Low-cost Multispectral Scene Analysis with Modality Distillation

Heng Zhang
Univ Rennes 1, IRISA, France
heng.zhang@irisa.fr

Elisa Fromont
Univ Rennes 1, IRISA, France
elisa.fromont@irisa.fr

Sébastien Lefevre
Univ Bretagne Sud, IRISA, France
sebastien.lefevre@irisa.fr

Bruno Avignon
ATERMES company, France
bavignon@atermes.fr

Abstract

Despite its robust performance under various illumination conditions, multispectral scene analysis has not been widely deployed due to two strong practical limitations: 1) thermal cameras, especially high-resolution ones are much more expensive than conventional visible cameras; 2) the most commonly adopted multispectral architectures, two-stream neural networks, nearly double the inference time of a regular mono-spectral model which makes them impractical in embedded environments. In this work, we aim to tackle these two limitations by proposing a novel knowledge distillation framework named Modality Distillation (MD). The proposed framework distils the knowledge from a high thermal resolution two-stream network with feature-level fusion to a low thermal resolution one-stream network with image-level fusion. We show on different multispectral scene analysis benchmarks that our method can effectively allow the use of low-resolution thermal sensors with more compact one-stream networks.

1. Introduction

Vision applications such as autonomous driving or remote surveillance need to maintain high reliability under various conditions, such as insufficient illumination or adverse weather. These situations are challenging for systems using only visible cameras, which is why multispectral systems introduce additional thermal cameras to provide supplementary information. In particular, visible cameras provide visual details of colour and texture, while thermal cameras are sensitive to temperature changes, thus their contributions are complementary and their combination can ensure reliable recognition performance round-the-clock.

Under the conventional settings of multispectral scene analysis, thermal cameras and visible ones must provide image pairs with identical perception fields and identical spatial resolution. The former requirement can be achieved through camera calibration. However, due to the extreme price gap between high-resolution visible and thermal cameras[1], the requirement of identical spatial resolution usually leads to either 1) visible image downsampling that may cause information loss or 2) high manufacturing costs for thermal cameras that prevent massive production. From a practical point of view, using a high-resolution visible camera and a low-resolution thermal one would be the best compromise in performance/price.

Another constraint from the current multispectral systems lies in the software part. Nowadays, deep learning-based methods dominate the field of (multispectral) scene analysis. Multispectral information fusion methods can be categorized into: image-level fusion, feature-level fusion and decision-level fusion. Architectures that implement a feature-level fusion, usually within a two-stream network architecture (one dedicated to each source), have been proven to outperform the other strategies, and are currently the most studied in the literature [18, 13, 14, 29, 27, 32, 26, 7, 21, 8]. However, the computational overhead provided by two-stream networks is huge, which is particularly undesirable for software deployment on embedded devices.

In this paper, we propose a novel knowledge distillation framework named *Modality Distillation* (MD) to tackle the aforementioned hardware and software constraints. This framework follows two steps: Firstly, a multispectral system with high-resolution visible and thermal cameras is used to collect training data and to learn a precise but complex two-stream neural network for scene analysis. This model will be used as a teacher model with fixed weights. Secondly, a more efficient *image-level fusion* student model is trained with high-resolution visible images and *downsampled thermal images* to simulate production systems

[1] A typical thermal camera of resolution 640×480 could cost more than 8,000 USD. When the resolution is reduced to 80×60, the price becomes much more affordable (around 200 USD).

that are equipped with more economical low-resolution thermal cameras. The knowledge from the teacher model is transferred to the student model to mimic the more accurate feature-level fusion architecture and to reconstruct high-resolution thermal details. We performed extensive experiments for multispectral pedestrian detection [11] and semantic segmentation [7]. In both tasks, our model strongly reduces thermal camera requirements (resolution divided by 16) and achieves substantial inference acceleration (runtime divided by, at least, 1.8), with minor precision drop compared to full-resolution two-stream teacher networks which makes the deployment of multiple low-cost student networks on embedded devices much more viable.

The rest of this paper is organized as follows: In Section 2, we provide some representative work on multispectral scene analysis and knowledge distillation. Section 3 details the proposed method. In section 4, we conduct various experiments to study the effects of MD under different thermal resolutions and compare our results to state-of-the-art methods. Section 5 concludes the paper.

2. Related work

Our review of the related work mainly focuses on multispectral scene analysis and knowledge distillation, the two critical techniques to build the proposed framework.

