear contrast with factors 0.6 and 1.4 (iii) gamma correction with parameter 0.6 and 1.5 (iv) Gaussian blur with parameter 1 and 2.5

C. Sample Statistics Computation

Accumulated predictions from N augmented images as $\mathcal{P} = \{\mathcal{P}_1, \cdots, \mathcal{P}_N\}$, where $\mathcal{P}_i = \{d_1, \cdots, d_n\}$ is a list of detections, and each detection $d_i = \{\mathbf{b}_i, \mathbf{s}_i\}$ is characterized by its bounding box coordinates, $\mathbf{b}_i \in \mathbb{R}^4$, and probability scores, $\mathbf{s}_i \in \mathbb{R}^K$, where K is the number of classes. After performing clustering by IoU on the accumulated predictions $\mathcal{P}$, we estimate bounding box and classification sample statistics for each individual model. Finally, fusing individual distributions can be done via Bayesian Inference. Fused detections are denoted as $\mathcal{D}_i = \{\mathcal{D}_{box}, \mathcal{D}_{cl}\}$, where $\mathcal{D}_{box}$ is the fused bounding box distribution and $\mathcal{D}_{cl}$ is the fused classification distribution.

For simplicity, the following derivations are made with respect to a single object in an image.

Gaussian Parameter Estimation: Bounding box coordinates are modeled as a Gaussian distribution. The mean $\overline{\mathbf{b}_i}$ and covariance Σ_i for a cluster of size t can be computed as

$$\overline{\mathbf{b}_i} = \frac{1}{t}\sum_{j=1}^{t}\mathbf{b}_j, \quad \Sigma_i = \frac{1}{t}\sum_{j=1}^{t}(\mathbf{b}_j - \overline{\mathbf{b}_i})(\mathbf{b}_j - \overline{\mathbf{b}_i})^T. \quad (1)$$

Similarly to [15], [6], the process of fusion for Gaussian distributions can be achieved by application of Bayes' theorem:

$$p(\mathcal{D}_{box}|\boldsymbol{b}_1, \cdots, \boldsymbol{b}_M) \propto \prod_{m=1}^{M} p(\mathcal{D}_{box}|\boldsymbol{b}_m) \quad (2)$$

where the posterior is $p(\mathcal{D}_{box}|\boldsymbol{b}_1, \cdots, \boldsymbol{b}_M) \sim \mathcal{N}(\boldsymbol{\mu'}, \Sigma')$. The fused mean vector $\boldsymbol{\mu'}$ and fused covariance Σ' can be computed via the Bayesian update formulae

$$\Sigma' = \left(\sum_{m=1}^{M}(\Sigma_m)^{-1}\right)^{-1} \quad \boldsymbol{\mu'} = \Sigma'\left(\sum_{m=1}^{M}\Sigma_m^{-1}\overline{\boldsymbol{b}_m}\right) \quad (3)$$

where Σ_m and $\boldsymbol{\mu_m}$ are the covariance and mean of an object detected in model m.

An important detail of our implementation lies in the fact that only clusters with size greater than 4 are

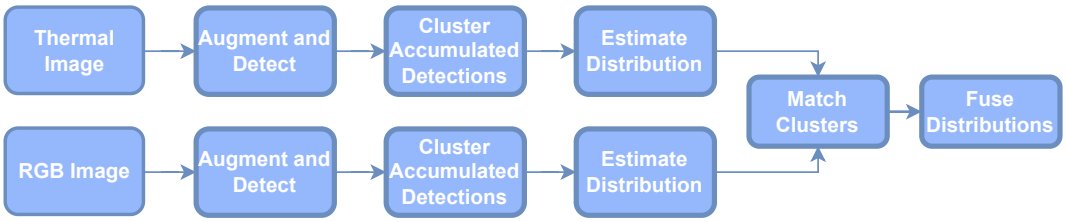

Fig. 2. Schematic overview of the proposed method

considered as detections. Intuitively, it makes sense that small clusters, i.e. an object detected in a small number of augmentations, would be a false detection. However, this choice is not arbitrary, since to estimate an n-dimensional non-singular sample covariance matrix, $n+1$ data points are needed. Even then, there are no guarantees that the covariance is non-singular, since in the case where all samples are equal, the result would be a zero-valued covariance. Nonetheless, this can be solved through additive smoothing of the diagonal values of the covariance matrix.

