# Boosting Cross-spectral Unsupervised Domain Adaptation for Thermal Semantic Segmentation

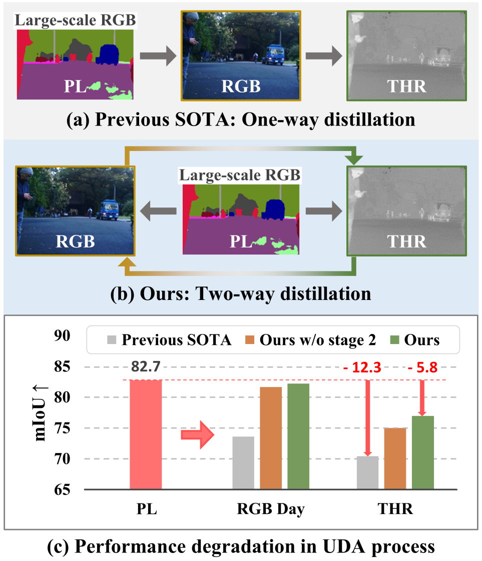
Seokjun Kwon[1*], Jeongmin Shin[1*], Namil Kim[2], Soonmin Hwang[3], Yukyung Choi[1†]

***Abstract*— In autonomous driving, thermal image semantic segmentation has emerged as a critical research area, owing to its ability to provide robust scene understanding under adverse visual conditions. In particular, unsupervised domain adaptation (UDA) for thermal image segmentation can be an efficient solution to address the lack of labeled thermal datasets. Nevertheless, since these methods do not effectively utilize the complementary information between RGB and thermal images, they significantly decrease performance during domain adaptation. In this paper, we present a comprehensive study on cross-spectral UDA for thermal image semantic segmentation. We first propose a novel masked mutual learning strategy that promotes complementary information exchange by selectively transferring results between each spectral model while masking out uncertain regions. Additionally, we introduce a novel prototypical self-supervised loss designed to enhance the performance of the thermal segmentation model in nighttime scenarios. This approach addresses the limitations of RGB pre-trained networks, which cannot effectively transfer knowledge under low illumination due to the inherent constraints of RGB sensors. In experiments, our method achieves higher performance over previous UDA methods and comparable performance to state-of-the-art supervised methods.**

## I. INTRODUCTION

In recent years, there has been a significant increase in research on robust semantic segmentation techniques for challenging environments. This surge is driven by the critical need for reliable performance in autonomous driving, as it directly impacts human safety. Nevertheless, previous approaches [15], [26], [27] relying solely on visual cues from RGB sensors often struggle in adverse scenarios such as low-light conditions, dense fog, and heavy rain. To tackle this issue, thermal sensors that capture the heat signatures of objects have been extensively utilized for achieving reliable semantic segmentation in challenging conditions [17], [18].

However, several challenges must be overcome in thermal image semantic segmentation to improve performance gain. Specifically, in these methods training the model is impeded by the lack of large-scale datasets, where pixel-level annotations are labor-intensive and costly to obtain. Moreover, thermal images often suffer from low quality such as textureless and low resolution, which could harm model training and performance gain despite the advantages of thermal cameras in low-illumination conditions. To tackle these issues, a few unsupervised domain adaptation (UDA) methods [14] have emerged in cross-spectral domains. The main idea of MS-UDA [14] is to leverage the knowledge learned from a large-scale RGB dataset with segmentation labels, which serves as the source domain, and transfer this knowledge to the thermal image domain where manual labels are not available, as illustrated in Fig. 1-(a).



(a) Previous SOTA: One-way distillation

(b) Ours: Two-way distillation

(c) Performance degradation in UDA process

Fig. 1. To handle the labeled data scarcity problem in the thermal domain, we employ a pre-trained model on a large-scale RGB dataset [2] to train student networks. (a) Previous SOTA method [14] performs one-way distillation from PL to THR, disregarding the characteristics of each spectral domain. (b) Our approach adopts two-way distillation, which appropriately transfers the complementary knowledge of each spectral domain. (c) To validate the potential of the distillation processes, we evaluate the performance of the teacher (i.e., PL) and each spectral student network. Our final method outperforms the previous SOTA method in both RGB and thermal domains despite leveraging the same pseudo-labels at the training phase. Due to the limitations of the RGB sensor at nighttime, previous SOTA and our method leverage only daytime RGB images for the distillation process (i.e., RGB Day). (PL: pseudo-labels, THR: thermal)

However, one limitation of MS-UDA lies in its reliance on a one-way knowledge distillation process that involves training an RGB student network using a pre-trained teacher model on the large-scale RGB dataset. The outputs of the RGB network are then used as training labels for the thermal network. Since the student network often inherits equal or lower performance relative to the teacher, this one-way distillation strategy restricts the model's potential for achieving higher performance, as depicted in Fig. 1-(c) In addition, the knowledge distillation process of MS-UDA relies solely on distilling the final predicted values across spectral domains. This disregard for domain-specific features

[1] Seokjun Kwon, Jeongmin Shin, and Yukyung Choi are with the Sejong University, South Korea {sjkwon, jmshin, ykchoi}@rcv.sejong.ac.kr

[2] Namil Kim is with NAVER LABS, South Korea namil.kim@naverlabs.com

[3] Soonmin Hwang is with the Department of Automotive Engineering, Hanyang University, South Korea soonminh@hanyang.ac.kr

*: Equal Contribution, †: Corresponding Author

inherent in RGB and thermal data impedes successful cross-spectral adaptation. Moreover, MS-UDA generates day-to-night synthetic thermal images for domain generalization in thermal images due to the inaccuracy of pseudo-labels from the RGB network in low-light conditions. These fake night images, however, might not fully reflect the whole distribution of real nighttime scenes, leading to limited performance improvements.