2.1. Multispectral scene analysis

The first dataset for pedestrian detection from visible-thermal image pairs was introduced in [11] and then various deep learning-based methods have been proposed to tackle this problem. In [22], the authors compared, on this dataset, two different fusion strategies: *image-level fusion* (called early fusion) and *feature-level fusion* (there called late fusion because it is at the last possible feature level). Early fusion combines the information from the two modalities by directly concatenating visible and thermal images. Late fusion methods usually apply a two-stream architecture which employs two separate feature extraction networks (for visible and thermal images respectively) and combine the multispectral information by feature concatenation. [22] concluded that feature-level fusion methods produce superior performance, whereas image-level fusion ones cannot even surpass traditional methods (such as Aggregated Channel Features (ACF) [4]). To the best of our knowledge, worldwide research on image-level fusion for multispectral image analysis has been mostly interrupted since these findings. The research focus has then shifted to feature-level fusion: [18, 13] studied the optimal fusion "timing" in the detection network and came to the same conclusion that halfway feature fusion produces better results; [14] introduced an auxiliary segmentation task on the basis of halfway feature fusion for further performance improvements. [29, 27] applied attention mechanisms to adaptively weigh the visible and thermal features in the feature fusion stage; [26, 32] alleviated the inconsistency between visible and thermal features to facilitate the optimization process of a dual-modality network. Apart from these studies on feature-level fusion, multiple *decision-level fusion* methods were suggested: [6, 15] used illumination information to guide the fusion of predictions (decisions) from visible/thermal images or from day/night sub-networks; [30] discussed a confidence-aware fusion mechanism, where the disagreement between visible and thermal predictions is used to reweigh visible contributions, which could also be regarded as a decision-level fusion approach.

Compared to multispectral pedestrian detection, multispectral semantic segmentation is a younger research topic, and most related methods are based on feature-level fusion. Some relevant datasets and baseline methods were introduced in [7]. Similarly to other feature fusion methods, the baseline method also adopts two separate feature extractors for visible and thermal images respectively. Moreover, short-cut blocks were designed to concatenate the extracted multispectral feature maps. Based on this, [21] adopted stronger feature extraction networks to further boost segmentation accuracy. However, because of this complexity, the inference speed of their models was much slower. On a different type of data, [8] tackled the RGB-D semantic segmentation task, the architecture of which is also applicable for visible-thermal inputs. Therefore, we also include this model for comparison in our experiments.

Previous studies on multispectral scene analysis aimed to improve detection/segmentation accuracy, neglecting the computational cost and the effectiveness of operational deployment. In this paper, we rather exploit the (forgotten) potentialities of image-level fusion architectures, which have a similar complexity as mono-spectral networks. Moreover, we also deal with the issue of reducing thermal camera resolution, which has never been discussed in the literature despite the great interest in practical multispectral applications.

2.2. Knowledge distillation

Knowledge distillation is a method for inheriting the knowledge learnt from one or multiple pre-trained teacher models to a student model. This concept was firstly introduced in [10], where the training objective of a student model is the prediction produced by the teacher models. The use of this method strongly improved the accuracy of the student model. [20] proposed to distil features instead of predictions. Concretely, features from internal layers of a student model should mimic that of a teacher model. [12] suggested that directly parsing features as guidance might be difficult, therefore, they transferred attention maps (generated in an unsupervised manner) that function as a summary of the whole feature maps.

Apart from image classification, knowledge distillation is also widely used in other scene analysis tasks. [1] proposed the first knowledge distillation framework for object detection, which includes a distillation loss on both the feature network and the detection head. Instead of distilling all feature maps, [16, 23] applied knowledge distillation only on ROI-sampled features or features near object bounding boxes to alleviate the extreme imbalance between foreground and background. In terms of semantic segmentation, [24] distilled both the fine annotated images and unlabelled auxiliary data to regularize the training of a student segmentation model; [19] proposed to distil the structured knowledge for a semantic segmentation task.

We take inspiration from previous work and propose two specific knowledge distillation methods for multispectral scene analysis: *Attention transfer* and *Semantic transfer*. The former generates teacher attention maps (masks) via *Guided Attentive Feature Fusion* (GAFF) [27] from a two-stream teacher network. These masks are then transferred to a one-stream student network (thus performing an image-level fusion). The latter tackles the imbalance problem in feature distillation through a novel "Focal Mean Square Error" loss. Moreover, contrary to previous works where the objective is to transfer knowledge from a "larger" teacher network into a "smaller" student network (e.g., from ResNet-101 to ResNet-18), our objective is instead to transfer from a *high thermal resolution two-stream multispectral network* into a *low thermal resolution one-stream multispectral network*, while the "base" network remains unchanged (e.g., we use ResNet-18 for all experiments).

3. Modality Distillation

This section starts with an overview of the proposed *Modality Distillation* (MD) framework. We then briefly provide the basic concepts of [27] that are used in our framework. Finally, the two proposed knowledge distillation modules are described in details.

3.1. Overview

As illustrated in Fig. 1(A), the proposed *Modality Distillation* (MD) framework includes a teacher model (upper model in blue) and a student model (lower model in orange). The teacher model takes *high-resolution* multispectral image pairs as input, and employs a *two-stream architecture* consisting of: two separate feature extraction networks, a GAFF [27] module for multispectral fusion and a task-specific network for pedestrian detection/semantic segmentation. Contrarily to the teacher model, the student model uses a *low-resolution* thermal input and a *one-stream* feature extraction network that takes as input the image-level fusion of both modalities (through different input channels). We also conduct distillation experiments for a student model without the thermal modality, i.e., in this particular case, we attempt to use a multispectral teacher to improve the performance of a visible-only student.