Dirichlet Parameter Estimation: Classification vectors obtained from object detectors are usually presented in the form of a categorical distribution. As noted in [6], methods like NMS or score averaging are are commonly used. However, they also note that NMS should not be considered a fusion method, since it only discards predictions. Additionally, it is stated that averaging will produce lower scores than single-modal cases, when it is intuitive that scores should be higher if a detection is made in more than one sensor. As such, we choose to model classification normalized score vectors as samples of a Dirichlet distribution, $Dir(\boldsymbol{\alpha})$, with $\boldsymbol{\alpha} \in \mathbb{R}^K$.

Unlike the Gaussian case, the parameters $\boldsymbol{\alpha}$ of a Dirichlet distribution cannot be computed in closed-form from a set of samples (normalized classification vectors). Iterative algorithms [19] have been proposed, but are found to be numerically unstable when the sample size is low, which is the case in this work. With this in mind, we estimate the posterior's parameters as proposed in [15]. When applying Bayes' theorem with a prior $p(\mathcal{D}_{cl}) \sim Dir(\boldsymbol{\alpha})$

$$p(\mathcal{D}_{cl}|\boldsymbol{s}_1, \cdots, \boldsymbol{s}_j) \propto p(\mathcal{D}_{cl}) \prod_{j=1}^{t} p(\mathcal{D}_{cl}|\boldsymbol{s}_j), \tag{4}$$

$\boldsymbol{\alpha}$ can be approximated in closed-form, for a single model, as

$$\boldsymbol{\alpha}' = \boldsymbol{\alpha} + \sum_{j=1}^{t} \boldsymbol{s}_j, \tag{5}$$

where $\boldsymbol{s}_j$ is the classification vector of the j^{th} cluster member.

Eq. (4) differs from Eq. (2) in the fact that a prior was used. Since the product of Gaussian distributions is also a Gaussian distribution (and a density function up to a scale factor), the use of a prior is not needed. Like in [15], the Dirichlet prior is chosen to be non-informative and its parameters are $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K) = (1/K, \cdots, 1/K)$, where K is the number of classes.

Contrary to the Gaussian case, the product of Dirichlet distributions is not, in general, a Dirichlet distribution. However, because we have access to the original data samples from all sensors, we do not need to compute the Dirichlet distributions of each individual sensor and then fuse the results. Instead, we can readily compute the fused distribution by estimating the overall Dirichlet distribution with the individual classification scores from matching clusters. Using the same prior, Eq.(4) and Eq.(5) can then be rewritten to account for this as

$$p(\mathcal{D}_{cl}|\boldsymbol{s}_{11}, \cdots, \boldsymbol{s}_{Mt_m}) \propto p(\mathcal{D}_{cl}) \prod_{m=1}^{M} \prod_{j=1}^{t_m} p(\mathcal{D}_{cl}|\boldsymbol{s}_{mj}), \tag{6}$$

$$\boldsymbol{\alpha}' = \boldsymbol{\alpha} + \sum_{m=1}^{M} \sum_{j=1}^{t_m} \boldsymbol{s}_{mj} \ . \tag{7}$$

where t_m denotes the size of a cluster in model m.

IV. Experiments and Results

A. Experimental Setup

In this work, we use YOLOv5s [20] as the object detector. We evaluate the proposed method on an aligned version of the FLIR dataset [21], introduced by Zhang et al. in [22]. Compared to the original, only the classes

person, *car*, and *bicycle* are considered. Particularly, we use the validation split containing 1013 RGB/thermal aligned image pairs with 640x512 pixel resolution, and the following class balance: 4107 *person*, 4124 *car*, 360 *bicycle*. Finally, we note that, while image pairs are well-alligned as far as we can tell, images are captured in different perspectives. Because of this, and since annotations are made only for thermal images, RGB detections will be less accurate.

