Thermal Semantic Segmentation using Unsupervised Domain Adaptation with Unpaired RGB and Thermal Images

Du, Muliang

Award date:
2021

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Automotive Technology
Signal Processing Systems Group

Thermal Semantic Segmentation using Unsupervised Domain Adaptation with Unpaired RGB and Thermal Images

Graduation Thesis

Student:	Muliang Du
Identity Number:	1279874
Thesis Committee:	Gijs Dubbelman, Sveta Zinger, Tanir Ozcelebi
Email:	m.du@student.tue.nl
Supervisors:	Pavol Jancura, Anweshan Das
Study Load:	45 ECTS
Year of Graduation:	2021

Eindhoven, October 2021

Declaration concerning the TU/e Code of Scientific Conduct for the Master's thesis

I have read the TU/e Code of Scientific Conduct[i].

I hereby declare that my Master's thesis has been carried out in accordance with the rules of the TU/e Code of Scientific Conduct

Date

22-10-2021

Name

Muliang Du

ID-number

1279874

Signature

Muliang Du

Submit the signed declaration to the student administration of your department.

[i] See: https://www.tue.nl/en/our-university/about-the-university/organization/integrity/scientific-integrity/
The Netherlands Code of Conduct for Scientific Integrity, endorsed by 6 umbrella organizations, including the VSNU, can be found here also. More information about scientific integrity is published on the websites of TU/e and VSNU

Thermal Semantic Segmentation using Unsupervised Domain Adaptation with Unpaired RGB and Thermal Images

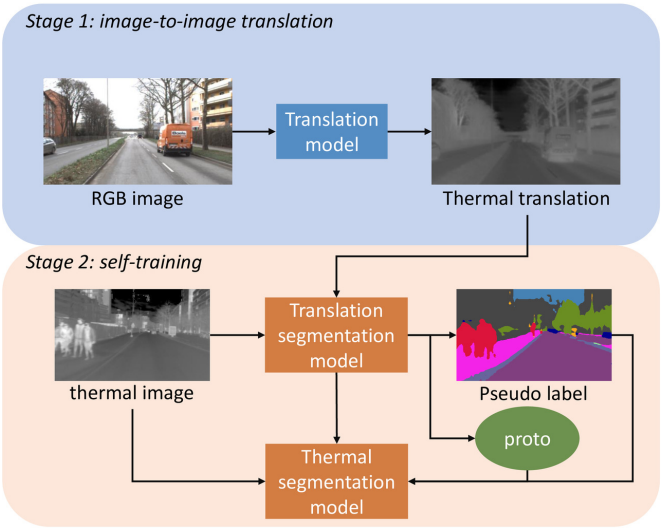
Muliang Du
Department of Automotive Technology
Einhoven University of Technology
Eindhoven, the Netherlands
m.du@student.tue.nl

***Abstract*—Thermal cameras interpret the captured scene by sensing the temperature of the objects, enabling thermal cameras to perceive barely visible objects in particular situations, such as poor illuminating and foggy environments. This motivates autonomous driving systems to include thermal cameras in the suite of the sensors, which compensates for the over-reliance of conventional RGB cameras on visible light. In recent years, a decent amount of researches pursued semantic segmentation tasks using RGB imaging, which aims to provide a class label to each pixel in an image and which requires costly annotation work. However, there has been little discussion on semantic segmentation using thermal imaging. Moreover, few relevant datasets have been published. Therefore, this work aims to train a semantic segmentation model using thermal imaging with an unsupervised domain adaptation (UDA) method. Our UDA approach consists of two stages: a) unpaired image-to-image translation; b) self-training with prototypical pseudo label denoising. Comprehensive experiments show our method can significantly improve the performance compared to without domain adaptation.**

I. Introduction

After nearly four decades of gestation, autonomous driving is now becoming a reality. This benefits from the accelerated growth of deep neural networks in computer vision in recent years. However, safety is one of the most frequently stated problems with autonomous driving. The study of [1] claims that 75% of pedestrian fatalities in the U.S. occurred in a dark environment. One of the key factors is the limitation of RGB cameras. RGB cameras are widely used environmental sensors for autonomous driving. However, they are unreliable in poor illumination environments, such as low and excessive lighting, resulting in possible hazards. A solution addressing this problem is applying thermal cameras instead of RGB cameras. The reason is that thermal cameras sense the transmitted infrared radiation of objects independent of the illumination. Therefore, incorporating a thermal camera in the suite of the sensor would fill the missed points in the interpretation of the environment.

Deep neural networks (DNN) have developed rapidly in recent years, which has led to their increasing use in computer vision. One of the major topics to be investigated in computer vision with DNN for autonomous driving is semantic segmentation. It aims to classify each pixel in an image into predefined classes, e.g., vehicles, buildings, roads, etc. Currently, many studies on semantic segmentation of RGB images have yielded remarkable achievements. Nonetheless, although thermal cameras are gradually becoming crucial support to autonomous driving, the semantic segmentation of thermal images has been sparsely discussed compare to RGB semantic segmentation. In this paper, we present a method to learn a thermal semantic segmentation model.



(a) Thermal image (b) before adaptation (c) Our prediction

Fig. 1: **Overview of the proposed UDA framework.** stage 1: a RGB-to-thermal translation model generates thermal images; stage 2: a target segmentation model is trained using a self-training method. (proto: prototypes).

Learning a semantic segmentation model with conventional supervised learning methods typically requires a large amount of pixel-wise annotation. Manually annotating a semantic segmentation dataset is highly laborious and time-consuming. Intuitively, one feasible way to reduce annotation cost is by training with existing annotated data that is used for a similar task. However, since deep neural networks are sensitive to

domain misalignment, the performance usually is not ideal. In order to effectively avoid duplication of labeling efforts, recent researches proposed unsupervised domain adaptation (UDA) methods for semantic segmentation [2]. The objective of UDA is to decrease the domain gap between the labeled source domain and the unlabeled target domain by some specific methods. Thereby a semantic segmentation model for the target domain can be trained with the annotated source data.

A considerable amount of approaches of UDA for semantic segmentation have been introduced. Among them, the generative-based methods and self-training strategy have promising performance [2]. The objective of the generation model is to transfer the style of the annotated source domain data to the style of the target domain while preserving the semantic information of the source domain data. In this way, the distribution of the original source domain is relocated and aligned to the distribution of the target domain. Since the source domain data are explicitly annotated, the semantic segmentation model of the target domain can be learned with the adapted source domain data.

One representative generative-based UDA framework is CycleGAN [3]. The framework contains two generative adversarial modules, executing image translation between two domains in both directions. CycleGAN framework can transfer the color and the texture of the target images to the source images. However, it disregards the semantic characteristics of the objects. This may result in a semantic inconsistency between the adapted source data and the target data. This will ultimately render the adapted source model incapable of correctly predicting the target domain data, even if it performs well on the adapted source domain data.

The adaptation of self-training-based strategies (e.g., [4]–[7]) is conducted in the process of training the segmentation model. On the basis of the source domain semantic segmentation model, pseudo-labels for the target domain data are generated by means of a confidence-based threshold criterion. The pseudo-labeled target domain data can then be used to retraining the source domain semantic segmentation model. New pseudo labels are thus iteratively obtained and then used to refine the semantic segmentation model. Gradually, the new semantic segmentation model generates more accurate pseudo labels, while the original source domain semantic segmentation model is adapted to the target domain. However, self-training-based methods require a robust pre-trained segmentation model to generate reliable pseudo labels. This is difficult to achieve for adaptation issues with large domain gaps.