In this paper, we present novel learning and loss strategies that significantly improve the performance of cross-spectral unsupervised domain adaptation framework for thermal image semantic segmentation. We first propose a novel masked mutual learning that distills the knowledge learned from a large-scale RGB dataset in a two-way, as depicted in Fig. 1-(b). Our proposed strategy encourages the RGB and thermal student network to learn complementary information from each spectral knowledge by filtering the uncertain pixels. Furthermore, our novel prototypical self-supervised loss enables training of the model when the pseudo-labels provided by the teacher network are unreliable due to challenging conditions such as poor illumination at night. Our method outperforms recent UDA methods on both MF [4] and KP [3] datasets. Furthermore, we achieve robust performance compared with state-of-the-art supervised methods.

## II. RELATED WORK

### A. Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) transfers knowledge acquired from a large-scale source domain to a target domain where labeled data are scarce. Many UDA methods for semantic segmentation leveraged adversarial learning frameworks [21]. These adversarial approaches [22], [23], [24] involved training model to generate features that could deceive a discriminator, effectively aligning the feature distributions of source and target domains. Recently, self-training approaches [25], [19], [11] have emerged to incrementally improve the pseudo-labels.

While most existing works consider UDA between two domains of the same modality (e.g., RGB), MS-UDA [14] proposed a UDA technique between the RGB-thermal domains to address the data scarcity issue in the thermal image dataset. To solve this issue, MS-UDA proposed one-way knowledge distillation strategies to train the thermal image semantic segmentation: i) large-scale RGB-to-RGB inter-domain adaptation, ii) RGB-to-thermal, and iii) thermal-to-thermal intra-domain adaptation. However, this distillation process simply transfers teacher predictions to the student without considering the characteristics of each domain, and the performance degradation for each distillation stage is significant. For effective UDA in cross-spectral domains, we design a novel UDA framework that leverages the complementary knowledge of the RGB-thermal domains by filtering out the inherent drawbacks associated with each spectral domain.

## III. METHOD

### A. Overview

**Overall Architecture.** Our goal is to train the thermal image semantic segmentation network while mitigating the spectral domain discrepancies. To perform this, we introduce several enhancements that are employed exclusively during the training phase to ensure persistent efficiency during inference. As illustrated in Fig. 2-(a), our method consists of three parts: spectral-specific encoder, weight-shared decoder, and prototype-fusion module. Specifically, an RGB image $x_R$ and a thermal image $x_T$ are fed into a spectral-specific encoder $E_\theta$, where $\theta = \{R, T\}$. The encoder $E_\theta$ extracts the multi-scale encoder feature map and then each spectral feature map is passed through a weight-shared decoder $D_s$ with a skip connection between encoder and decoder features, producing the multi-scale decoder features and segmentation output. We then calculate cross-entropy loss $L_{seg}$ between the pseudo-labels $y_{PL}$ and the prediction of each spectral stream $P_\theta$ as follows:

$$\begin{aligned} L_{seg} &= L_R + L_T \quad \text{where} \\ L_\theta &= -\frac{1}{N}\sum_{p=1}^{H\times W}\sum_{c=1}^{C} y_{PL}^{(p,c)} \log\left(P_\theta^{(p,c)}\right) \end{aligned} \tag{1}$$

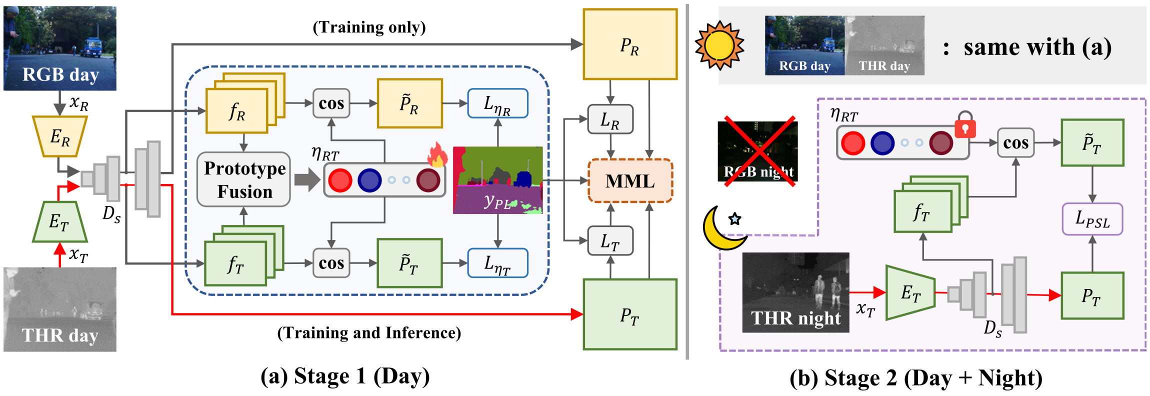
where H, W, N, and C denote the height, width, the number of pixels, and classes. Pseudo-labels $y_{PL}$ are generated by HRNet [26] pre-trained on a large-scale RGB dataset [2], as done in MS-UDA.