The model is pre-trained on the COCO dataset [23], which specifies 80 classes. We note that thermal images are **not** found in the training set. Since the aligned FLIR dataset only contains three classes, we discard detections whose class does not match any of these. Pratically, this choice does not have a big impact because 85% of detections are within the admissible set. Additionally, we consider the COCO classes *truck* and *bus* to be equal to *car*, and *motorcycle* to be equal to *bicycle*.

B. Evaluation Metrics

In both individual models and fusion, we consider a detection's effective bounding box as the mean vector of its Gaussian distribution, and its predicted class the one with highest Dirichlet expected value

$$p_i = \frac{\alpha_i}{\sum_{j=1}^{K} \alpha_j}. \tag{8}$$

Detections with IoU larger than 0.5 and correct predicted class are considered True Positive (TP). If any of these criteria is not met, a detection is considered False Positive (FP). Finally, if a ground-truth box does not have a matching detection, it is counted as a False Negative (FN).

Feng et al. [24] remark that there is no standard metric to evaluate probabilistic object detectors, but note that mean Average Precision (mAP) is the most used metric to evaluate object detectors, even if it does not factor in uncertainty estimations. To evaluate the quality of estimated probabilistic distributions, they use the Negative Log Likelihood (NLL), which we will also use in this work. For a Gaussian distribution parameterized by mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, its NLL to a ground-truth vector Z can be computed as

$$\text{NLL}(\boldsymbol{\mu}, \Sigma, Z) = \frac{1}{2}(Z - \boldsymbol{\mu})^T \Sigma^{-1} (Z - \boldsymbol{\mu}) + \frac{1}{2} \log \det \Sigma. \tag{9}$$

For the classification task, we compute the Dirichlet categorical NLL, NLL_{dir}, as well as the average classification score NLL, NLL_{avg}, as

$$\text{NLL}(\boldsymbol{p}, \boldsymbol{y}) = \sum_{i=1}^{K} -y_i \log p_i, \tag{10}$$

where $\boldsymbol{y}$ is the ground-truth class encoded as a one-hot vector and p_i are the per-class Dirichlet expected values or the per-class average score, for the Dirichlet case and average case respectively.

The value for mAP is achieved by averaging AP across multiple IoU thresholds. For clarity, we denote this as IoU_{AP}. In particular, the COCO data set benchmark computes mAP in the interval 0.5:0.95 in steps of 0.05. Given the different perspective misalignment introduced by the data set, higher values for the IoU threshold would heavily impact the overall results, by considering many RGB and fused true predictions as false. Because of this, we choose to compute AP at IoU_{AP}=0.5, denoted as AP@50, as well as mAP in the more relaxed interval 0.5:0.75 with steps of 0.05. For the sake of simplicity, we still refer to this simply as mAP.

A less common metric in object detection, but often found in sensor fusion research is Miss Rate (MR). We evaluate the capability of the proposed method of correctly identifying more objects by computing MR, defined as

$$MR = \frac{FN}{FN + TP}. \tag{11}$$

C. Experiments

The first experiment we conduct aims to find the best values for the IoU thresholds both for detection clustering, IoU_c, and cluster matching, IoU_m. In [18], the impact of different augmentations on RGB images is studied, and it is concluded that a combination of brightness and contrast augmentations yielded the best results in the used data set. Inspired by that work, we do the same experiment on both types of images of the FLIR data set. Finally, we evaluate the results obtained for both the individual modalities and the sensor fusion.