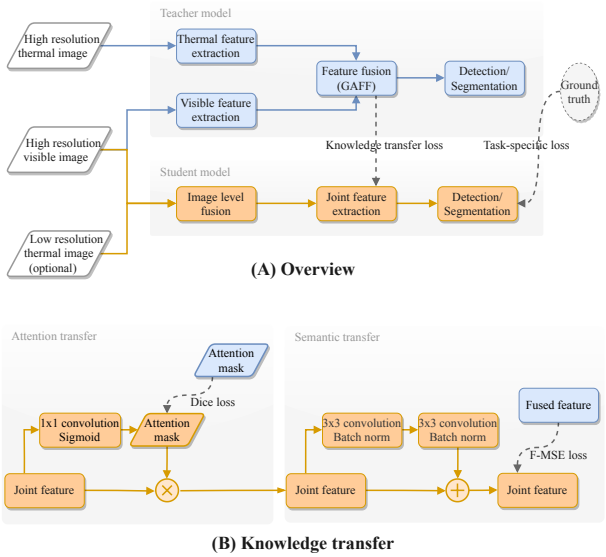

Figure 1. Overview of the proposed method (A) and details on knowledge transfer modules (B). Blue and orange blocks represent components from teacher and student models, respectively.

The proposed MD framework includes two training stages. In the first stage, we train the teacher model and fix its weights, such that the fused features from GAFF module contain the rich semantics of high-resolution thermal-visible image pairs. These features are used to guide the training of the student model; In the second stage, the optimization of the student model is supervised by a task-specific loss (e.g., pedestrian detection or semantic segmentation loss) as well as the knowledge transfer loss. The objective of the knowledge transfer loss is two-fold: *using a more efficient one-stream network to mimic a more precise two-stream network* and *using the more available low-resolution thermal images to reconstruct high-resolution thermal details*. Finally, we obtain a student model that takes low-resolution thermal images as input, and the required parameters and calculations are greatly reduced. Meanwhile, the precision of the low thermal resolution one-stream student model is supposed to be close to the high thermal resolution two-stream teacher model.

3.2. Guided attentive feature fusion

We carefully follow the implementations of GAFF module from [27] for multispectral feature fusion in the teacher model. GAFF learns the intra- and inter-modality attention masks to adaptively enhance important areas and to identify reliable modalities, respectively. Intra-modality attention is used to distinguish the foreground and background for each individual modality, and the inter-modality attention is used to select features among visible and thermal modalities ac-

cording to the dynamic comparison of their prediction quality of the former mask.[2]. The weighted visible and thermal features are obtained via:

$$f^{visible}_{weighted} = f^{visible} \otimes (1 + m^{visible}_{intra}) \otimes (1 + m^{visible}_{inter})$$
$$f^{thermal}_{weighted} = f^{thermal} \otimes (1 + m^{thermal}_{intra}) \otimes (1 + m^{thermal}_{inter}) \quad (1)$$

where $f^{visible}$ and $f^{thermal}$ denote features from visible and thermal feature extraction branches. m_{intra} and m_{inter} are predicted intra- and inter-modality attention masks from GAFF. Their superscript indicates the modality.

The fusion of the weighted visible and thermal features are assigned as the teacher features:

$$f_{teacher} = \frac{f^{visible}_{weighted} + f^{thermal}_{weighted}}{2} \quad (2)$$

3.3. Knowledge transfer modules

To preserve the knowledge learnt from the teacher model to the maximum extent, we apply two knowledge transfer strategies: Attention transfer that *guides the one-stream model to mimic the two-stream attentive fusion*, and Semantic transfer that *rebuilds high-resolution visual details from low-resolution thermal images.*

Attention transfer. GAFF significantly improves the scene analysis performance in a two-stream teacher model. However, such a multispectral feature fusion module does not exist in a one-stream student model. Thus, as illustrated in the left part of Fig. 1(B), we design the *Attention transfer* module to simulate this attentive fusion in a one-stream model. The teacher attention mask is the combination of intra- and inter-modality attention masks. To keep the architecture simple, the student attention mask is generated by a 1×1 convolution followed by a Sigmoid activation, and is supervised by minimizing the Dice loss [3] between the student and teacher attention masks:

$$m_{teacher} = m^{visible}_{intra} \otimes m^{visible}_{inter} + m^{thermal}_{intra} \otimes m^{thermal}_{inter}$$
$$m_{student} = \mathcal{F}(f_{student})$$
$$L_{attention} = 1 - \frac{2|m_{student} \otimes m_{teacher}|}{|m_{student}| + |m_{teacher}|} \quad (3)$$

where $L_{attention}$ denotes the *Attention transfer* loss; $m_{teacher}$ and $m_{student}$ represent the teacher and student attention masks respectively; $f_{student}$ denotes the student features acquired from the joint feature branch; $\mathcal{F}$ represents a 1×1 convolution followed by a Sigmoid activation; $\otimes$ and $||$ represent respectively the pixel-wise multiplication and summation operation.