**Cross-Spectral Prototypes.** In addition, inspired by previous methods [19], [11], [20], our UDA framework leverages prototypes to incorporate semantic information at the pixel level. First, we generate cross-spectral prototypes that share the latent space of two spectral domains. Specifically, these cross-spectral prototypes $\eta_{RT}$ are generated by using intermediate decoder features $f_\theta \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}}$ as follows:

$$\begin{aligned} \eta_\theta^{(c)} &= \frac{\sum_{p\in\Omega} f_\theta^{(p)} \odot \mathbb{1}[y_{PL}^{(p,c)} = 1]}{\sum_{p\in\Omega} \mathbb{1}[y_{PL}^{(p,c)} = 1]} \\ \eta_{RT}^{(c)} &= \frac{\eta_R^{(c)} + \eta_T^{(c)}}{2} \end{aligned} \tag{2}$$

where $\Omega$ denotes a set of pixels that has a higher confidence score than 0.1, $\odot$ is an element-wise multiplication, $\mathbb{1}[\cdot]$ is an indicator function and $y_{PL}^{(p,c)}$ is one-hot labels, i.e., 1 if the class label at position $p$ corresponds to $c$ and 0 otherwise. These prototypes are considered approximated representational centroids shared by both spectral representations for each class. We also leverage the cross-spectral prototypes to perform contrastive learning for each spectral domain. Specifically, we encourage each pixel of the $f_\theta^{(p)}$ to be attracted toward prototypes of the same class while pushing away different classes. We calculate prototypical contrastive loss $L_\eta$ as follows:

$$\begin{aligned} L_\eta &= L_{\eta_R} + L_{\eta_T} \quad \text{where} \\ L_{\eta_\theta} &= -\sum_{p=1}^{H\times W}\sum_{c=1}^{C} y_{PL}^{(p,c)} \log \frac{\exp(s(f_\theta^{(p)}, \eta_{RT}^{(c)})/\tau)}{\sum_c \exp(s(f_\theta^{(p)}, \eta_{RT}^{(c)})/\tau)} \end{aligned} \tag{3}$$



(a) Stage 1 (Day)

(b) Stage 2 (Day + Night)

Fig. 2. An overview of our framework. (a) In stage 1, both RGB and thermal networks are trained in the daytime using pseudo-labels $y_{PL}$ generated by HRNet [26] pre-trained on a large-scale RGB dataset [2]. Simultaneously, Masked Mutual Learning (MML) is applied between these student networks, and cross-spectral prototypes $\eta_{RT}$ are gradually updated during training time. (b) In stage 2, the same learning process is performed for daytime as in (a). We impose prototypical self-supervised loss $L_{PSL}$ using our cross-spectral prototypes to address the absence of reliable annotations for nighttime training.

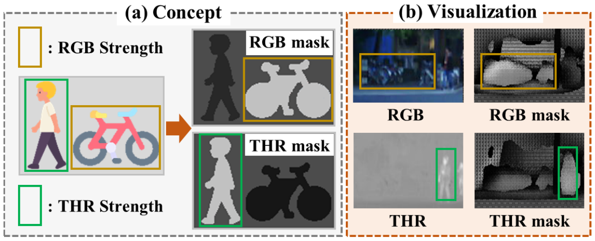


Fig. 3. We present a conceptual illustration of masks in Masked Mutual Learning (MML) (a), and the visualization results (b). (a) Our proposed MML explicitly incorporates the unique characteristics of each spectral domain during the training process. We leverage RGB and THR masks to extract complementary information from cross-spectral images. Simultaneously, these masks exploit the strengths of each domain while mitigating its inherent limitations (i.e., person in RGB and bicycle in THR). (b) The visualization results of our masks during training.

where $\tau$ and $s(\cdot,\cdot)$ refer to the temperature and cosine similarity, respectively. $\tau$ is set to 1. By leveraging cross-spectral prototypes, our UDA approach achieves effective incorporation of pixel-wise semantic information. Moreover, cross-spectral prototypes are progressively updated using an exponential moving average with a momentum parameter of 0.9 during the training to capture richer semantic knowledge.

### *B. Boosting Cross-Spectral UDA for Thermal Segmentation*

**Masked Mutual Learning.** RGB and thermal cameras exhibit complementary strengths, with RGB sensors providing rich visual information such as color and texture, while thermal sensors facilitate robust perception in low-light environments. Based on this aspect, we propose Masked Mutual Learning (MML), which encourages spectral-specific mutual learners to share their advantageous knowledge. The main idea of MML is to maximize the complementary information between cross-spectral images during training by eliminating the inherent drawback within each spectral feature. We leverage each spectral-wise mask in the MML strategy, as depicted in Fig. 3-(a). The RGB network is guided by the output of the thermal network, focusing on regions where the thermal mask is activated. The thermal network is also guided in the same manner by the RGB mask. To accomplish this, we design uncertainty-aware masking that eliminates uncertain training signals during mutual learning. Specifically, uncertainty-aware masks are calculated by incorporating intra-spectral and inter-spectral relationships.