D. Results

The step of accumulating predictions (third module in Fig.2) is achieved by using threshold IoU_c. To determine the optimal value for IoU_c, we compute RGB and thermal mAP for values of IoU_c in the interval 0.5:0.85 with steps of 0.05 (Fig. 3 and Fig. 4). Maximum

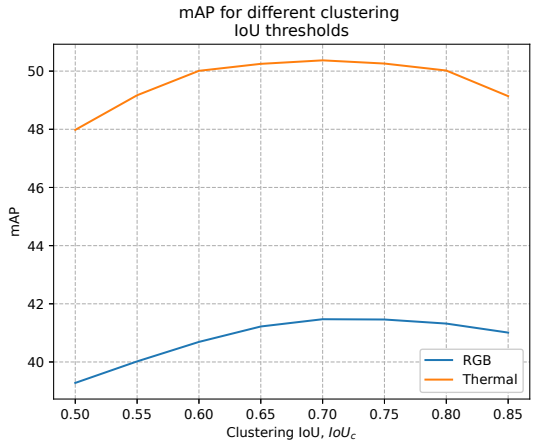

Fig. 3. RGB and Thermal mAP for different IoU_c values

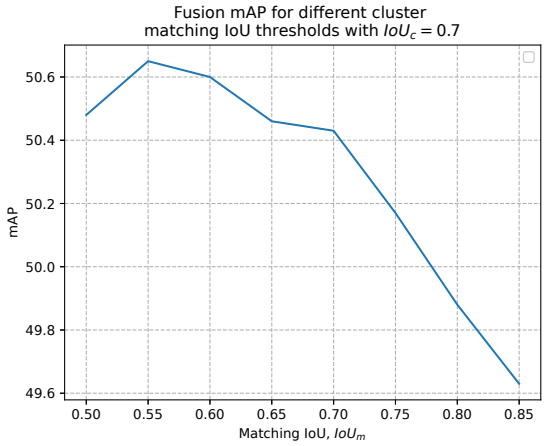

Fig. 4. Fused mAP for different IoU_m values with $IoU_c = 0.7$

RGB and thermal mAP was achieved for $IoU_c = 0.7$. Fixing this value, we now sweep values for IoU_m in the same interval and compute fusion mAP. Best results were obtained for $IoU_m = 0.55$. As one could expect, a more lenient value for IoU_m is needed to achieve the best results, considering the perspective mismatch between RGB/thermal image pairs. For the final results, we considered the proposed method used with $IoU_c = 0.7$ and $IoU_m = 0.55$.

We assess the impact in performance of each type of augmentation both RGB and thermal images, and show results in Table I. We conclude that Gaussian blur should be avoided because of the large drop in performance in detection compared to other augmentations.

We now evaluate the impact in performance of the other augmentations (brightness, gamma and contrast) individually or in combinations. We find that, for thermal images, gamma augmentation alone performs the best in all considered augmentation scenarios. Qualitatively, this can be observed in Figure 5. Gamma augmentation can preserve edges and other details better than brightness and contrast augmentations. For RGB images, results show that different augmentations yield better results in different metrics. For the final evaluation, we consider RGB brightness as the best augmentation because despite being showing the best results in only one metric, does not heavily sacrifice performance in the remaining metrics, i.e. shows the most balanced trade-off in all metrics.

Our method shows considerable improvements in MR, scoring 13pp and 11pp less missed detections than the base RGB and thermal models, respectively. For classification uncertainty, we compare the proposed Dirichlet fusion method with score averaging. We note that since the type of distribution used for these cases is different (Dirichlet and Categorical respectively), NLL values should only be compared relatively to values in the same column, i.e. NLL values obtained from the same type of distribution. In particular, we show that, compared to the base models, Dirichlet class fusion gives *better* estimates, while score average fusion results in *worse* estimates. NLL_{avg} shows better metric results than NLL_{dir} because we only use predictions made by the detector post-NMS, meaning that classification vectors will naturally show very high probability to one class and very low values for the remaining classes. This also means that the predicted class will often be correct, resulting in better NLL. Additionally, the use of a symmetric Dirichlet prior means that incorrect classes will, from the beginning, assume a non-negligible probability.