While some UDA approaches for semantic segmentation (e.g., [8]–[10]) have been carried out on some paired domains (e.g., synthetic-to-real), there have been few empirical investigations into adaptation between RGB and thermal image domains. Hence, this work explores the ways to realize RGB-to-thermal unsupervised domain adaptation. Densely discussed domain adaptation tasks (e.g., synthetic-to-real, day-to-night) deal with domain gaps caused by virtual bias or illumination. Therefore, the domain gaps are weakly dependent on semantic contents. Differently, thermal images and RGB images represent semantic contents in entirely different ways. An RGB camera captures the visible spectrum, while a thermal camera records the emitted infrared radiation from the objects. This results in RGB images representing the captured sceneries by their colors, while thermal images present the temperature of the pixel locations in the scene as different shades of gray. This difference in representation yields a significant domain gap between the RGB domain and the thermal domain. Consequently, RGB-to-thermal is more challenging than synthetic-to-real domain adaptation.

In this paper, we aim to solve this UDA problem by a combination of a generative-based method and a self-training-based method. We propose that these two approaches can compensate for each other's shortcomings. The generative-based method can generate images close to the style of the target domain. Still, due to the lack of semantic consistency, these generated images do not represent the actual semantic features of the target domain properly. Meanwhile, the self-training-based method is able to adapt at the semantic level, but a reasonable pre-trained model is essential due to the requirement for the correctness of the pseudo labels. By combining these two approaches, the self-training method can further refine the model semantically on the basis of the generative-based method, which in turn can provide a reliable pre-trained model for the self-training-based method.

With both techniques, the thermal semantic segmentation model learned with our method improves significantly over the model without domain adaptation. We found that the generative-based-method was able to translate RGB images into thermal images effectively, and to a certain extent achieved semantic consistency after the translation. The self-training-based method can further promote the domain adaptation based on the generative-based method and attain a better semantic segmentation performance. Our contributions are:

- We propose a RGB-to-Thermal UDA method of the combination of a generative-based method and a self-training-based method. The generative-based method generates annotated thermal data. After learning a thermal semantic segmentation model with the generated target data, the self-training-based method can further increase the performance of the thermal segmentation model.
- We show the effectiveness of the generative-based method and the self-training-based method.
- The proposed method achieves 5-fold mIoU increase in segmentaion task on Freiburg IR thermal data when compared to the thermal augmentation without adaptation from Freiburg RGB data.

II. Related works

We first show generic semantic segmentation studies. After this, we focus on semantic segmentation for thermal images. After having a background on semantic segmentation, we will present existing approaches in the field of unsupervised domain adaptation research.

A. Semantic segmentation

Benefiting from the rapid development of deep convolutional neural networks, a considerable amount of semantic

segmentation model have been proposed. The first architecture for semantic segmentation is FCN [11]. Inspired by it, recent researches such as PSPNet [12], Deeplab [13] and its variations [14], [15] have achieved compelling performance. However, their performance strictly relies on the quality and quantity of training datasets. Great efforts have been paid on collecting and labeling semantic segmentation datasets (e.g., Cityscapes [16], PascalVOC [17]). One naive way to reduce annotation cost is to train models using data from a similar domain without considering the domain gap. However, degrades the performance of segmentation method considerably.

B. Semantic segmentation of thermal images

The research of semantic segmentation of thermal images can be divided into two main directions: unimodal thermal semantic segmentation and multimodal semantic segmentation including the thermal domain. The studies of unimodal thermal semantic segmentation are dedicated to developing networks to accommodate the features of thermal images. Li *et al.* [18] adaptively incorporate edge prior knowledge as guidance to train the semantic segmentation network with a dataset of diverse indoor and outdoor scenarios. Xiong *et al.* [19] propose a multi-level correction network with a multi-level attention module, which aims to capture the inter-class and intra-class contextual dependencies by a multi-level correction process, and a multi-level edge enhancement module combines precise context information and edge prior knowledge in each level to correct the final feature representation. They evaluate their approach on their own dataset. Different from unimodal thermal semantic segmentation researches, multimodal focus on learning networks that can adapt to multiple domains, such as thermal and RGB. Sun *et al.* [20] proposed an encoder-decoder architecture containing two encoders to obtain the features of thermal and RGB images separately. Ha *et al.* [21] developed an architecture for multi-spectral image segmentation, combining the thermal and visible light information to obtain boosted segmentation performance.

Different from the aforementioned works, we train a semantic segmentation model of thermal images without requiring any manual annotation work. Instead, we apply unsupervised domain adaptation method to adapt labelled RGB domain dataset, then use it to train a thermal semantic segmentation model.

C. Domain adaptation for semantic segmentation

UDA can be broadly divided into seven categories: classifier discrepancy [22], [23], domain adversarial discriminative [24], [25], generative-based [3], [9], [26], [27], self-training [5]–[7], entropy minimization [28], [29], curriculum learning [30], [31] and multi-tasking [32], [33]. Among these, generative-based and self-training based methods are the most relevant to our approach.

Generative-based approaches employ a strategy of generative adversarial learning. This includes the training of both the generator and the discriminator. The learning process is that the discriminator strives to correctly distinguish whether the image is real or generated, while the generator aims to generate fake images that can fool the discriminator. Liu and Tuzel [26] apply this idea to build the Coupled Generative Adaversarial Networks, which can generate corresponding pairs of images in different domains from same random noise samples. Yu *et al.* [27] proposed Simulated Generative Adversarial Networks, whose generation is conditioned on the source data. The objective of SimGAN is to refine the simulated images into real ones, which initially shows the idea of adapting source data to a target domain (synthetic-to-real). Bousmalis *et al.* [34] proposed a task oriented domain adaptation approach Pixel DA, where the learning process of the domain adaptation and task is decoupled. Notably, generative adversarial networks lack constraints on image content during domain adaptation. To overcome this, CycleGAN [3] applies a bidirectional architecture to implement domain adaptation in both source-to-target and target-to-source to constrain the image-level translation on a modest scale. Building on the success of CycleGAN, Cycada [9] further introduced semantic consistency into the construction of CycleGAN. It uses a pre-trained semantic segmentation model to classify images before and after translation, and then encourages the two classifications to be consistent as a way to preserving semantic information during translation. However, this approach of semantic consistency may not achieve positive results for adaptation tasks with large domain gaps.

The self-training approach involves employing highly confident network predictions estimated from unlabeled data to produce pseudo-labels, which are then used to retrain the network. At the same time, the pseudo-labels are iteratively updated in the training process. Zou et al. [5] firstly proposed to use pseudo labels in self-training for semantic segmentation. Yet, it has the shortcoming of frequently generating noisy labels. To overcome this, in their paper [6], Zou *et al.* developed confidence regularized self-training framework. In this approach, pseudo-labels are treated as continuous latent variables that are jointly optimized using alternating optimization. Nevertheless, in the aforementioned self-training methods, pseudo labels only get updated after the entire training stage. Without updating the pseudo labels on time, the model can easily overfit the noisy labels. Therefore, Zhang *et al.* [7] proposed a self-training strategy with an online pseudo label denoising technique, which can rectify the pseudo labels based on the distance between the sample and class features.