To quantify uncertainty, we first calculate cross-entropy loss maps $L_\theta \in \mathcal{R}^{H\times W\times 1}$ for each spectral using student and teacher predictions. Subsequently, we concatenate the loss maps along the channel axis and apply softmax to generate an inter-spectral mask as follows:

$$\mathcal{M}_R^{inter}, \mathcal{M}_T^{inter} = (\text{split} \circ \text{softmax} \circ \text{cat})(-L_R, -L_T). \tag{4}$$

The inter-spectral mask assigns weights to spectral prediction during mutual learning based on which spectral-loss map is larger than the counterparts. Moreover, the intra-spectral uncertainty mask is generated by applying a sigmoid function $\sigma$ as follows:

$$\mathcal{M}_R^{intra} = \alpha(1-\sigma(L_R)),\ \mathcal{M}_T^{intra} = \alpha(1-\sigma(L_T)) \tag{5}$$

where $\alpha$ is set to 2. This mask can filter out inaccurate predictions even when they outperform the results of their counterparts. Finally, our masked mutual loss $L_{MML}$ is imposed via KL divergence loss with intra- and inter-spectral uncertainty masks as follows:

$$\begin{aligned} L_{MML} &= L_{KL}(P_R M_T, P_T M_T) \\ &\quad + L_{KL}(P_T M_R, P_R M_R) \\ M_\theta &= \mathcal{M}_\theta^{intra} \odot \mathcal{M}_\theta^{inter}. \end{aligned} \tag{6}$$

**Prototypical Self-Supervised Learning.** Due to the absence of manual annotation for training, it is essential to obtain reliable pseudo-labels from the pre-trained network. However, the susceptibility to domain shift and the inherent limitations of RGB sensors in low-light conditions can lead to inaccurate pseudo-labels, hindering the effective training of student networks. While MS-UDA employed CycleGAN [16] to generate a daytime thermal image to nighttime synthetic image for training, it is inherently limited by the model's inability to learn the distribution of real night conditions.

To solve this, we propose a prototypical self-supervised loss that employs the segmentation output of the decoder as pseudo-labels and transfers them to outputs of prototypes. This strategy promotes the model to learn the segmentation on the real night distribution despite the absence of manual annotations and the inherent limitations of pre-trained RGB models in low-light conditions. As illustrated in Fig. 2-(b), the model is trained on both daytime and nighttime images during the second learning stage. The daytime learning process proceeds in the same manner as the initial learning process described in Fig. 2-(a). In the night condition, a night thermal image is fed into the thermal encoder and decoder, resulting in segmentation outputs from the decoder and the prototypes, respectively. The decoder's output is then employed as a pseudo-label, and the KL divergence loss is imposed as follows:

$$L_{PSL} = L_{KL}(\tilde{P}_T, P_T) \tag{7}$$

where $\tilde{P}_T$ and $P_T$ refer to segmentation outputs generated by the prototypes and decoder, respectively.

Since our cross-spectral prototypes can be updated using RGB and thermal features as shown in Eq. (2), we only selectively update our cross-spectral prototypes in the daytime training set and freeze the prototypes when the model is trained on nighttime-condition images.

### C. Training Loss

**Stage 1.** As shown in Fig. 2-(a), stage 1 of our proposed framework leverages daytime images for training. We define a training loss as follows:

$$L_{stage1} = L_{seg} + \lambda_1 L_\eta + \lambda_2 L_{MML} \tag{8}$$

where $\lambda_1$ and $\lambda_2$ are set to 0.2, 20, respectively.

**Stage 2.** In stage 2, we impose a prototypical self-supervised loss for nighttime in the absence of reliable pseudo-labels, as depicted in Fig. 2-(b). The same process is performed for daytime images, as in Fig. 2-(a). We calculate the training loss for stage 2 as follows:

$$L_{stage2} = L_{stage1} + \lambda_3 L_{PSL} \tag{9}$$

where $\lambda_3$ is set to 20. $L_{stage1}$ is computed using daytime cross-spectral images, while $L_{PSL}$ is calculated from nighttime single thermal image.

## IV. EXPERIMENTS

### A. Implementation Details

**Dataset and Evaluation Metric.** For training our framework, we generate pseudo-labels with a teacher network [26] trained on Cityscapes [2], a large-scale RGB dataset (i.e., 5,000 images) that consists of 19 classes, the same setting in MS-UDA [14]. We conduct our experiments on two public RGB-T paired datasets: the MF dataset [4] and the KAIST Multispectral Pedestrian Detection (KP) dataset [3].

As the MF dataset is annotated in 9 classes, we only report the mean intersection over union (mIoU) on the three classes that overlap with Cityscapes (i.e., car, person, and bicycle). In contrast to mixed day-night evaluation protocols in previous methods, MS-UDA exclusively trains on 820 daytime images and tests on 749 nighttime images. Due to this setting, it is difficult to make fair comparisons with other previous methods. Therefore, we re-implemented MS-UDA with 410 training daytime images and marked it as MS-UDA* in the experiment table. All the evaluations were conducted on 393 day-night testing images.