Finally, Table I verifies the hypothesis introduced in Section IV-B, that large IoU_{AP} values are detrimental to fusion, as evidenced by the small improvements in mAP but more considerable gains in AP@50. This happens because fused bounding boxes computed using Eq.(3) are a weighted average of detections from individual models, whose predictions suffer from perspective misalignment. This also has a negative impact in regression NLL since fused bounding boxes will be less accurate, resulting in worse NLL_{reg}.

However, it is possible that the fused bounding box is closer to the ground-truth, i.e. lower squared error $(Z - \boldsymbol{\mu})^T(Z - \boldsymbol{\mu})$ as defined in Eq.(9), yet has worse

TABLE I

EVALUATION OF THE PROPOSED METHOD USING DIFFERENT AUGMENTATIONS. VALUES IN BOLD REPRESENT THE BEST VALUES FOR EACH CASE (RGB/THERMAL/FUSED). VALUES IN ITALIC REPRESENT THE BEST VALUE OVERALL.

	mAP↑	**AP@50↑**	**MR↓**	**NLL$_{reg}$↓**	**NLL$_{dir}$↓**	**NLL$_{avg}$↓**
RGB (All)*	41.47	56.65	37.59	**182.427**	0.0873	0.0150
RGB (Brightness)	41.81	**57.51**	36.84	230.500	0.0845	0.0143
RGB (Contrast)	41.46	56.7	37.3	257.644	0.0857	0.0152
RGB (Gamma)	**41.89**	57.12	**36.67**	358.967	**0.0824**	**0.0130**
RGB (Gaussian Blur)	40.51	54.48	42.07	241.101	0.0897	0.0172
RGB (Brightness + Contrast)	41.42	56.8	36.91	236.385	0.0846	0.0144
RGB (Brightness + Gamma)	41.55	57.04	36.81	243.043	0.0842	0.0139
RGB (Contrast + Gamma)	41.43	56.69	36.84	246.921	0.0848	0.0143
Thermal (All)*	50.37	61.42	35.29	***134.093***	0.0907	0.0136
Thermal (Brightness)	50.76	61.73	34.91	196.737	0.0834	0.0121
Thermal (Contrast)	51.16	62.34	34.11	178.922	0.0856	0.0135
Thermal (Gamma)	**51.83**	**63.60**	**32.73**	260.058	**0.0828**	***0.0128***
Thermal (Gaussian Blur)	42.69	50.07	49.66	146.053	0.0928	0.0159
Thermal (Brightness + Contrast)	50.8	62.11	34.16	146.917	0.0870	0.0134
Thermal (Brightness + Gamma)	51.17	62.73	33.34	164.694	0.0845	0.0124
Thermal (Contrast + Gamma)	51.46	62.95	33.29	152.763	0.0854	0.0135
Fused (RGB All / Thermal All)*	50.65	65.2	25.74	**229.360**	0.0693	0.0171
Fused (RGB Brightness / Thermal Gamma)	***51.84***	***66.52***	***24.34***	353.189	***0.0637***	**0.0155**

*Baseline cases use all augmentations as described in Section III-B

Fig. 5. Gamma augmentation (bottom-right) preserves more details than brightness (top-right) and contrast (bottom-left), even for large parameter values. Effects are noticeable on pedestrians on the left and right regions of the images. Top-left shows the original image for comparison.

regression NLL. This happens because fused precision is larger than individual models, as evidenced in Eq.(3), and it is a penalizing factor in Eq.(9). In a way, both individual models make accurate predictions, yet these are different by virtue of perspective misalignment, resulting in a fused model with very high prediction but a not accurate enough prediction. Two examples of this can be observed in Fig. 6 and Fig. 7. This problem could be better handled if images had the same perspective or by using feature-fusion detectors, which are able to perform implicit alignment [6].

V. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a method to probabilistically fuse predictions made by neural network object detectors in RGB/thermal image pairs based on TTA. This method is shown to greatly improve the rate of correctly identified objects.

We also show that modeling classification as a Dirichlet distribution when fusing predictions from multiple models yields better uncertainty estimates than the individual base models, which was noted by some authors [6] to be a shortcoming of current late-stage fusion methods such as score averaging.