The above approaches have been applied to many domain adaptation tasks, for example, synthetic-to-real domain adaptation, art style transfer, day-to-night domain adaptation, etc. However, the unsupervised domain adaptation problem of RGB-to-thermal has been rarely discussed. Vertens *et al.* [35] indirectly achieve the adaptation from RGB to the thermal domain for semantic segmentation, but this is based on their perfectly aligned RGB and thermal images. This paper thus investigates the feasibility of unsupervised domain adaptation for the semantic segmentation task of thermal images.

III. Methodology

This section describes our approach to thermal semantic segmentation using unsupervised domain adaptation techniques, transferring knowledge of RGB data to thermal data.

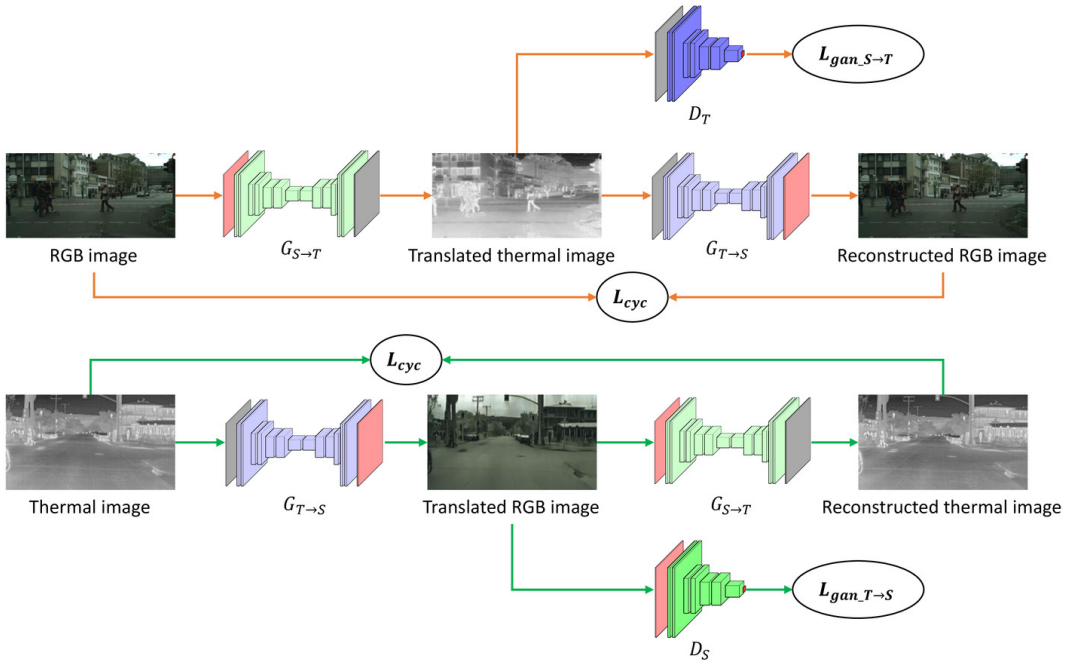


Fig. 2: Cycle-consistent architecture for RGB to thermal image-to-image translation. Upper part of the architecture is the cycle architecture for RGB to thermal domain translation, while the lower part is for thermal to RGB domain translation.

We define the RGB domain as source domain, while thermal domain as target domain. We use the source data X_S, source labels Y_S, target data X_T without target labels Y_T to approach the unsupervised domain adaptation problem. Our objective is to learn a model f that makes accurate predictions on the target data X_T. We go over our methodology in more depth below.

A. *Unpaired image-to-image translation*

Simply learning a source domain model f_S in a supervised fashion can achieve decent performance on the source data X_S. However, the source domain model cannot perform well when evaluated with target data X_T due to the domain gap between the target and source domain. Therefore, to alleviate the performance difference caused by the domain gap, we align the target and source domain distribution through an image-to-image translation approach. Specifically, our goal is to learn a mapping model that can translate source data into the target domain without losing content information. With the target domain oriented translations of source data and their corresponding labels, a target domain model f_T can be learned.

We denote this mapping model as $G_{S\to T}$. $G_{S\to T}$ can be learned through the adversarial learning technique, where we train it to translate the source domain data x_s into target domain style samples $G_{S\to T}(x_s)$ that can fool a target domain adversarial discriminator D_T. Concurrently, D_T is trained to correctly distinguish $G_{S\to T}(x_s)$ from real target data x_t. We apply adversarial loss to train $G_{S\to T}$ and D_T, which can be expressed as Equation 1.

$$\begin{aligned}\mathcal{L}_{\text{GAN\_S}\to\text{T}} = &\mathbb{E}_{x_t\sim X_T}\left[\log D_T(x_t)\right] \\ &+ \mathbb{E}_{x_s\sim X_S}\left[\log\left(1 - D_T(G_{S\to T}(x_s))\right)\right]\end{aligned} \tag{1}$$

$$\begin{aligned}\mathcal{L}_{\text{GAN\_T}\to\text{S}} = &\mathbb{E}_{x_s\sim X_S}\left[\log D_S(x_s)\right] \\ &+ \mathbb{E}_{x_t\sim X_T}\left[\log\left(1 - D_S(G_{T\to S}(x_t))\right)\right]\end{aligned} \tag{2}$$

While the learning objective of D_T is to maximize the adversarial loss by making correct domain classification, G_{S-T} tries to minimize the loss by generating more realistic target domain style samples. Subsequently, we can learn a target model f_T in a supervised fashion with the convincing translations $G_{S-T}(X_S)$ and Y_S.

Nevertheless, simply applying the aforementioned adversarial training approach mentioned above does not assure that G_{S-T} will generate valid target domain style samples for learning the target model f_T, due to the fact that G_{S-T} is not restricted to retaining content information but only performs style transfer.

To further encourage $G_{S\to T}(x_S)$ to preserve content and structure information, we apply the cycle-consistency architecture from [3]. This architecture realizes the preservation through reproducing the original image x_S from its translation $G_{S-T}(x_S)$, i.e. $G_{T\to S}\left(G_{S\to T}\left(x_s\right)\right) \approx x_s$. The target to source mapping model $G_{T\to S}$ can be learned in synchronisation with $G_{S\to T}$ in the same way (see Equation 2), together with source domain discriminator D_S. Specifically,

the reconstructions of two domain samples is trained with the cycle-consistency loss (Equation 3), where the reconstruction errors are penalized with the L1 penalty.

$$\begin{aligned}\mathcal{L}_{\text{cyc}} = &\mathbb{E}_{x_s \sim X_S}\left[\left\|G_{T\to S}\left(G_{S\to T}\left(x_s\right)\right) - x_s\right\|_1\right] \\ &+ \mathbb{E}_{x_t \sim X_T}\left[\left\|G_{S\to T}\left(G_{T\to S}\left(x_t\right)\right) - x_t\right\|_1\right]\end{aligned} \tag{3}$$

Altogether, the overall training loss for the image-to-image translation architecture can be concluded as Equation 4, where λ is the coefficient for weighting the contribution of the cycle-consistency loss. Similar to [3], we set it to 10. The architecture and overall training scheme is shown as Figure 2.

$$\mathcal{L}_{trans} = \mathcal{L}_{\text{GAN\_S}\to\text{T}} + \mathcal{L}_{\text{GAN\_T}\to\text{S}} + \lambda\mathcal{L}_{\text{cyc}} \tag{4}$$

After we learned a reasonable mapping model G_{S-T}, we can generate target domain style data $G_{S-T}(X_S)$. We denote the translated data as X_{TR}. With X_{TR} and Y_S, the target model f_T can be learned by minimizing a cross-entropy loss:

$$\begin{aligned}p_{tr} &= f_T\left(x_{tr}\right) \\ \mathcal{L}_{\text{ce}}(p_{tr}, y_s) &= -\mathbb{E}_{(x_{tr},y_s)\sim(X_{TR},Y_S)} \sum_{k=1}^{K} \mathbb{1}_{[k=y_s]} \log\left(\sigma\left(p_{tr}^{(k)}\right)\right)\end{aligned} \tag{5}$$

where $\mathbb{1}$ is the indicator function and σ denotes the softmax function.