TABLE I
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON MF DAY-NIGHT EVALUATION SET [4]. (D+N: TRAINING WITH DAYTIME AND NIGHTTIME IMAGES ; R+T: TESTING WITH RGB-T IMAGE PAIRS)

| | Method | Train | Test | Car | Person | Bicycle | **mIoU↑** |
|---|---|---|---|---|---|---|---|
| Sup. | MFNet [4] | D+N | R+T | 65.9 | 58.9 | 42.9 | 55.9 |
| | RTFNet [6] | D+N | R+T | 86.3 | 67.8 | 58.2 | 70.7 |
| | MDBFNet [7] | D+N | R+T | 85.9 | 69.2 | 58.9 | 71.3 |
| | CENet [5] | D+N | R+T | 85.8 | 70.0 | 61.4 | 72.4 |
| | EAEFNet [8] | D+N | R+T | 86.8 | 71.8 | 62.0 | 73.5 |
| | CMX [9] | D+N | R+T | 90.1 | 75.2 | 64.5 | 76.6 |
| | CRM [10] | D+N | R+T | 90.0 | 75.1 | 67.0 | 77.4 |
| **UDA** | ProCA [11] | D | T | 50.8 | 36.8 | 14.2 | 33.9 |
| | DAFormer [12] | D | T | 52.0 | 51.6 | 38.9 | 47.5 |
| | MS-UDA* [14] | D | T | 82.1 | 73.4 | 55.6 | 70.4 |
| | Ours w/o stage 2 | D | T | 85.7 | 78.4 | 61.0 | 75.0 |
| | ProCA [11] | D+N | T | 48.9 | 47.1 | 15.4 | 37.1 |
| | DAFormer [12] | D+N | T | 48.9 | 55.4 | 28.9 | 44.4 |
| | HeatNet [13] | D+N | R+T | 56.4 | 68.8 | 33.9 | 53.0 |
| | EKNet [5] | D+N | T | 78.6 | 67.5 | 51.9 | 66.0 |
| | Ours | D+N | T | **86.2** | **80.3** | **64.3** | **76.9** |

* We re-implemented MS-UDA with 410 training daytime images

The KP dataset consists of RGB-T paired 95K video frames (62.5K for daytime and 32.5K for nighttime) on the urban driving scene for pedestrian detection. For semantic segmentation, MS-UDA manually annotated 950 images with the same 19 class ground-truth labels as Cityscapes. Recently, Shin *et al.* [10] divided 950 annotated images into 499, 140, and 311 for training, validation, and testing, respectively, and we followed their setting.

**Training Details.** As in MS-UDA [14], we adopt basic structure of RTFNet [6] consists of an encoder-decoder network. We also employed ImageNet [1] pre-trained ResNet50 as an encoder. All experiments were conducted with a batch size of 8 on 2 NVIDIA GeForce RTX 3090 GPUs.

At the initial learning stage (i.e., Ours w/o stage 2 on the experiment tables), we train the model with $120k$ iterations for the MF dataset and $500k$ iterations for the KP dataset. We adopt the same training data split as MS-UDA, utilizing 410 unlabeled daytime images for training the MF dataset and 3,283 unlabeled daytime images for training the KP dataset. We also utilize fake night thermal images generated from CycleGAN [16] with a 50% probability. In the case of the second learning stage, we train the model using only 30% of the iterations trained in stage 1. We utilize 374 and 3,095 additional unannotated nighttime images for training and fake night images used in stage 1 are not employed.

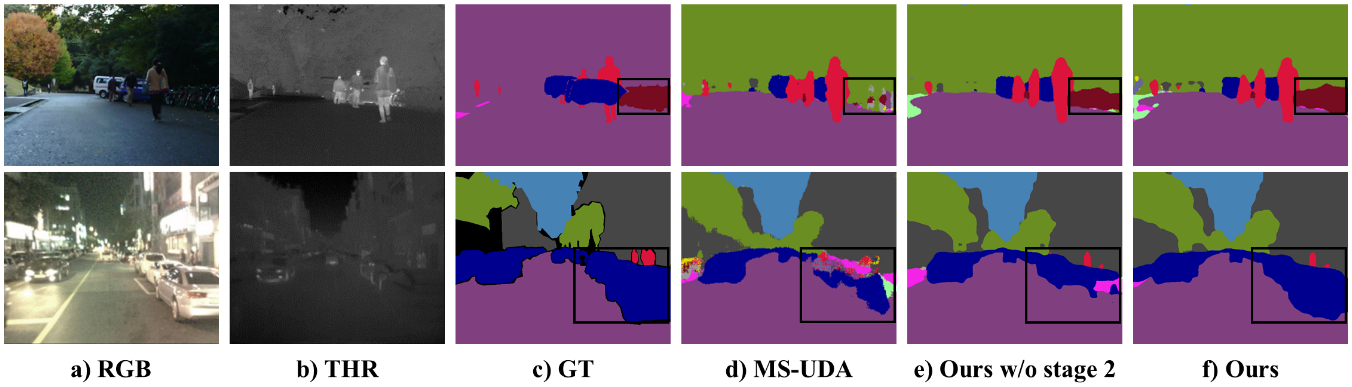
### B. Main Results

**Quantitative Results on the MF Dataset [4].** Table I demonstrates the effectiveness of our framework on the MF dataset. To accurately compare the effectiveness of each learning strategy, the networks for all UDA methods except HeatNet [13] are set to RTFNet [6] with the ResNet50-based