[2]Due to space constraint, we refer the reader to [27] for more details.

Semantic transfer. To compensate for the resolution reduction of thermal input images, the *Semantic transfer* module performs an implicit super-resolution of student feature maps. As shown in the right part of Fig. 1(B), we use a basic residual block [9] to increase the details in the joint features. *Semantic transfer* aims to minimize the distance between the student (joint) and teacher (fused) feature maps. However, optimizing this distance has proven to be difficult. At first glance, this is due to the extreme imbalance between the foreground and background areas [16, 23]. Inspired by the Focal loss [17], *we argue that the true problem lies in the extreme imbalance between easily-mimic and hardly-mimic areas*. Therefore, we propose the Focal Mean Square Error (F-MSE) loss defined as:

$$d = (f_{student} - f_{teacher})^2$$
$$L_{semantic} = \sum_w \sum_h \frac{1}{n} (\delta(\sum_n d) \times \sum_n d) \quad (4)$$

where $L_{semantic}$ denotes the *Semantic transfer* loss; d is the squared L2 distance between student and teacher feature maps. δ signifies the Softmax function; w, h, n represent the width, height and depth of feature maps, respectively.

The major difference between the proposed F-MSE loss and the standard MSE loss (used in [20]) is the spatial re-weighting based on feature-mimicking errors. Concretely, the Softmax function generates a 2-D re-weighting mask, where each value reflects the difficulty of feature mimicking on a specific area, and the summation of all values on the mask is equal to 1. In such a manner, the optimization adaptively "focuses" on mis-predicted areas, and the imbalance problem is therefore solved.

4. Experiments

In this section, we conduct experiments on two multispectral benchmarks [11, 7] and two scene analysis tasks (pedestrian detection and semantic segmentation) to evaluate the effectiveness of the proposed *Modality Distillation* (MD) framework.

4.1. Datasets

KAIST. The KAIST multispectral pedestrian detection dataset [11] (denoted as KAIST Dataset) focuses on the *pedestrian detection* task based on aligned multispectral image pairs. These image pairs are collected during daytime and nighttime. This dataset contains 21,622 annotated image pairs for training, and 2,252 image pairs for testing. Due to the problematic annotations from the original dataset, we adopt the improved annotations proposed by [30] and [18] for training and evaluation, respectively. Following previous works, we adopt the log-average Miss Rate (computed by averaging the miss rate on false positive per-image points sampled within the range of $\left[10^{-2}, 10^{0}\right]$, lower is better)

under a "reasonable" setting [5] (pedestrians that are occluded or shorter than 55 pixels are eliminated) as the evaluation metric.

MFNet. The Multispectral semantic segmentation dataset [7] (denoted as MFNet Dataset) targets the *semantic segmentation* of street scenes for autonomous vehicles. The segmentation labels consist of eight classes: car, person, bike, curve, car stop, guardrail, colour cone and bump. It provides 1,568 aligned multispectral image pairs in the training set, 392 pairs in the validation set and 393 pairs in the test set. Among each subset, half of the image pairs are taken during daytime, and the other half during nighttime. To evaluate the segmentation accuracy, we report the class-wise Mean Accuracy, calculated by averaging the ratio between the number of true positive pixels and the sum of true positive and false negative pixels for each class.

4.2. Implementation details

Network architecture. For all experiments, we apply ResNet-18 [9] as the feature extraction network, RetinaNet [17] (its label assignment is optimized via Mutual Guidance strategy [25]) as the pedestrian detection network and PSPNet [31] as the semantic segmentation network. For the teacher model (Fig. 1(A) upper model), GAFF [27] is adopted for the attentive fusion of visible and thermal features. For the student model (Fig. 1(A) lower model), visible and thermal images are concatenated to generate 6-channel multispectral inputs, i.e., 3 channels from each modality. The first convolution layer is modified to suit the 6-channel input[3]. Note that in lieu of generating 4-channel input as done in [22], we duplicate the single-channel thermal images into 3 channels to balance the contribution of visible and thermal spectrum in the first convolution layer.

Input resolution. The resolution of visible input images are set identical to previous methods for fair comparisons. More specifically, the resolution is 640×512 on KAIST Dataset and 640×480 on MFNet Dataset. To simulate the low-resolution thermal camera in actual products, we downsample the high-resolution thermal images (e.g., from 640×512 to 160×128) through bilinear interpolation. These downsampled small thermal images are then re-scaled to the original spatial size (e.g., from 160×128 back to 640×512) to concatenate with the visible images. Note that the high-resolution thermal details are already lost in the first interpolation operation. Considering the camera price and the containing number of pixels, 16 times thermal resolution downsampling is regarded as the most practical case (e.g., downsampling from 640×512 to 160×128).