B. Self-training with prototypical pseudo label denoising

Although the unpaired image-to-image translation method can generate target domain style images with consistent structure and content, the semantic information of the source data is not considered during the domain adaptation process. That may lead to a performance drop of target model f_T when evaluating with real target data X_T. To further enhance the performance of f_T, we apply a self-training approach on the resulting target model obtained by the unpaired image-to-image translation method, which is denoted as f_{T0}.

Traditional self-training approaches optimize a model through iteratively generating pseudo labels of the target data X_T and training with samples that have a high-confidence pseudo label. The high-confidence pseudo labels are determined by a hard confidence threshold, namely only the pixels whose prediction confidence is higher than the threshold would be utilized to train the model. We denote those valid pseudo labels as hard pseudo labels.

However, the traditional self-training approaches using hard pseudo labels have some limitations. Firstly, the hard pseudo labels filtered by a fixed confidence threshold can be over-confident, resulting in a false interpretation of the target domain. Secondly, the hard pseudo labels only get updated after the whole training process. In addition, when updating the pseudo labels, although pseudo labels with low confidence are not definitely incorrect, they are never considered since their confidence is always lower than the threshold. This may lead to dispersed features in the target domain.

To alleviate the lackings of traditional self-training approaches, we apply the online prototypical pseudo label denoising technique proposed in [7]. The overall workflow is demonstrated as Figure 3. Instead of using hard pseudo labels, we apply fixed soft pseudo labels and adjust them with weights obtained by a clustering method. The fixed soft pseudo labels are the class-wise softmax scores of all pixels. More specifically, we express the weighted pseudo labels as:

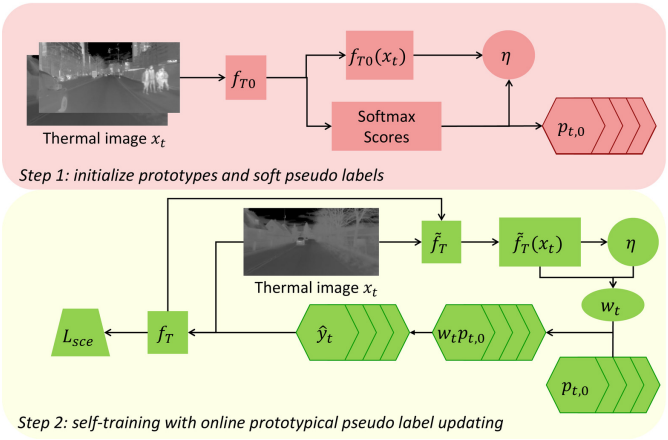


Fig. 3: Overall workflow of self-training with prototypical pseudo label denoising

$$\hat{y}_t^{(i,k)} = \begin{cases} 1, & \text{if } k = \arg\max_{k'}\left(w_t^{(i,k')} p_{t,0}^{(i,k')}\right) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $\hat{y}_t^{(i,k)}$ represents the hard weighted pseudo label of pixel $x_t^{(i)}$, $\omega_t^{(i,k)}$ is the weight for adjusting the probability of belonging to class k and the $p_{t,0}^{(i,k)}$ is the soft pseudo label, which is initialized by the pre-trained target model f_{T0} and fixed throughout the complete training process. Finally, the weighted soft pseudo label is converted as a hard pseudo label in one-hot form.

The prototype $\eta^{(k)}$ is the feature centroids of class k, which is defined based on the fact that the prototypes are located closer to the centroids of the underlying clusters. Thus, we assign a higher weight for class k if the distance between $f_T\left(x_t\right)^{(i)}$ and $\eta^{(k)}$ is relatively shorter. Conversely, classes with longer distances are assigned with lower weights. In general, this weight assigning strategy can be implemented in the form of the softmax score over the distances between features and prototypes. Equation 7 is the formula for computing the weights.

$$\omega_t^{(i,k)} = \frac{\exp\left(-\left\|\tilde{f}_T\left(x_t\right)^{(i)} - \eta^{(k)}\right\|\right)}{\sum_{k'}\exp\left(-\left\|\tilde{f}_T\left(x_t\right)^{(i)} - \eta^{(k')}\right\|\right)} \tag{7}$$

Similar to the work from [7], instead of extracting the features of training data with the updated target segmentation model f_T, we also apply the momentum encoder [36] as the feature extractor, denoted as $\tilde{f}_T$ in Equation 7. The momentum encoder $\tilde{f}_T$ is updated following:

$$\theta_{a,n+1} = m\theta_{a,n} + (1-m)\theta_{b,n}, m \in [0,1), n \in \mathbb{N} \tag{8}$$

where θ_a is the parameter of the momentum encoder $\tilde{f}_T$ and θ_b is the parameter of the target segmentation model f_T. n is the number of training epochs. When n equals 0, both θ_a and θ_b are equal to the parameter values of the pre-trained target model f_{T0}. m is the momentum coefficient which we set as 0.999. During training, f_T is updated by backpropagation, while $\tilde{f}_T$ itself is not updated, but fine-tuned by f_T. This makes the transformation of $\tilde{f}_T$ smoother, which in turn ensures that $\tilde{f}_T$ always extracts reliable data features throughout the training process.

The prototypes $\eta^{(k)}$ are initialized by using features of the complete target dataset extracted by the pre-trained target segmentation model f_{T0}. As mentioned above, we compute the average feature values of the target domain data as the prototypes $\eta^{(k)}$. More specifically, $\eta^{(k)}$ are calculated as shown in Equation 9, where $\mathbb{1}$ is the indicator function whose purpose is to count and calculate only the pixels in the dataset classified as k.

$$\eta^{(k)} = \frac{\sum_{x_t \in \mathcal{X}_t} \sum_i f_T\left(x_t\right)^{(i)} * \mathbb{1}\left(\hat{y}_t^{(i,k)} == 1\right)}{\sum_{x_t \in \mathcal{X}_t} \sum_i \mathbb{1}\left(\hat{y}_t^{(i,k)} == 1\right)} \tag{9}$$

Nevertheless, in order to update prototypes online during training, it is very time-consuming to compute prototypes for the entire target dataset. Hence, we calculate the prototypes of a batch of data within each iteration and then update the overall prototypes by moving the average. We denote the prototypes of a batch of data as $\eta_b^{(k)}$. So we update the overall prototypes $\eta^{(k)}$ as:

$$\eta^{(k)} \leftarrow (1-\lambda)\eta^{(k)} + \lambda \eta_b^{(k)} \tag{10}$$

Where λ is the moving-average momentum coefficient. We set it to 0.0001.

