RGB-only day-night semantic segmentation using domain adaptation with thermal images

Sun, Jia

Award date:
2022

Technische Universiteit
Eindhoven
University of Technology

Automotive Technology, Department of Mechanical Engineering
Signal Processing Systems Group, Department of Electrical Engineering

RGB-only day-night semantic segmentation using domain adaptation with thermal images

Master Thesis Report
Report Number: 1438

Jia Sun, 1486683

Assessment Committee

Chairman	prof.dr.ir. Peter de With
Member 1	dr. Gijs Dubbelman
Member 2	dr. Jos Elfring
Advisor	dr. Pavol Jancura
Supervisors	
Supervisor 1	dr. Gijs Dubbelman
Supervisor 2	dr. Pavol Jancura

Eindhoven, March 2022

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 3/15/22

Name Jia Sun

ID-number 1486683

Signature Jia Sun

Insert this document in your Master Thesis report (2nd page) and submit it on Sharepoint

[i] See: http://www.tue.nl/en/university/about-the-university/integrity/scientific-integrity/
The Netherlands Code of Conduct for Academic Practice of the VSNU can be found here also.
More information about scientific integrity is published on the websites of TU/e and VSNU

RGB-only day-night semantic segmentation using domain adaption with thermal images

1st Jia Sun 1486683
Automotive Technology, Mechanical Engineering
Eindhoven University of Technology
Eindhoven, the Netherlands
j.sun@student.tue.nl

***Abstract*—Most current algorithms and models are optimized for daytime scenes, but little attention is paid to the nighttime situations. However, in real-world driving scenarios, vehicles need to face both situations and even some adverse weather or other weakly-illuminated scenes. Therefore, in this work, we propose to use a multi-head network to perform semantic segmentation on both daytime and nighttime scenarios. Due to the lack of annotation for nighttime data, we adopt domain adaptation technique, specifically adversarial learning with a domain discriminator, to narrow the domain gap. In this process, we first apply the multi-head network to the RGB-to-Thermal adaptation task, and then transfer it to solving the shift between day and night with a modified decoder structure, which helps to boost the performance on nighttime domain by 2%. During our study, we refer to the work of HeatNet [3], which proposes a multimodal network to bridge the day-night gap with thermal images. As a final result, we compare our multi-head approach as the unimodal network with HeatNet RGB-only model on Freiburg dataset, and our network can achieve the same performance as it. However, to create the unimodal network, we do not need to train a multimodal network at first except pre-training a RGB and a thermal teacher models. Compared to HeatNet, our multi-head network is also more efficient to train because there are less parameters.**

I. Introduction

An autonomous car is a vehicle capable of sensing its environment and operating without human involvement. It can go anywhere a traditional car goes and do everything that an experienced human driver does. To achieve this goal, a precise and robust semantic segmentation for urban scenes plays an important role in the perception of a self-driving system. However, there are still many factors that hinder the popularization of autonomous driving. One crucial challenge is how to help the vehicle perceive its surroundings quickly and precisely.

A vehicle needs to face both day and night, well and weakly illuminated conditions while driving. Currently, great progress has been made in RGB image semantic segmentation for autonomous driving scenarios [1] [2], but most of them are present in well illumination or daytime conditions. Due to the large gap between day and night within RGB modality, these methods are difficult to generalize well to nighttime scenes. Besides, there are few datasets for nighttime urban scenes, and this also hinders the development at this area. Nowadays, most vehicles use optical cameras to detect objects on the roads, but they require sufficient illumination. As consequence, they cannot work well at night and darkness. In this case,

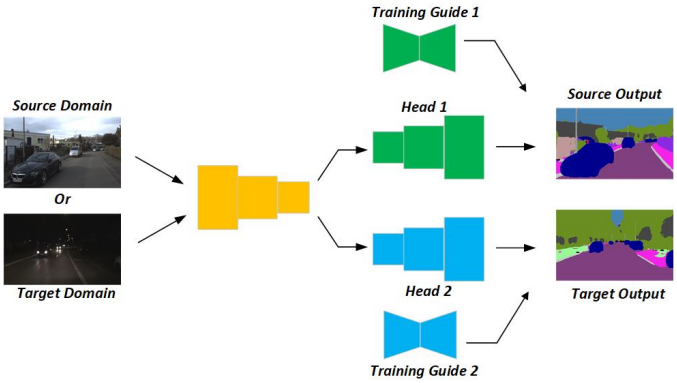

Fig. 1: Our multi-head network makes use of two head networks to segment the input images from different domains. In each domain, a pre-trained model guides the training of the segmentation network.