[3]Concretely, the pretrained ImageNet [2] weights for the first convolution layer are duplicated along the input channel dimension, and the values are halved.

Fusion stage	Miss Rate (%)			Runtime
	Day	Night	All	
Visible-only	16.95	35.15	22.84±0.77	**6.48ms**
Image-level	10.73	6.61	9.40±0.39	6.55ms
Feature-level	9.37	4.71	**7.77±0.07**	10.97ms
Decision-level	10.74	9.25	10.34±0.32	12.96ms

Table 1. Different fusion methods on KAIST Dataset. For fair comparisons, all listed methods use the same feature extraction network (ResNet-18) and detection network (RetinaNet).

Fusion stage	Mean Accuracy (%)			Runtime
	Day	Night	All	
Visible-only	50.83	52.26	55.06±0.21	**4.57ms**
Image-level	54.97	56.07	59.42±0.22	4.68ms
Feature-level	57.21	62.18	**63.45±0.24**	8.94ms
Decision-level	51.72	53.37	56.21±0.14	9.14ms

Table 2. Different fusion methods on MFNet Dataset. For fair comparisons, all listed methods use the same feature network (ResNet-18) and segmentation network (PSPNet).

Training details. All models (including teacher and student models) are trained on a single GPU with 16 multispectral image pairs per mini-batch, and with an initial learning rate of 1e-2. To stabilize the training at the beginning, we adopt the warm-up strategy, where the learning rate is linearly increased from 1e-6 to 1e-2 within the first 500 iterations, then decayed with a cosine annealing. Pedestrian detection models are trained for 3,500 iterations, and semantic segmentation models are trained for 14,000 iterations. The whole project is coded in PyTorch 1.20. For fair runtime comparisons, all models' runtimes are measured on an Nvidia GTX 1080Ti GPU. We repeat each training 3 times with different random seed values and report the average performance as well as the standard error (only for the "all" setting because of the space restriction). The best results are shown in bold.

4.3. Experimental results

Baseline results. Image-level, feature-level and decision-level are the three major fusion methods for multispectral scene analysis. We list in Tab. 1 and 2 their prediction accuracy and inference time on KAIST Dataset [11] and on MFNet Dataset [7], respectively. The visible-only results are also listed for reference. In order to fairly compare these fusion methods, we use the same feature network (ResNet-18 [9]) and detection/segmentation network (RetinaNet [17]/PSPNet [31]). Specifically, we adopt GAFF [27] as the feature-level fusion method. For simplicity, we average the prediction from visible and thermal images for decision-level fusion. The tables show that, regardless of the information fusion stage (image-level, feature-level or decision-level), multispectral methods greatly improve the detection/segmentation accuracy

Thermal resolution	MD	Miss Rate (%)		
		Day	Night	All
Full resolution		10.73	6.61	9.40±0.39
	✓	9.45	4.61	**7.78±0.28**
4x downsample		11.57	6.51	9.84±0.72
	✓	9.39	5.07	**7.91±0.11**
16x downsample		12.09	6.73	10.17±0.42
	✓	9.85	4.84	**8.03±0.19**
64x downsample		14.92	10.66	13.37±0.30
	✓	10.75	7.07	**9.50±0.06**
Visible-only		16.95	35.15	22.84±0.77
	✓	14.74	34.13	**21.08±0.21**

Table 3. Comparison between native models and distilled models on KAIST Dataset under different thermal resolution settings (from full thermal resolution to no thermal scenario). All listed methods use an image-level fusion architecture.

Thermal resolution	MD	Mean Accuracy (%)		
		Day	Night	All
Full resolution		54.97	56.07	59.42±0.22
	✓	59.71	62.78	**64.93±0.11**
4x downsample		53.89	55.18	57.93±0.27
	✓	58.32	61.88	**64.25±0.11**
16x downsample		53.85	55.43	58.21±0.46
	✓	58.46	61.37	**63.52±0.87**
64x downsample		53.68	53.68	57.06±0.18
	✓	57.58	59.67	**62.62±0.81**
Visible-only		50.83	52.26	55.06±0.21
	✓	57.74	56.67	**60.62±0.42**

Table 4. Comparison between native models and distilled models on MFNet Dataset under different thermal resolution settings (from full thermal resolution to no thermal scenario). All listed methods use an image-level fusion architecture.

compared to the visible-only model, especially for nighttime detection/segmentation. Feature-level and decision-level fusion methods almost double the execution runtimes (as well as the number of parameters, not reported here because of the space restriction) of a visible-only model. In contrast, the computational overhead for the image-level fusion is negligible, which shows the relevance of this fusion method when fewer computational resources are available.