To update the target segmentation model f_T, we apply symmetric cross-entropy loss, which has tolerance to the noise of pseudo labels. It can be expressed as:

$$\mathcal{L}_{sce} = \alpha \mathcal{L}_{ce}\left(p_t, \hat{y}_t\right) + \beta \mathcal{L}_{ce}\left(\hat{y}_t, p_t\right) \tag{11}$$

where α and β are the balancing coefficients, which are set to 0.1 and 1 respectively similar to [7]. $\mathcal{L}_{ce}$ is defined in Equation 5.

C. two-stage training scheme

In our process, we first apply the unpaired image-to-image translation technique to translate an annotated RGB dataset into the thermal image style. Subsequently, we train a semantic segmentation model with the translations of the RGB dataset in a supervised fashion. The performance of the trained semantic segmentation model may not be optimal due to remained domain gap. Thus we further optimize the semantic segmentation model using the self-training algorithm with prototypical pseudo label denoising technique.

IV. Experiments

This section begins with an introduction to the experimental setup, which includes the construction of the training model for image-to-image translation, the setup of the semantic segmentation model we employ, the training details in the two stages and the dataset we use for training and evaluating. After this, we present our experimental results qualitatively and quantitatively and discuss the effects of our method by comparing them with baselines.

A. Implementation

1) Network Architecture: We apply the architecture of CycleGAN from [3] as our image-to-image translation model. Differently, since a thermal image has one channel, while an RGB image has three channels, we set the input layer input channel numbers of the mapping models and discriminators as shown in Table I:

Network	n_{in}	n_{out}
$G_{S\to T}$	3	1
$G_{T\to S}$	1	3
D_S	3	1
D_T	1	1

TABLE I: Mapping models and discriminators channel numbers configuration

where n_{in} is the input channel number of the input layer, n_{out} is the output channel number of the output layer.

Regarding the segmentation model, we adapt Deeplabv2 [13] with ResNet101 [37] as the backbone. The penultimate convolutional layer is configured to contain 256 filters, in order to create a feature space of a reasonable amount of features.

2) Training Details: During the training process of the image-to-image translation model, the training image is firstly cropped to a random size and aspect ratio, and then resized to 512 $\times$ 256. In addition, the cropped image is normalized with mean and standard deviation of 0.5 and 0.5 respectively. Similar to the strategy from [3], we randomly select samples from an image pool that contains 50 previously generated images to train the discriminators. We train the image-to-image translation model for 130 epochs. Both generators and discriminators are optimized by the Adam optimizer with a learning rate of 0.0001 and a batch size of 4.

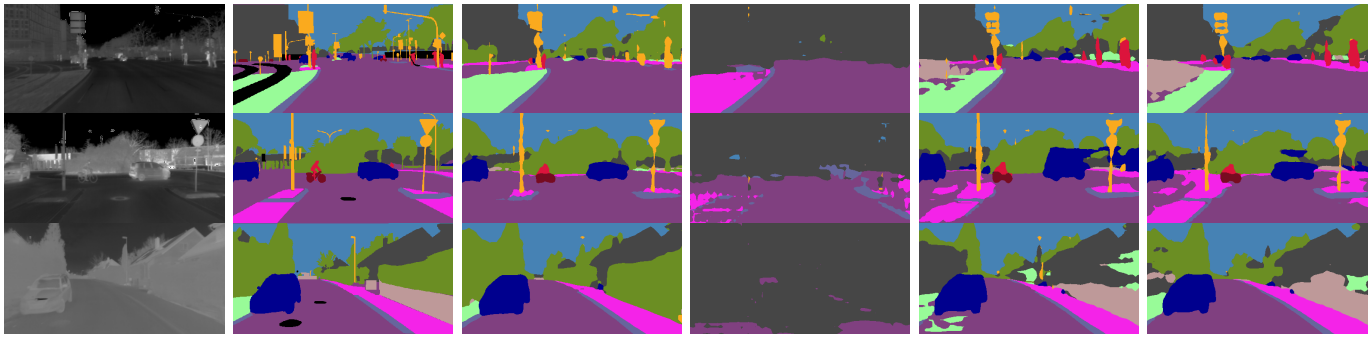
After the image-to-image translation model is trained, we translate the complete source domain dataset into the target domain style. The image is resized to 512 $\times$ 256 and also normalized. The output image is denormalized and resized to the original size.

We train the thermal semantic segmentation model with the translation of the source domain dataset. Similar to the image translation model training process, the translations are randomly cropped and resized to 512 $\times$ 256. We apply the Adam solver with the initial learning rate of 0.0001, reduced by a factor of 0.1 following a scheduler with a patience of 10 epochs.

Once the process of learning the thermal semantic segmentation model with translations is complete, we further refine

Model	Train on	Test on	Road	Sidewalk	Building	Curb	Fence	Pole	Vegetation	Terrain	Sky	Person	Car	Bicycle	mIoU	gain
Lower-b	FR-RGB	FR-T	53.6	5.3	21.5	6.9	0	0.2	4.2	0	0	0	0	1.1	8.4	-
Upper-b	FR-T	FR-T	89.4	65.8	69.6	53.5	42.2	42.5	69.2	59.1	82.8	67.2	86.9	63.2	66.2	-
CycleGAN	FR-Tr	FR-Tr	88.2	56.7	69.3	44.5	37.6	42.7	68.2	52.8	78.7	61.4	87.7	54.7	62.5	-
CycleGAN	FR-Tr	FR-T	84.6	47.2	61.2	37.6	25.7	**34.3**	57.2	33.8	70.8	58.2	62.9	**51.4**	52.1	+43.7
Self-training	FR-Tr + FR-T*	FR-T	**86.7**	**51.4**	**65.0**	**42.5**	**26.3**	31.5	**60.2**	**43.0**	**73.8**	**58.9**	**74.9**	33.7	**55.8**	+47.4
Joint	FR-Tr + FR-T*	FR-T	87.0	51.4	65.5	44.3	26.6	34.1	59.5	43.9	73.4	57.4	76.9	52.0	56.4	+48.0
Joint	FR-Tr + FR-T*	FR-T-night	84.6	53.7	68.9	46.3	21.0	33.6	49.9	44.3	51.7	41.8	81.5	46.0	52.5	+44.1

TABLE II: Quantitative performance comparison of our thermal semantic segmentation model and baselines on the Freiburg thermal dataset. The mIoU gains of baselines are marked as dash(-). **CycleGAN**: segmentation model trained with RGB image translations; **Self-training**: refined segmentation model with self-training; **Joint**: joint segmentation model; **FR-RGB**: Freiburg daytime RGB dataset; **FR-T**: Freiburg daytime thermal dataset; **FR-T***: Freiburg daytime thermal dataset without labels but soft pseudo labels and prototypes; **FR-Tr**: translation of FR-RGB; **FR-T-night**: Freiburg thermal night-time test dataset



(a) Thermal image (b) Groundtruth (c) Upper-bound (d) Lower-bound (e) CycleGAN (f) Self-training

Fig. 4: Qualitative performance comparison of our thermal semantic segmentation model and baselines on the Freiburg thermal dataset.

the model with the self-training approach with online pseudo label denoising for 10 more epochs. We adapt the Adam solver with the initial learning rate of 1e6, reduced by a factor of 0.1 following a scheduler with a patience of 500 iterations.