a thermal camera can provide complementary information, since it captures infrared radiation that objects emit to the surroundings so that its performance is less influenced by weather and illumination. Nonetheless, compared to optical cameras, thermal ones are usually more expensive, and they are not always installed on vehicles. These factors motivate us to pay more attention to semantic segmentation for nighttime RGB images. At the same time, inspired by HeatNet [3], we aim to create a unimodal segmentation network that only uses RGB images as inputs to make semantic prediction. In this process, we also notice that people usually utilize one common head network which is also known as the decoder to fulfil the adaptation task of computer vision, such as classification and segmentation, between different domains. However, in this case, it is assumed that the decoder needs to make a compromise to restore the information from different fields, which may impair the performance on a certain domain. Encouraged by this assumption, we propose a multi-head structure, and it is expected to bring positive influences to achieving adaptation between different domains [4].

As mentioned above, it is difficult to find a dataset that is published specifically for nighttime urban scene. Recently, the authors of HeatNet [3] also released the corresponding Freiburg dataset that contains RGB and thermal images of day- and night-time automotive scenes. Since there is no annotation

for nighttime data, domain adaptation technique is adopted to transfer the knowledge of daytime to the nighttime domain. Such an approach [21] [22] is usually used to narrow the domain gap between a labeled source domain and a target domain, where supervised learning is impossible.

In this work, we first reproduce the result of HeatNet, and also consider it as a reference for our own approach. Then, we propose the multi-head network to perform RGB-only semantic segmentation on day and night with the help of adversarial learning. To bridge these two domains, we also employ thermal features during the training.

The main contribution of this work is the application of multi-head neural network architecture to perform semantic segmentation for daytime and nighttime RGB images. We propose a training strategy that first pre-trains the architecture by RGB-to-Thermal domain adaptation step, and then it is trained to perform daytime-to-nighttime domain adaption. We summarize all the contributions as follows:

- The thesis tries to reproduce the work of HeatNet method. Although the results are partially reproduced, the work is used for a comparison with our approach.
- The thesis proposes a multi-head neural network architecture for semantic segmentation task which can narrow the domain gap between different domains.
- The thesis demonstrates the effectiveness of the proposed multi-head approach as the final unimodal network that can well perform the semantic segmentation on daytime and nighttime RGB images. It achieves the same performance as HeatNet but with a more efficient training process.

II. Related work

A. Domain adaptation for semantic segmentation

Classical machine learning techniques assume that the training and test samples are from the same distribution so that the model trained on the training set can be applied to the unknown samples directly. However, this assumption can not always hold in all cases, especially when the training and test data are collected from different sources. Due to the domain discrepancy, the trained model may not generalize well to the target domain. To solve this type of problem, domain adaptation (DA) is proposed and explored in many sub-fields within machine learning. Unsupervised domain adaptation (UDA) [18] is one branch within DA, and it aims at the specific situation when there are only labels available for training data but few or no labels for domain of interest.

There are already many works that study on UDA for semantic segmentation [23] [24]. Markus Wulfmeier et al. [5] proposed adversarial domain adaptation. They trained a supervised task module and encoder to maximize the likelihood of source labels given the source inputs. At the same time, to adapt the network to the unknown target domain, they trained the encoder to confuse a domain discriminator that is responsible for distinguishing the domain of inputs. This method has been considered as a benchmark in UDA. Based on this work, Yi-Hsuan Tsai et al. [6] proposed to adapt the structured output space instead of feature space to transfer the structured spatial knowledge. This paper also argued that multi-level adaptation can improve the segmentation result further. Apart from adversarial learning technique, there is another type of method to achieve UDA. Parallel CNN architectures such as Siamese network have been verified to be effective for learning invariant features [7] [17]. Therefore, in addition to the loss of the supervised task, an extra domain loss is applied to minimizing the distance between domains. The distribution divergence is usually measured by Maximum Mean Discrepancy (MMD) [19] [20], but it lacks the strong semantic representation so that it is rarely used in the field of semantic segmentation. To adapt this method to semantic segmentation, ADVENT employed Entropy Minimization, which constrains the model such that it produces high-confident prediction on target-like samples, in their work [8].