Distillation results. We list in Tab. 3 and 4 the comparisons between native image-level fusion models and distilled image-level fusion models (i.e., student models) on KAIST Dataset [11] and MFNet Dataset [7], respectively. It can be observed that MD strategy brings important improvements for all thermal resolutions for both datasets.

Specifically, on the multispectral pedestrian detection task (Tab. 3), our full thermal resolution result with MD is already close to that of the feature-level fusion model (i.e., teacher model) shown in Table 1 (7.78% versus 7.77%), while the inference time is almost halved (10.97ms versus 6.55ms). When it comes to the most practical case where thermal resolution is 16 times lower than visible resolution, MD strategy brings 2.14% of Miss Rate improvement, and

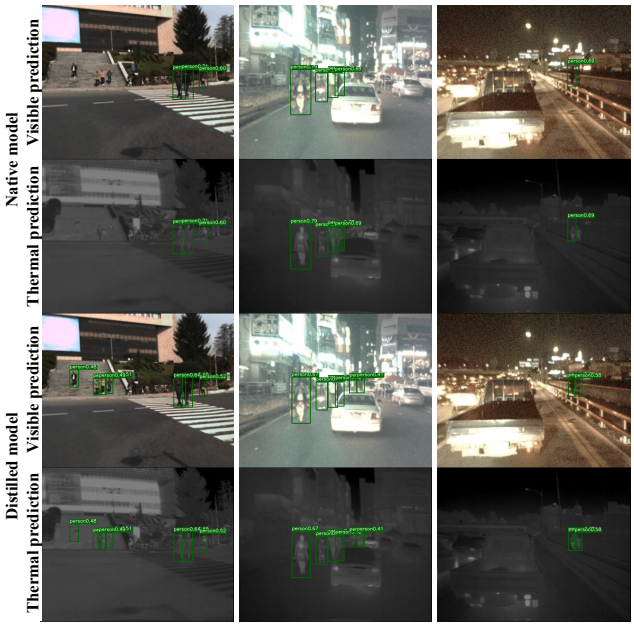

Figure 2. Visual improvements on KAIST Dataset.

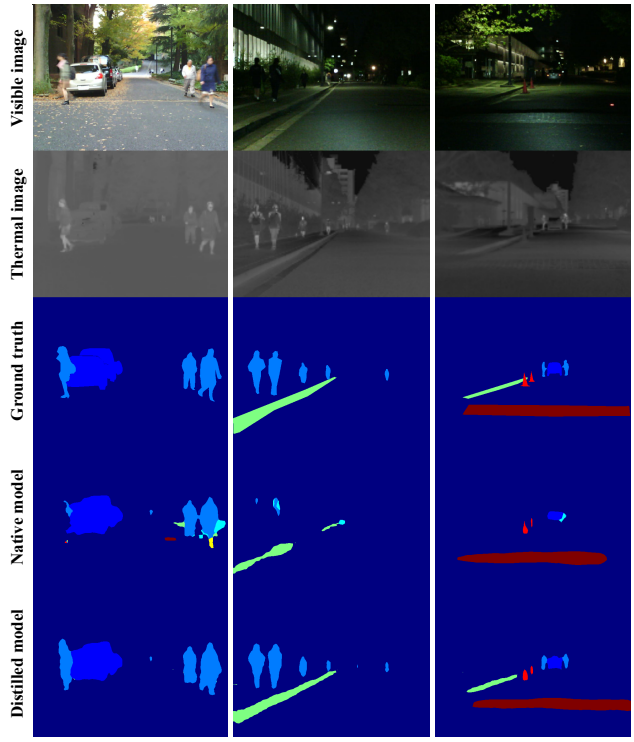

Figure 3. Visual improvements on MFNet Dataset.

the performance difference compared to the teacher model is only 0.26% (8.03% versus 7.77%). We show some detection results from native model and distilled model for this practical case in Fig. 2, and it can be observed that our distilled model provides more precise detection results. Interestingly, the nighttime detection precision is boosted by 27.06% (7.07% versus 34.13%) even if the thermal resolution is reduced to 80×64 (i.e., 64 times downsampled), proving the necessity of the thermal modality in nighttime detections. Moreover, our strategy remains helpful when the thermal image is completely removed (e.g., the Miss Rate for visible-only model is reduced from 22.84% to 21.08%). Here the multispectral knowledge from the teacher model allows the visible-only student to perform pseudo-multispectral detection, which is the main reason of improvements.