TABLE II

QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON KP DAY-NIGHT EVALUATION SET [3]. (D+N: TRAINING WITH DAYTIME AND NIGHTTIME IMAGES ; R+T: TESTING WITH RGB-T IMAGE PAIRS

| | Method | Train | Test | Road | Sidewalk | Building | Wall | Fence | Pole | Traffic light | Traffic sign | Vegetation | Terrain | Sky | Person | Rider | Car | Truck | Bus | Train | Motorcycle | Bicycle | **mIoU↑** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sup. | MFNet [4] | D+N | R+T | 93.5 | 23.6 | 75.1 | 0.0 | 0.1 | 9.1 | 0.0 | 0.0 | 69.3 | 0.2 | 90.4 | 24.0 | 0.0 | 69.6 | 0.3 | 0.3 | 0.0 | 0.0 | 0.6 | 24.0 |
| | RTFNet [6] | D+N | R+T | 94.6 | 39.4 | 86.6 | 0.0 | 0.6 | 0.0 | 0.0 | 0.0 | 81.7 | 3.7 | 92.8 | 58.4 | 0.0 | 87.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 | 28.7 |
| | CMX [9] | D+N | R+T | 97.7 | 53.8 | 90.2 | 0.0 | 47.1 | 46.2 | 10.9 | 45.1 | 87.2 | 34.3 | 93.5 | 74.5 | 0.0 | 91.6 | 0.0 | 59.7 | 0.0 | 46.1 | 0.2 | 46.2 |
| | CRM [10] | D+N | R+T | 99.0 | 61.9 | 91.8 | 0.0 | 58.7 | 50.6 | 39.2 | 55.3 | 89.2 | 23.2 | 94.3 | 85.2 | 2.9 | 95.3 | 0.0 | 80.5 | 0.0 | 66.2 | 54.6 | 55.2 |
| **UDA** | DAFormer [12] | D | T | 67.5 | 0.1 | 0.4 | 0.0 | 0.1 | 1.9 | 0.0 | 3.6 | 22.3 | 0.4 | 4.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 5.3 |
| | ProCA [11] | D | T | 71.4 | 5.0 | 44.3 | 0.0 | 2.0 | 2.1 | 0.0 | 1.3 | 26.2 | 2.2 | 15.6 | 16.8 | 0.0 | 40.8 | 0.0 | 0.1 | 0.0 | 0.1 | 0.0 | 12.0 |
| | MS-UDA [14] | D | T | 97.2 | 25.7 | 86.3 | 0.0 | 31.4 | 16.5 | 0.0 | 28.5 | 83.1 | 25.1 | 92.5 | 60.8 | 0.0 | 85.1 | 0.0 | 78.3 | 0.0 | 0.0 | 0.0 | 37.4 |
| | Ours w/o stage 2 | D | T | 97.7 | 33.0 | **88.7** | 0.0 | 35.5 | **36.4** | 14.9 | 46.2 | 85.6 | **24.5** | **94.4** | **72.2** | 4.9 | 88.4 | 0.0 | **82.9** | 0.0 | 35.3 | **4.3** | 44.5 |
| | DAFormer [12] | D+N | T | 57.6 | 0.1 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 4.9 | 28.8 | 3.5 | 19.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 6.1 |
| | ProCA [11] | D+N | T | 74.0 | 5.3 | 48.8 | 0.0 | 1.4 | 2.3 | 0.0 | 0.3 | 21.1 | 2.3 | 11.6 | 17.9 | 0.0 | 22.5 | 0.0 | 0.3 | 0.0 | 0.0 | 0.0 | 10.9 |
| | Ours | D+N | T | **97.8** | **34.4** | 88.6 | 0.0 | **36.6** | 35.0 | **17.9** | **47.3** | **85.7** | **24.5** | **94.4** | 71.9 | **5.5** | **89.3** | **0.2** | 81.7 | 0.0 | **56.0** | 2.6 | **45.8** |



a) RGB

b) THR

c) GT

d) MS-UDA

e) Ours w/o stage 2

f) Ours

Fig. 4. Qualitative comparison with MS-UDA [14] on the MF [4] and KP [3] datasets. We visualize the prediction result of the daytime image for the MF dataset (first row) and the nighttime image for the KP dataset (second row). In comparison to MS-UDA, our approach shows robust performance for both daytime and nighttime across all classes.

encoder.

Compared with existing methods, our UDA framework outperforms most of the previously supervised and all UDA methods. Furthermore, we achieve comparable performance to the state-of-the-art supervised method [10] only with a single thermal image as input. Specifically, ProCA [11] and DAFormer [12] achieve poor results in cross-spectral images despite demonstrating superior performance in the RGB domain. These methods indicate a substantial mIoU drop for specific classes, especially the bicycle class in thermal images, due to inherent deficiencies of thermal cameras (e.g., difficulty in capturing objects with low thermal emissivity). These results imply the necessity for a cross-spectral UDA strategy that can effectively leverage the complementary knowledge of the RGB and thermal domains. On the other hand, our method achieves superior performance compared to existing state-of-the-art approaches in single-spectral [11], [12] and cross-spectral domains [13], [14], [5] even when employing only stage 1 (i.e., Ours w/o stage 2). Moreover, our final model (i.e., Ours) exhibits a remarkable improvement in mIoU of 6.5 over the baseline [14], indirectly demonstrating the effect of our prototypical self-supervised loss in enabling segmentation training without manual annotations or a pre-trained teacher network.