B. The work of HeatNet

Johan Vertens et al. [3] proposed a novel multimodal approach for daytime and nighttime image segmentation with the help of adversarial learning, leveraging both RGB and thermal images while not requiring annotations for nighttime RGB or thermal infrared images. Fig. 2 illustrates the specific architecture of HeatNet.

The whole architecture is based on PSPNet [1] with ResNet [12] backbone. In the backbone network, there are two parallel encoders to extract features from RGB and thermal inputs, respectively. They consist of the first three stages of ResNet. After that, two features are concatenated and then passed through the remaining layers to produce the final segmentation prediction. The discriminator network is a fully convolutional network, and its main function is to distinguish which domain the input is from. However, the objective is to confuse it so that it can not discriminate the input image's origin. If this confusion can be achieved, the segmentation network trained on daytime domain can be transferred to nighttime. This is fulfilled by adding another adversarial loss, which is applied only to nighttime images.

As for the training scheme, it can be separated into two steps, which is very similar to the generative adversarial network [13] [25] [26]. In the first step, the parameters of the discriminator are frozen, the segmentation network is trained using segmentation loss and adversarial loss; in the second step, the case is reversed, and the discriminator is trained using discriminator loss while the segmentation network is frozen.

The above is the basic architecture and training scheme of HeatNet. As an extension, the paper also proposed two-stage training. Since there is a large shift between day and night within RGB modality, the gap within thermal modality is much smaller. The authors proposed to train the segmentation network merely in the first stage with supervision provided by the RGB teacher for daytime images and the thermal teacher for nighttime images. In the second stage, the normal training procedure including domain adaptation, as described above, is continued.

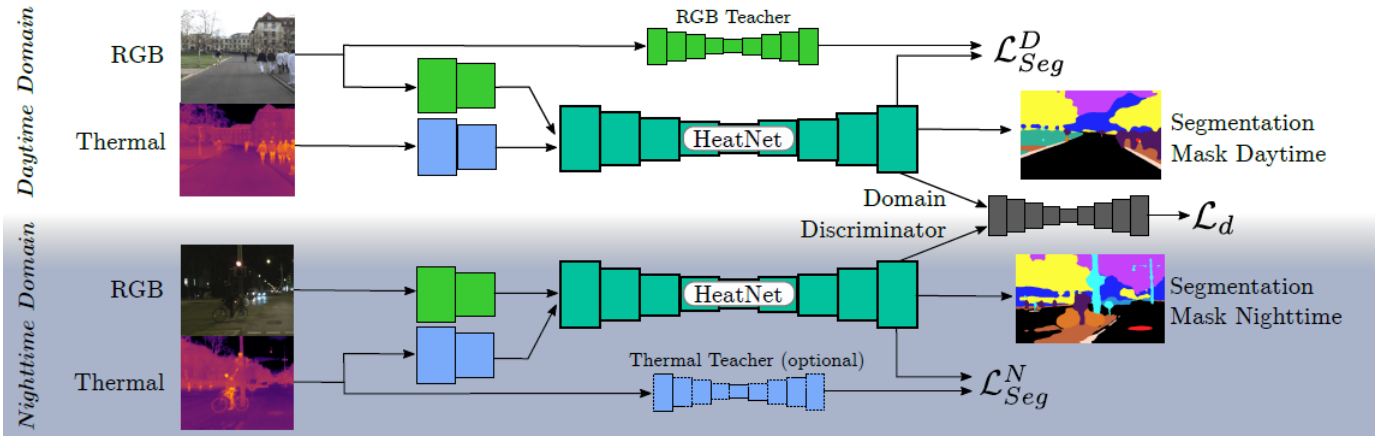

Fig. 2: The architecture of HeatNet [3]. HeatNet is a multimodal network that uses RGB and thermal images as inputs. The network is trained with the supervision produced by the pre-trained RGB teacher model for the daytime images and the supervision produced by the pre-trained thermal teacher model for the nighttime images. The discriminator is used to narrow the domain gap by training with the adversarial loss.