3) Dataset: We conduct the experiments using the Freiburg dataset [35]. The dataset contains 12501 daytime thermal images and also 12501 time-synchronized RGB images. Although the thermal images and RGB images are perfectly aligned, we ignore this correspondence and disrupt them as two separate datasets to investigate the adaptation between the two domains. The dataset provides RGB image annotations. We use the annotated RGB images as the source dataset and the thermal images as the target dataset. In addition, the dataset contains a test dataset with 32 annotated thermal images. We use this test dataset to evaluate the performance of our model. The pixel-wise semantic annotation of the training dataset and test dataset contains 13 classes, including Road, Sidewalk, Building, Curb, Fence, Pole/Signs, Vegetation, Terrain, Sky, Person/Rider, Car/Truck/Bus/Train, Bicycle/Motorcycle, and Background. The thermal images have full black masking on both sides, we intercept the part with information in the middle and crop the RGB images in the same area. The original thermal images have a bit-depth of 16 bits. Similar to [35], we crop the thermal images to relevant range [21800, 25000] and normalize them to [0, 1].

B. Baseline Comparison

We qualitatively and quantitatively compare the performance of our method with the lower-bound and upper-bound baselines. We use the source domain model without domain adaptation as the lower-bound baseline. With respect to the upper-bound baseline, a target domain model train with fully-supervised method is used. The model uses Freiburg thermal images and their corresponding RGB labels as training data. The quantitative results are listed in Table II. Unlike extensively studied domain adaptation tasks, such as synthetic to real domains, we observe that in the absence of domain adaptation, RGB domain models are incapable of making correct predictions for thermal domain data. The mIoU score of this model is only 8.4. This poor performance is to be expected since there is a large domain gap between the RGB and thermal domains due to the difference in object representation. In contrast, the mIoU score of the thermal

(a) RGB images (b) Thermal images (c) CycleGAN translations (d) GAN translations

Fig. 5: Image-to-image translation results. The white clothes in the first row and the white car in the second row are translated as low temperature (dark color in thermal images). This demonstrates the tendency of the image-to-image translation model to generate inverse colors when translating RGB images to thermal images.

model trained supervisedly is 66.2. Our model achieves a mIoU score of 55.8, which has a boost of 564% compare to the lower-bound baseline. Nevertheless, there is still a considerable performance gap between our method to fully-supervised method.

We also used the translation of the RGB test dataset to evaluate the segmentation model trained with the translation of the RGB train dataset. The third row of Table II shows it reaches a mIoU score of 62.5, which shows that the translations are informative enough to train a segmentation model with them, although during translation, there is still an information leakage of 3.7 mIoU score compare to the semantic segmentation model trained with the real thermal dataset.

V. Discussion

In this section, we first discuss the effectiveness of the unsupervised domain adaptation in the two stages separately. Secondly, the results of an optimised solution are discussed. Finally, we demonstrate the ability of our approach to address the ultimate goal, i.e. to classify barely visible objects under poor illumination conditions.

A. The effectiveness of Unpaired image-to-image translation

Figure 5 shows some thermal images translated using our image-to-image translation model. As a comparison, we also offer images translated by a model trained with the traditional generative adversarial learning method. Benefits from the correspondence between the RGB data and the thermal data of the Freiburg dataset, we are also able to compare the translated images with the real thermal images.

It can be seen that a model trained with the traditional GAN approach does not generate convincing thermal images. Still, with the use of the cycle consistent architecture, the translated thermal images are closer to the real thermal images, while the structure and content information of the original RGB images are significantly preserved. This indicates that visually our method is effective in adapting the RGB domain to the thermal domain. However, by comparing them with real thermal images, we observe that there are still differences in how the objects are presented, although they look similar. The presentation of objects in a translated image is not defined purely by the temperature of the objects, but also by the shade of their color. Specifically, objects with a darker color tend to be translated as higher temperature objects, i.e., they appear whiter on the thermal image.

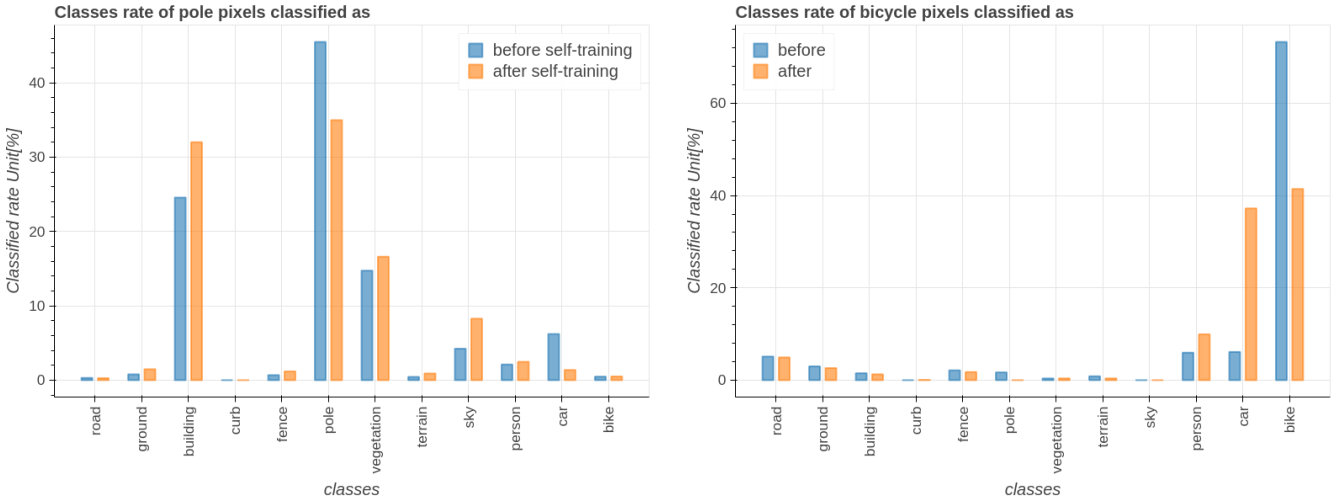
Positively, this difference in object representation did not have a disproportionate impact on the semantic segmentation task. We train the target domain segmentation model with the translations generated by the translation model. When the model is evaluated with the test dataset, the model achieves a mIoU score of 52.1. Class-wise IoU score can be seen in Table II. Some qualitative results are shown in Figure 4. By observing the translated dataset, we found the following explanation for this phenomenon. The translated thermal images are not generated strictly according to the temperature but are influenced by the color of the objects at the same time. This may cause similar objects to appear in the translation as having significant temperature differences. Yet, these objects exist in different shades of color, such as the vehicles in the picture, this ensures that the correct temperature conversion exists in the translation. In other words, we can summarise this as the target domain is a subset of the domain in which the translated thermal images are located. In this way, we can use the segmentation model trained with translated thermal images to correctly classify the real thermal images as well.

B. The effectiveness of prototypical pseudo label denoising

The results in Table II (row 4 and 5) show that after continuing to refine the target segmentation model using the self-training with prototypical pseudo label denoising method, we get a 3.7 improvement in mIoU scores with respect to only using CycleGAN. Furthermore, the individual IoU scores for most of the classes have improved. This proves that this method is effective for RGB-to-thermal domain adaptation.