**Quantitative Results on the KP Dataset [3].** In Table II, we also present the performance on the KP dataset. Consistent with the observations on the MF dataset, the state-of-the-art methods for single and cross-spectral UDA show insufficient segmentation results since they do not leverage the advantages of mutual information between cross-spectral images during knowledge distillation. On the other hand, our method significantly surpasses the advanced cross-spectral unsupervised domain adaptation method [14], despite employing the same model as in  [14] (37.4 vs 44.5 in mIoU).

**Qualitative Results.** Fig. 4 shows qualitative comparisons on the MF and KP datasets. As shown in the black box of row 1, MS-UDA inaccurately predicts the bicycle class. In contrast, our proposed method provides successful prediction when employing MML and cross-spectral prototypes only (i.e., Ours w/o stage 2). These results are attributed to our masked mutual learning that encourages the student models to effectively transfer complementary information within cross-spectral images to each other by filtering out the inherent limitations associated with each spectral domain. Moreover, compared to MS-UDA, our model makes good predictions in nighttime conditions, as illustrated in row 2. These results show that our prototypical self-supervised loss with the cross-spectral prototype is effective in enhancing performance at nighttime where RGB pre-trained networks struggle with reliable annotation generation.

## C. Ablation Study

**One-Way Distillation vs Mutual Learning.** In Table IV, we present an ablation analysis for each component of

TABLE III

Ablation results for each masking strategy applied to Masked Mutual Learning (MML) on MF dataset [4]

| Masking Strategy | Car | Person | Bicycle | **mIoU↑** |
|---|---|---|---|---|
| w/o Mask | 82.7 | 76.7 | 58.9 | 72.8 |
| Intra | **85.3** | **78.0** | 56.1 | 73.1 |
| Inter | 83.9 | 77.8 | 57.2 | 72.9 |
| Intra & Inter | 85.0 | 76.6 | **59.0** | **73.6** |

TABLE IV

Ablation results of components in our proposed framework. We report the mIoU on MF dataset [4]

| MML | Cross-spectral Prototypes | Stage 2 | Thermal | | | RGB |
|---|---|---|---|---|---|---|
| | | | Day | Night | **All** | **All** |
| One-way distillation [14] | | | 71.6 | 67.0 | 70.4 | 61.7 |
| Two-way distillation | | | 75.3 | 67.9 | 72.8 | 69.4 |
| ✓ | | | 76.4 | 67.4 | 73.6 | 69.2 |
| ✓ | ✓ | | 77.2 | 70.5 | 75.0 | 71.1 |
| ✓ | ✓ | ✓ | **78.8** | **73.2** | **76.9** | **73.4** |

our framework on the MF dataset. We first replace the sequential one-way distillation in MS-UDA with a two-way distillation framework. To achieve this, we train the student models by imposing $L_{seg}$ and $L_{MML}$ without applying our domain-wise masks $M_R$ and $M_T$. Interestingly, this two-way distillation method achieves remarkable mIoU gains for both thermal (70.4 vs 72.8) and RGB segmentation (61.7 vs 69.4). This suggests that complementary information between cross-spectral domains is important.

**The Effectiveness of Masked Mutual Learning.** Although mutual learning facilitates efficient model learning between students, it may also lead to the dissemination of potentially unreliable knowledge. To address this issue, our masked mutual learning can be employed to refine the knowledge by considering both inter and intra-spectral dependencies. To validate the effect of MML, we ablate the uncertain-aware masking components. In Table III, our intra and inter-spectral masks bring minor performance gains when applied independently (72.8 → 72.9, 73.1). However, when they are combined, we achieve a significant performance gain compared to the mutual learning model without a mask (72.8 → 73.6 in mIoU). This implies that it is crucial to consider both the intra and inter-spectral components for calculating uncertain masks.

**The Benefits of Cross-Spectral Prototypes.** We also evaluated the cross-spectral prototypes in stage 1 and achieved an impressive improvement of mIoU in both thermal (73.6 → 75.0) and RGB domains (69.2 → 71.1) as shown in the fourth row of Table IV. This implies that our method effectively learns scant semantic information (e.g., a person in the RGB domain and a bicycle in the thermal domain) inherently limited to individual domains by capturing complementary knowledge between spectral domains through cross-spectral prototypes.

**The Effectiveness of Prototypical Self-Supervised Loss.** As mentioned in Section III, we designed a prototypical self-supervised loss in stage 2 for self-supervision at nighttime due to the absence of reliable pseudo-labels. In Table IV, we observe that our prototypical self-supervised loss leads to a significant performance boost in mIoU for the day (77.2 → 78.8) and night times (70.5 → 73.2). Interestingly, despite that our prototypical self-supervised loss is only imposed on a thermal student network, the RGB student model also achieves meaningful improvement (71.1 → 73.4). These results suggest that our prototypical self-supervised loss is effective in training the model despite the absence of reliable ground truth labels at nighttime.