III. Methodology

A. Single-head network

Single-head network is the most frequently used framework in domain adaptation currently. It consists of one common encoder and also a shared decoder. In the work [27] [28], it can be found that only a single decoder is placed in the network. In this work, a single-head network is considered as a baseline. We first apply this method to study the task of RGB-to-Thermal adaptation. We consider the following two domain adaptation strategies for a single-head neural network:

1) Maximum squares loss: In the work of ADVENT, entropy minimization is used to enforce the network to produce high-confident predictions on target-like images, which is formulated as follows:

$$\mathcal{L}_{ent}(x_t) = \frac{-1}{log(C)} \sum_{h,w} \sum_{c=1}^{C} P_{x_t}^{(h,w,c)} log P_{x_t}^{(h,w,c)}, \quad (1)$$

where C denotes the number of object classes, and x_t denotes the input image from target domain. $P_{x_t}^{(h,w,c)}$ represents the C-dimensional prediction from the segmentation network. However, it brings a problem: probability imbalance. Specifically speaking, in the training process, the gradient is dominated by samples with high probability so that the training on those with lower probability is relatively ignored. To solve this problem, Minghao Chen et al. [9] proposed an improved loss for domain adaptation of semantic segmentation: Maximum Squares Loss. The loss is formulated as:

$$\mathcal{L}_T(x_t) = -\frac{1}{2N} \sum_{n=1}^{N} \sum_{c=1}^{C} (p_t^{n,c})^2, \quad (2)$$

where N denotes the total number of pixels in an image, C denotes the total number of object classes in the dataset, and $p_t^{n,c}$ represents the model prediction probability of the class c at point n for sample x_t. For the simplicity of analysis, we just take the binary case for an example. In this case, the gradient curves of two functions are plotted in Fig. 3.

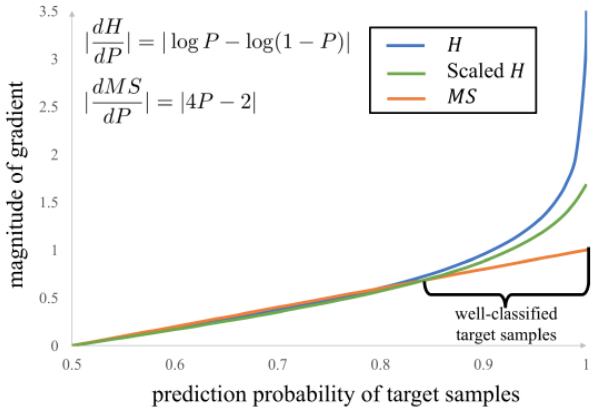

Fig. 3: The gradient curves of entropy minimization and maximum squares loss [9].

In Fig. 3, it can be seen that within the modality of entropy minimization, the gradient of points with high prediction probability is much larger than those lie in other ranges. Hence, the training effect is dominated by samples with high certainty and those with low confidence are ignored. In contrast, maximum squares loss has a linear gradient curve. Hence, there is no large difference between the gradients of points with high or low probability, while a point with higher probability still preserves the higher gradient than a point of less probability. Therefore, in this work, we decide to use maximum squares loss as our first baseline experiment for the single-head method.

2) Adversarial learning based method: Adversarial learning is a machine learning technique that attempts to deceive models by providing deceptive inputs or pseudo labels. Gen-

erally speaking, the objective is to train the network so that it can not distinguish which domain the input image is from. This strategy has been illustrated in the work of HeatNet, and we adopt this directly to execute our second baseline experiment. This time, we adopted DeepLab v2 [14] as our basic framework, and the architecture of the discriminator is the same as HeatNet.

Our training process is also separated into two stages. In the first stage, we just optimize the segmentation network using segmentation loss on RGB images; then we apply the adversarial learning strategy. As an extension, we execute another experiment where we also calculate the segmentation loss for thermal images using the supervision provided by the RGB teacher model in the first stage as well.