(a) Thermal image (b) Groundtruth (c) Iteration 0 (d) Iteration 1000 (e) Iteration 5000

Fig. 6: Iteratively degenerated poles in the self-training process.



(a) Rate of pole pixels classified as different classes (b) Rate of bike pixels classified as different classes

Fig. 7: Rate of pole and bike pixels classified as different classes.

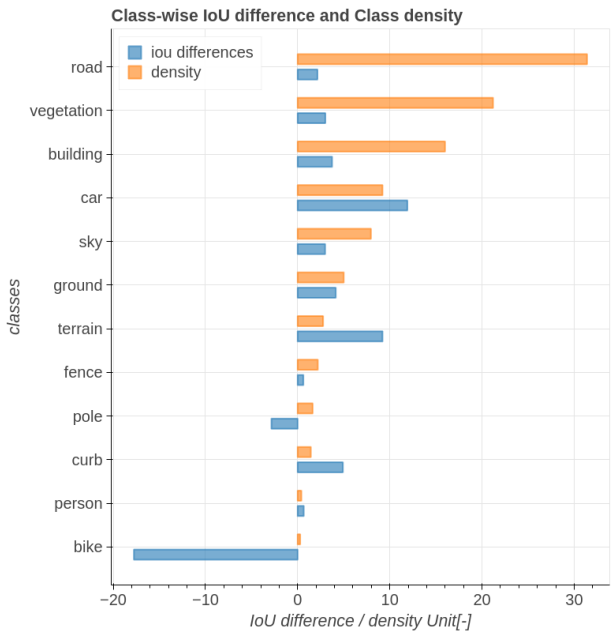


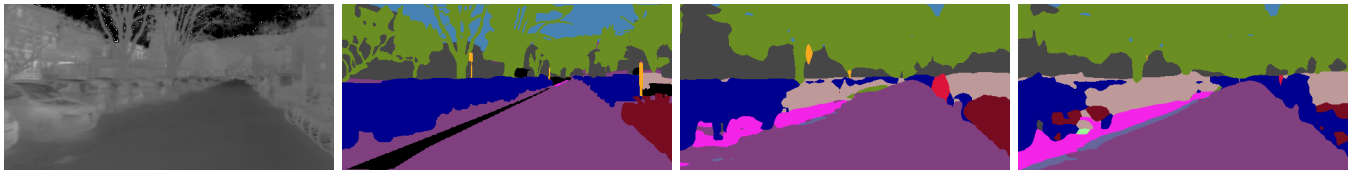
Fig. 8: Class-wise IoU differences and class densities. Classes are sorted by density. Low density classes pole and bike show negative IoU differences before and after self-training.

With the benefit of rectification based on prototypes, we were able to produce more reasonable pseudo labels. Also, the online update of prototypes allows us to further correct the pseudo labels. However, it is worth noting that two classifications, namely poles and bicycles, underwent degradation.

Figure 8 shows that these two classes are two of the most sparse classes in the Freiburg dataset. Sparse classes are those with low densities, where density is the number of pixels belonging to a particular class as a proportion of the total number of pixels in the dataset. Additionally, Figure 7a demonstrates that the pixels of poles are mostly misclassified as buildings and vegetation as measured by their class rate. The class rate is the fraction of pixels belonging to a pole or a bike that are classified as different classes. This is legitimate since buildings and vegetation have a lot of overlap with poles in the actual scene. This can also be observed in Figure 6, where poles are gradually degenerated to buildings. Similarly, we also find in Figure 7b that after fine-tuning, more pixels belonging to bicycles are misclassified as vehicles and people, which again have a lot of overlap with bicycles. More examples can be seen in Figure 9 and 10. We can therefore conclude that the prototypical pseudo label denoising method has a positive effect on most classes in the task of thermal translation-to-thermal domain adaptation, but can cause some sparse classes to degenerate into overlapped ones.

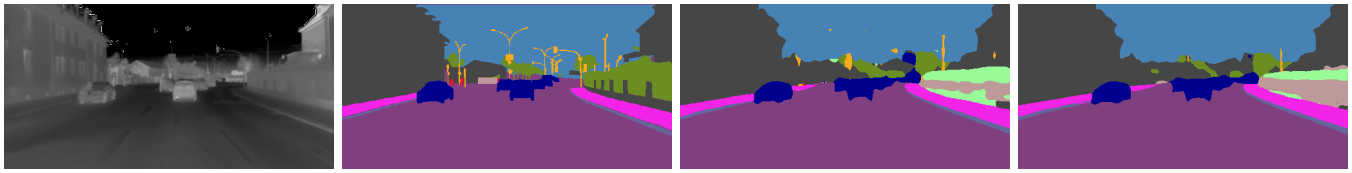
C. Joint classification and night time thermal images classification

Considering the degeneration of the pole and bike classes after self-training, we combined the models before and after self-training, using the model without self-training to classify pole and bike alone, while the other classes were classified using the model after self-training. This gives us the best solution we can produce, which is a mIoU score of 56.3.



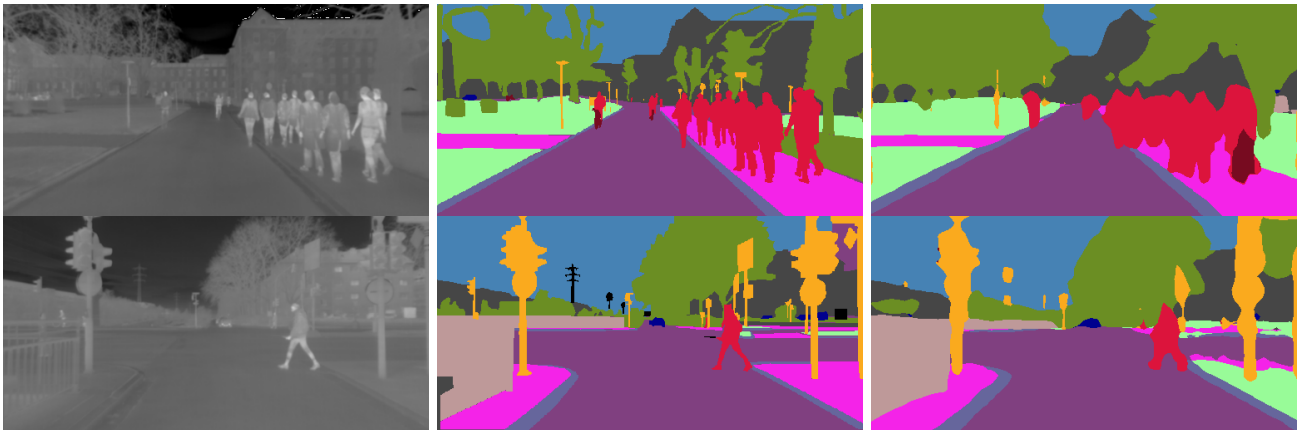
(a) Thermal image (b) Groudtruth (c) CycleGAN classification (d) Self-training classification

Fig. 9: Degeneration of bicycle. The bikes on the right hand side are misclassified as cars after self-training.



(a) Thermal image (b) Groudtruth (c) CycleGAN classification (d) Self-training classification

Fig. 10: Degeneration of pole. The poles in the middle of the image are misclassified as sky, building and vegetation.