On the multispectral semantic segmentation task, the improvements are more important (around 5% for all thermal resolutions using MD). It is noteworthy that the performance of the distilled visible-only model is even better than that of the native full-resolution image-level fusion model (60.62% versus 59.42%). Here, our assumption is that the multispectral semantic segmentation task is more critical for the choice of fusion architecture, e.g., according to Tab. 2, native image-level fusion performs 4.03% worse than feature-level fusion. This may be the reason why rare previous work use image-level fusion for multispectral semantic segmentation. However, our MD strategy makes the student model mimic a feature-level fusion teacher model, which compensates its established disadvantage of image-level fusion architecture, and thus brings tremendous accuracy improvements. For the practical case (16 times thermal resolution downsampling), the mean accuracy difference with the teacher model is minor (63.52% versus 63.45%). We visualize in Fig. 3 the segmentation results from the native model and our distilled model for the practical case, and it could be noted that the improvement from MD strategy on segmentation quality is obvious.

Comparing with state-of-the-art. We compare the results of our distilled models (which adopt the more efficient image-level fusion) with state-of-the-art methods (all adopting the more cumbersome feature-level fusion) on KAIST Dataset (Tab. 5) and MFNet Dataset (Tab. 6). Note that our teacher models use [27] and this method has already been shown to give better results than its competitors. However, our goal here is to show how our student models (which are supposed to be less good than their teachers) perform compared to their competitors. Specifically, we provide our one-stream student models' results using full thermal resolution (same condition as our competitors, denoted as "full") and using 16 times downsampled thermal resolution (denoted as "practical"). We also list our two-stream teacher models' results (denoted as "teacher") for reference. We consider "practical" the most interesting setting for actual multispectral applications.

Method	Miss Rate (%)			Runtime
	Day	Night	All	
ACF [11]	42.57	56.17	47.32	2730ms
Halfway Fusion [18]	24.88	26.59	25.75	430ms
FusionRPN+BF [13]	19.57	16.27	18.29	800ms
IAF R-CNN [15]	14.55	18.26	15.73	210ms
IATDNN+IASS [6]	14.67	15.72	14.95	250ms
RFA [28]	16.78	10.21	14.61	80ms
CIAN [29]	14.77	11.13	14.12	70ms
MSDS-RCNN [14]	10.53	12.94	11.34	220ms
AR-CNN [30]	9.94	8.38	9.34	120ms
MBNet [32]	8.28	7.86	8.13	70ms
Ours (teacher)	9.37	4.71	**7.77**	11ms
Ours (full)	9.45	4.61	7.78	**7ms**
Ours (practical)	9.85	4.84	8.03	**7ms**

Table 5. Comparison between state-of-the-art multispectral pedestrian detection methods and ours on KAIST Dataset. Our competitors' results are taken from [32].

Method	Mean Accuracy (%)			Runtime
	Day	Night	All	
MFNet [7]	42.6	41.4	45.1	4.35ms
FuseNet [8]	49.5	48.9	52.4	**3.92ms**
RTFNet-50 [21]	57.3	59.4	62.2	11.25ms
RTFNet-152 [21]	60.0	60.7	63.1	29.35ms
Ours (teacher)	57.2	62.2	63.5	8.94ms
Ours (full)	59.7	62.8	**64.9**	4.92ms
Ours (practical)	57.6	61.4	63.5	4.92ms

Table 6. Comparison between state-of-the-art multispectral semantic segmentation methods and ours on MFNet Dataset. Our competitors' results are taken from [21].

On the multispectral pedestrian detection task, the achieved Miss Rate from the distilled "practical" model is already better than that of the best feature-level fusion methods in the literature [32] (8.03% versus 8.13%). It should be noted that our "practical" model takes downsampled thermal images as input and adopts a much simpler architecture (one-stream networks for "full"/"practical" and two-stream networks for others). It is also worth noting that our nighttime detection performance surpasses all previous methods, which proves that the thermal information has been well-preserved in the student model.

On the multispectral semantic segmentation task, thanks to the substantial accuracy improvements from MD (about 5%), our distilled "practical" model also surpasses the best previous result [21] (63.5% versus 63.1%). For this dataset as well, all our trained models (including the "practical" model with downsampled thermal input) have obvious ad-

S(M)	S(F)	A	Miss Rate (%)		
			Day	Night	All
			12.09	6.73	10.17±0.42
✓			11.92	6.69	10.09±0.06
	✓		10.24	6.93	9.11±0.15
		✓	10.61	5.83	8.99±0.23
	✓	✓	9.85	4.84	**8.03±0.19**

Table 7. Ablation experiments on KAIST Dataset. We study the effects of Semantic transfer (with MSE or F-MSE loss) and Attention transfer modules in the proposed MD framework.