TABLE V

Comparison of generalization performance across diverse segmentation network architectures on MF dataset [4]

| Network | Method | Car | Person | Bicycle | **mIoU↑** |
|---|---|---|---|---|---|
| RTFNet [6] | MS-UDA* [14] | 82.1 | 73.4 | 55.6 | 70.4 |
| | Ours w/o stage 2 | 85.7 | 78.4 | 61.0 | 75.0 |
| | Ours | **86.2** | **80.3** | **64.3** | **76.9** |
| DeepLab-V3 [15] | MS-UDA* [14] | 77.4 | 74.7 | 46.4 | 66.2 |
| | Ours w/o stage 2 | 82.2 | 75.6 | 50.1 | 69.3 |
| | Ours | **82.6** | **75.7** | **53.0** | **70.5** |

* We re-implemented MS-UDA with 410 training daytime images

**Generalizability Across Diverse Segmentation Networks.** Consistent with MS-UDA [14], we leverage RTFNet [6] as our segmentation network. To assess the generalizability of our proposed learning strategy, we conduct an experiment utilizing DeepLab-V3 [15], a widely employed network architecture for segmentation tasks. As shown in Table V, our method outperforms MS-UDA across both RTFNet and DeepLab-V3 architectures, highlighting its generalizability and robustness to model architectures.

## V. Conclusions

In this letter, we present a cross-spectral unsupervised domain adaptation (UDA) approach for thermal image semantic segmentation. Our proposed Masked Mutual Learning (MML) strategy facilitates effective UDA by enabling the transfer of essential information between RGB and thermal domains. Moreover, we introduce cross-spectral prototypes to incorporate pixel-wise semantic knowledge. These prototypes are subsequently employed within a novel prototypical self-supervised loss function, enabling robust training even under unreliable nighttime conditions. Experimental results demonstrate the effectiveness of our framework, which significantly outperforms previous UDA methods while achieving competitive results with state-of-the-art supervised methods. In the future, we will investigate novel pseudo-label enhancement methodologies to refine our framework.

## Acknowledgment

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) program(IITP-2024-RS-2022-00156345) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation)

# References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[2] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213-3223.

[3] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1037-1045.

[4] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 5108-5115.

[5] Z. Feng, Y. Guo, and Y. Sun, "CEKD: Cross-modal edge-privileged knowledge distillation for semantic scene understanding using only thermal images," in *IEEE Robot. Automat. Lett.*, vol. 8, no. 4, 2023, pp. 2205-2212.

[6] Y. Sun, W. Zuo, and M. Liu, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," in *IEEE Robot. Automat. Lett.*, vol. 4, no. 3, 2019, pp. 2576–2583.

[7] W. Liang, C. Shan, Y. Yang, and J. Han, "Multi-branch differential bidirectional fusion network for rgb-t semantic segmentation," in *IEEE Trans. Intell. Veh.*, 2024, pp. 1-11.

[8] M. Liang, J. Hu, C. Bao, H. Feng, and T. L. Lam, "Explicit attention-enhanced fusion for RGB-thermal perception tasks," in *IEEE Robot. Automat. Lett.*, vol. 8, no. 7, 2023, pp. 4060-4067.

[9] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers," in *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, 2023, pp. 14679–14694.

[10] U. Shin, K. Lee, I. S. Kweon, and J. Oh, "Complementary Random Masking for RGB-Thermal Semantic Segmentation," in *arXiv preprint arXiv:2303.17386*, 2023.

[11] Z. Jiang, Y. Li, C. Yang, P. Gao, Y. Wang, T. Tai, and C. Wang, "Prototypical contrast adaptation for domain adaptive semantic segmentation," in *Proc. Eur. Conf. Computer Vis.*, 2022, pp. 36-54.

[12] L. Hoyer, D. Dai, and L. Van Gool, "Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9924-9935.

[13] J. Vertens, J. Zürn, and W. Burgard, "Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 8461–8468.

[14] Y. H. Kim, U. Shin, J. Park, and I. S. Kweon, "MS-UDA: Multi-spectral unsupervised domain adaptation for thermal image semantic segmentation," in *IEEE Robot. Automat. Lett.*, vol. 6, no. 4, 2021, pp. 6497-6504.

[15] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," in *arXiv preprint arXiv:1706.05587*, 2017.

[16] J. Y. Zhu, Park. T, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. Int. Conf. Computer Vis.*, 2017, pp. 2223-2232.

[17] C. Li, W. Xia, Y. Yan, B. Luo, and J. Tang, "Segmenting objects in day and night: Edge-conditioned CNN for thermal image semantic segmentation," in *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, 2020, pp. 3069-3082.

[18] H. Xiong, W. Cai, and Q. Liu, "MCNet: Multi-level correction network for thermal image semantic segmentation of nighttime driving scene," in *Infr. Phys. Technol.*, vol. 113, Art. no. 103628, 2021.

[19] P. Zhang, B. Zhang, T. Zhang, D. Chen, Y. Wang, and F. Wen, "Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12414-12424.

[20] G. Lee, C. Eom, W. Lee, H. Park, and B. Ham, "Bi-directional contrastive learning for domain adaptive semantic segmentation," in *Proc. Eur. Conf. Computer Vis.*, 2022, pp. 38-55.

[21] Y. Ganin et al., "Domain-adversarial training of neural networks," in *J. Mach. Learn. Res.*, vol. 17, no. 59, 2016, pp. 1-35.

[22] J. Hoffman et al., "Cycada: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1989-1998.

[23] Y. H. Tsai, W. C. Hung, S. Schulter, K. Sohn, M. H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7472-7481.

[24] T. H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2517-2526.

[25] Y. Zou, Z. Yu, B. V. K. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. Eur. Conf. Computer Vis.*, 2018, pp. 289-305.

[26] J. Wang et al., "Deep high-resolution representation learning for visual recognition," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.43, no. 10, 2020, pp. 3349-3364.

[27] S. Ainetter, and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from rgb," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 13452-13458.