Fig. 6. Fused box (yellow): squared error = 11.269, NLL_{reg} = 102.022. RGB box (blue): squared error = 12.124, NLL_{reg} = 75.701. Thermal box (red): squared error = 18.654, NLL_{reg} = 16.519. Ground-truth (green) for comparison.

Fig. 7. Fused box (yellow): squared error = 10.817, NLL_{reg} = 693.187. RGB box (blue): squared error = 14.353, NLL_{reg} = 523.103. Thermal box (red): squared error = 17.176, NLL_{reg} = 208.560. Ground-truth (green) for comparison.

As future work, we leave some suggestions to further enhance detection results. First, given that this method can be used with an arbitrary number of detection models, feature-fusion with multiple sensors can be explored. Some works [10][9] already use sensors with heterogeneous types of data like RGB images with radar and LiDAR. Our method could be integrated with such methods to create multi-level fusion schemes. We also do not use estimated distributions to filter possible false positives. Intuitively, if a detection has high uncertainty, it also has a higher chance of being incorrect. Future work can focus on creating decision methods to further refine predictions.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014.

[4] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, *et al.*, "A survey of uncertainty in deep neural networks," *arXiv preprint arXiv:2107.03342*, 2021.

[5] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in neural information processing systems*, vol. 31, 2018.

[6] Y.-T. Chen, J. Shi, Z. Ye, C. Mertz, D. Ramanan, and S. Kong, "Multimodal object detection via probabilistic ensembling," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pp. 139–158, Springer, 2022.

[7] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, p. 2140, 2021.

[8] F. Farahnakian and J. Heikkonen, "Deep learning based multimodal fusion architectures for maritime vessel detection," *Remote Sensing*, vol. 12, no. 16, p. 2509, 2020.

[9] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," in *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, pp. 1–7, IEEE, 2019.

[10] H. Li and X.-J. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2018.

[11] Y. Bar-Shalom, F. Daum, and J. Huang, "The probabilistic data association filter," *IEEE Control Systems Magazine*, vol. 29, no. 6, pp. 82–100, 2009.

[12] B. Shahian Jahromi, T. Tulabandhula, and S. Cetin, "Real-time hybrid multi-sensor fusion framework for perception in autonomous vehicles," *Sensors*, vol. 19, no. 20, p. 4357, 2019.

[13] M.-H. Haghbayan, F. Farahnakian, J. Poikonen, M. Laurinen, P. Nevalainen, J. Plosila, and J. Heikkonen, "An efficient multi-sensor fusion approach for object detection in maritime environments," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2163–2170, IEEE, 2018.

[14] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.

[15] A. Harakeh, M. Smart, and S. L. Waslander, "Bayesod: A bayesian approach for uncertainty estimation in deep object detectors," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 87–93, IEEE, 2020.

[16] J. Choi, D. Chun, H. Kim, and H.-J. Lee, "Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 502–511, 2019.

[17] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34–45, 2019.

[18] R. Magalhães and A. Bernardino, "Quantifying object detection uncertainty in autonomous driving with test-time augmentation," *IEEE Intelligent Vehicles Symposium*, 2023.

[19] T. Minka, "Estimating a dirichlet distribution." https://vismod.media.mit.edu/pub/tpminka/papers/minka-dirichlet.ps.gz, 2000. Technical report, MIT.

[20] G. Jocher, "YOLOv5 by Ultralytics (v7.0)." https://github.com/ultralytics/yolov5, 2020. License GPL-3.0, doi: 10.5281/zenodo.3908559.

[21] Teledyne FLIR, "Teledyne FLIR Thermal Dataset for Algorithm Training." https://www.flir.com/oem/adas/adas-dataset-form/. Accessed: April 2023.

[22] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, "Multispectral fusion for object detection with cyclic fuse-and-refine blocks," in *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 276–280, IEEE, 2020.

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755, Springer, 2014.

[24] D. Feng, A. Harakeh, S. L. Waslander, and K. Dietmayer, "A review and comparative study on probabilistic object detection in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 9961–9980, 2021.