S(M)	S(F)	A	Mean Accuracy (%)		
			Day	Night	All
			53.85	55.43	58.21±0.46
✓			56.04	57.87	60.41±0.28
	✓		57.49	58.47	61.16±0.13
		✓	57.55	58.46	61.51±0.30
	✓	✓	57.58	61.37	**63.52±0.87**

Table 8. Ablation experiments on MFNet Dataset. We study the effects of Semantic transfer (with MSE or F-MSE loss) and Attention transfer modules in the proposed MD framework.

vantage in nighttime prediction. It should be pointed out that both our distilled models with one-stream ResNet-18 feature network even outperform RTFNet-152 [21] with two-stream ResNet-152 feature network, demonstrating the high efficiency of our distilled models (ours are about 6 times faster than RTFNet-152). More surprisingly, we can see in Tab. 6 that the full student model gives slightly better results than the teacher model. The student model's feature extraction network is the same as the teacher's one, so the student could theoretically achieve similar performance, and the student model has more sources of supervision, i.e., the ground truth and the knowledge learnt from the teacher model, which we believe is the reason for the higher performance shown by the student model.

Ablation experiments. To explore the effects of the proposed *Attention transfer* and *Semantic transfer* modules, we conduct ablation experiments on KAIST Dataset (Tab. 7) and MFNet Dataset (Tab. 8), under the most practical case (thermal images are 16 times downsampled). The "S" and "A" denote *Semantic transfer* and *Attention transfer* modules as illustrated in Fig. 1(B) right and left parts, respectively. We conduct comparative experiments between the traditional MSE loss (denoted as "M") and the proposed F-MSE loss (denoted as "F") in the *Semantic transfer* module. According to our experimental results, the latter provides better performance for both tasks. Moreover, we visualize some examples of the 2-D spatial re-weighting mask from F-MSE loss (Eq.4) in Fig. 4, and it can be observed that: 1) the imbalance between easily-mimic and hardly-mimic area is grave, where the former occupies most of a given image;

Figure 4. Visualization of the visible-thermal image pairs and the 2-D spatial re-weighting masks from the proposed F-MSE loss. The first two lines of multispectral images pairs come from KAIST Dataset, and the last two lines come from MFNet Dataset.

2) with F-MSE loss, the optimization on the feature mimicking is automatically "focused" on more important areas, e.g., pedestrians, vehicles and colour cones. This specific loss tackles the imbalance problem in the *Semantic transfer* module. In conclusion, according to our ablation experiments on two datasets, both the proposed *Semantic transfer* and *Attention transfer* modules bring notable improvements and their combination leads to the best performance.

5. Conclusion

In this paper, we identify the hardware and software constraints in today's multispectral scene analysis systems and propose a novel *Modality Distillation* framework to tackle these constraints. Specifically, this framework distils the knowledge from a high thermal resolution two-stream network with feature-level fusion to a low thermal resolution one-stream network with image-level fusion. The distilled model could perform prediction on widely available low-resolution thermal cameras and shows similar complexity with the mono-spectral models. In this framework, we present two knowledge transfer modules named *Attention transfer* and *Semantic transfer* specifically for multispectral learning. Extensive experiments for multispectral pedestrian detection and semantic segmentation demonstrate the efficiency of the proposed framework. In the future, we plan to extend our method for scene analysis with even more modalities such as depth sensor, Doppler radar, LiDAR, etc.

References

[1] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 742–751, 2017.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[3] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[4] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE transactions on pattern analysis and machine intelligence*, 36(8):1532–1545, 2014.

[5] 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.

[6] 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.

[7] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multispectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017.

[8] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian conference on computer vision*, pages 213–228. Springer, 2016.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[11] 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.

[12] Nikos Komodakis and Sergey Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

[13] Daniel Konig, Michael Adam, Christian Jarvers, Georg Layher, 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.

[14] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Multispectral pedestrian detection via simultaneous detection and segmentation. In *British Machine Vision Conference (BMVC)*, September 2018.

[15] 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.

[16] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 6356–6364, 2017.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[18] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N Metaxas. Multispectral deep neural networks for pedestrian detection. *arXiv preprint arXiv:1611.02644*, 2016.

[19] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019.

[20] Adriana Romero, Samira Ebrahimi Kahou, Polytechnique Montréal, Y. Bengio, Université De Montréal, Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *in International Conference on Learning Representations (ICLR*, 2015.

[21] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, 2019.

[22] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *ESANN*, volume 587, pages 509–514, 2016.

[23] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4933–4942, 2019.

[24] Jiafeng Xie, Bing Shuai, Jian-Fang Hu, Jingyang Lin, and Wei-Shi Zheng. Improving fast segmentation with teacher-student learning. *arXiv preprint arXiv:1810.08476*, 2018.

[25] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Localize to classify and classify to localize: Mutual guidance in object detection. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[26] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 276–280. IEEE, 2020.

[27] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Guided attentive feature fusion for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 72–80, 2021.

[28] Lu Zhang, Zhiyong Liu, Xiangyu Chen, and Xu Yang. The cross-modality disparity problem in multispectral pedestrian detection. *arXiv preprint arXiv:1901.02645*, 2019.

[29] 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.

[30] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5127–5137, 2019.

[31] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[32] Kailai Zhou, Linsen Chen, and Xun Cao. Improving multispectral pedestrian detection by addressing modality imbalance problems. *arXiv preprint arXiv:2008.03043*, 2020.