(a) Thermal image (b) Groudtruth (c) Joint classification

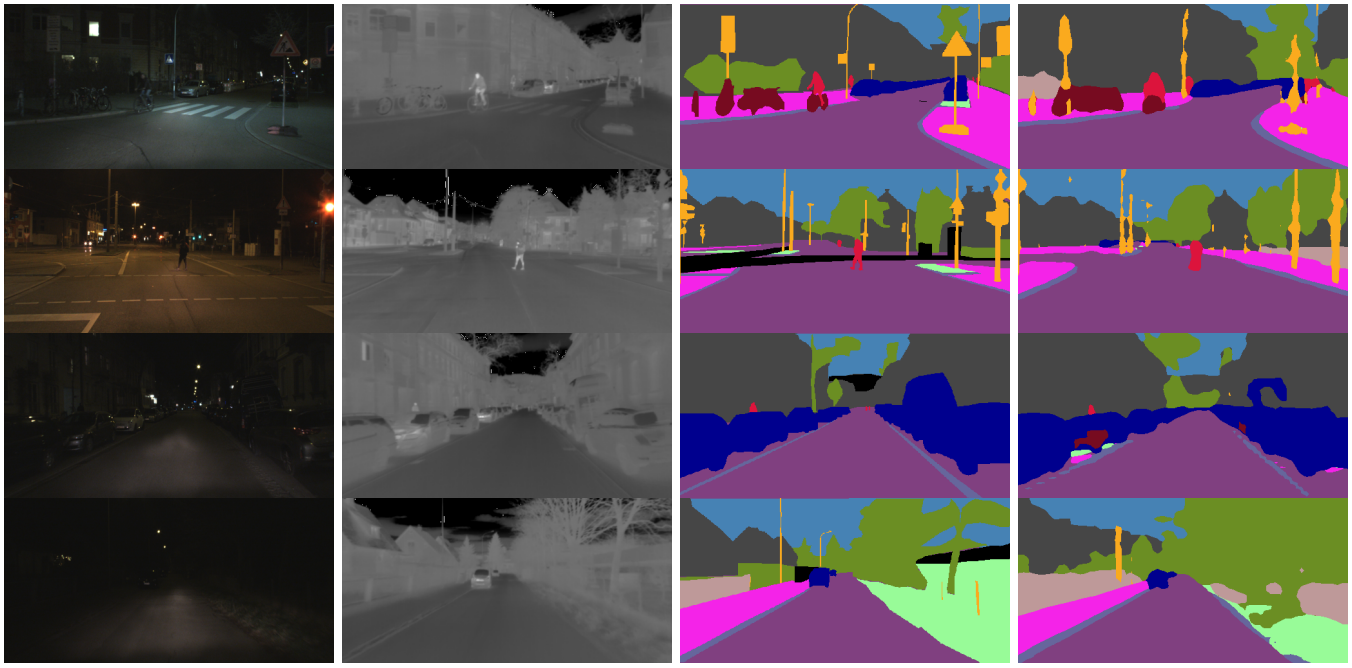
Fig. 11: Joint classification result

The qualitative and quantitative results are shown in Figure. 11 and Table. II, respectively. It is necessary to be aware that as the optimal solution requires two models to predict the data simultaneously, the required computing power and computation time are doubled.

The ultimate goal of the study of thermal semantic segmentation is to train models that can effectively understand their surroundings in the presence of poor illumination. We hence also use night-time thermal test data to evaluate our optimal model. Due to the difference in temperature between daytime and nighttime, thermal images captured at different times of the day give a different interpretation of the scene. In thermal images, an object tends to have a darker colour at night relative to daytime, i.e. it has a lower temperature. However, the difference between daytime and nighttime for thermal images is much smaller than for RGB images. The results show that although the mIoU scores (52.5) have dropped relative to the daytime thermal test data (56.3), our segmentation model still can make decent classifications (see Figure 12).

VI. Conclusion

In this paper, we propose an unsupervised domain adaptation method for thermal semantic segmentation, in which we adapt the RGB domain data to the thermal domain. This allows training a thermal semantic segmentation network without requiring per-pixel labeling of thermal images nor needing to have a 1-on-1 matching RGB images for each thermal image. Our approach resorts to unpaired image-to-image translation and self-training with prototypical pseudo label denoising technique. The proposed method essentially increases the semantic segmentation performance compared to without domain adaptation. It is worth noting that the method used in this article has limited performance for minority classes, such as pole and bicycle. Future research could be undertaken to constrain the self-training process for minority classes to ensure that the learning of these classes is successful. Secondly, the current approach ends up learning a joint domain of translations of the RGB domain and the thermal domain. Thus, another research direction could investigate narrowing down the learned domain to the real thermal domain, or attempting to generate more realistic translations.



(a) RGB image (b) Thermal (c) Groundtruth (d) Classification

Fig. 12: Night-time thermal images classification

REFERENCES

[1] A. Williams and P. Fischer, *Pedestrian traffic fatalities by state: 2017 preliminary data*. DC: Governors Highway Safety Association, 2017.

[2] M. Toldo, A. Maracani, U. Michieli, and P. Zanuttigh, "Unsupervised domain adaptation in semantic segmentation: a review," *Technologies*, vol. 8, no. 2, p. 35, 2020.

[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[4] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6936–6945.

[5] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 289–305.

[6] Y. Zou, Z. Yu, X. Liu, B. Kumar, and J. Wang, "Confidence regularized self-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5982–5991.

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 414–12 424.

[8] P. Li, X. Liang, D. Jia, and E. P. Xing, "Semantic-aware grad-gan for virtual-to-real urban scene adaption," *arXiv preprint arXiv:1801.01726*, 2018.

[9] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. PMLR, 2018, pp. 1989–1998.

[10] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4085–4095.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[14] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

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

[16] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[17] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[18] 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," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[19] H. Xiong, W. Cai, and Q. Liu, "Mcnet: Multi-level correction network for thermal image semantic segmentation of nighttime driving scene," *Infrared Physics & Technology*, vol. 113, p. 103628, 2021.

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

[21] 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 *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 5108–5115.

[22] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Adversarial dropout regularization," *arXiv preprint arXiv:1711.01575*, 2017.

[23] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3723–3732.

[24] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[25] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[26] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," *Advances in neural information processing systems*, vol. 29, pp. 469–477, 2016.

[27] S. Yu, H. Dong, F. Liang, Y. Mo, C. Wu, and Y. Guo, "Simgan: Photo-realistic semantic image manipulation using generative adversarial networks," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 734–738.

[28] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526.

[29] M. Chen, H. Xue, and D. Cai, "Domain adaptation for semantic segmentation with maximum squares loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2090–2099.

[30] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2020–2030.

[31] D. Dai, C. Sakaridis, S. Hecker, and L. Van Gool, "Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding," *International Journal of Computer Vision*, vol. 128, no. 5, pp. 1182–1204, 2020.

[32] K.-H. Lee, G. Ros, J. Li, and A. Gaidon, "Spigan: Privileged adversarial learning from simulation," *arXiv preprint arXiv:1810.03756*, 2018.

[33] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Dada: Depth-aware domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7364–7373.

[34] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3722–3731.

[35] J. Vertens, J. Zürn, and W. Burgard, "Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 8461–8468.

[36] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.