B. Multi-head network

1) The basic architecture: In the network of semantic segmentation, such as PSPNet and DeepLab, the head network, also known as the decoder, usually consists of a module that is formed by multi-scale dilated convolutions and an interpolation function to restore the resolution. However, in most works of domain adaptation involving multiple domains, only one decoder is placed in the network, but some works have proven that multi-head configuration is of much potential. The work of Shota Masaki et al. [4] proposes a semantic segmentation model that involves using a multi-head network. In their method, for each domain, an output head is assigned to it. By preparing an output head specific to each domain, datasets with different object classes can be trained simultaneously. Inspired by this layout, we propose to employ multi-head network to fulfil adaptation between different domains.

Since inputs from different domains are passed through the common encoder, it ensures the encoder is learning domain-invariant features from both inputs. However, when only a decoder is placed there, it may lead to an underlying competition because it needs to learn how to restore information for different domains, which may impair the performance on a certain domain. It is assumed that two decoders can help improve the result further because they only focus on their own features so that they are able to extract meaningful information from both domains. Such a design could give the decoders more freedom to learn their interested information. Therefore, we place two identical decoders in the network and each is responsible for one modality, respectively. In this work, we explore the effect of the multi-head network to domain adaptation on two adaptation tasks: RGB-to-Thermal and Day-to-Night.

Fig. 4 shows the overall architecture of the multi-head network. The general framework is based on DeepLab v2 with ResNet backbone. A decoder contains an ASPP module and a bilinear interpolation function subsequently. In Fig. 4, it can be observed that images from both source and target domains, namely I_{source} and I_{target}, are passed through a common encoder to get the extracted features at first. After that, the features are sent to their unique decoders to produce the final prediction P_{source} and P_{target}. With the prediction, segmentation loss can be calculated on both domains using the supervision generated by the teacher models. As for the discriminator C, it has as inputs the softmax activation S_{source} or S_{target} of the segmentation model, and its function and the computation of loss are the same as before.

We also adopt two-stage training to train this network. In the first stage, we only optimize the network with segmentation loss on both domains, and the loss is calculated using:

$$\mathcal{L}_s = -\frac{1}{HW}\sum_{h,w}\sum_{c=1}^{C} P^{(h,w,c)} log P^{(h,w,c)}, \tag{3}$$

where C denotes the total number of classes and H, W denotes the height and width of the output, respectively. $P^{(h,w,c)}$ represents the C-dimensional prediction from the segmentation network.

In the second stage, specifically, the optimization objective of the multi-head network can be formulated as:

$$\mathcal{L} = \begin{cases} \mathcal{L}_s^S + \lambda \mathcal{L}_{adv}, & \textit{the first step} \\ \mathcal{L}_d, & \textit{the second step} \end{cases} \tag{4}$$

$$\mathcal{L}_d = \frac{1}{HW}\sum_{h,w} \begin{cases} [0 - C(S_X)]^2, & X{=}S \\ [1 - C(S_X)]^2, & X{=}T \end{cases} \tag{5}$$

$$\mathcal{L}_{adv} = \frac{1}{HW}\sum_{h,w} [0 - C(S_X)]^2, \quad X{=}T \tag{6}$$

where $\mathcal{L}_s^S$ represents the segmentation loss on the input from the source domain, $\mathcal{L}_{adv}$ represents the adversarial loss on the input from the target domain, and $\mathcal{L}_d$ denotes the loss of the discriminator. λ is the weighting factor of the adversarial loss.

As for the training process, it is worth mentioning that we define two optimizers for the two heads, respectively. Specifically, when a source image is inputted, Optimizer 1 updates the parameters for the common backbone and the source-domain head; when it turns to a target input, Optimizer 2 updates the parameters for the common backbone and the target-domain head.

2) The modified architecture: Apart from the above basic multi-head network, we also make a change to the network to improve its performance. In previous networks, the decoder consists of a module formed by multi-scale dilated convolution layers and a following interpolation function. Because the down-sampling ratio of the backbone is 8, the interpolation function also up-samples the feature eight times directly. To improve the learning capability of the decoder, some additional convolutional layers are added.

Fig. 5 shows the architecture of the modified decoder. The feature is up-sampled for three times, and each time the ratio is two. After each up-sampling operation, a convolutional layer is inserted which will not change the dimension of the feature. The inserted convolution operation is expected to learn how to make segmentation prediction further.

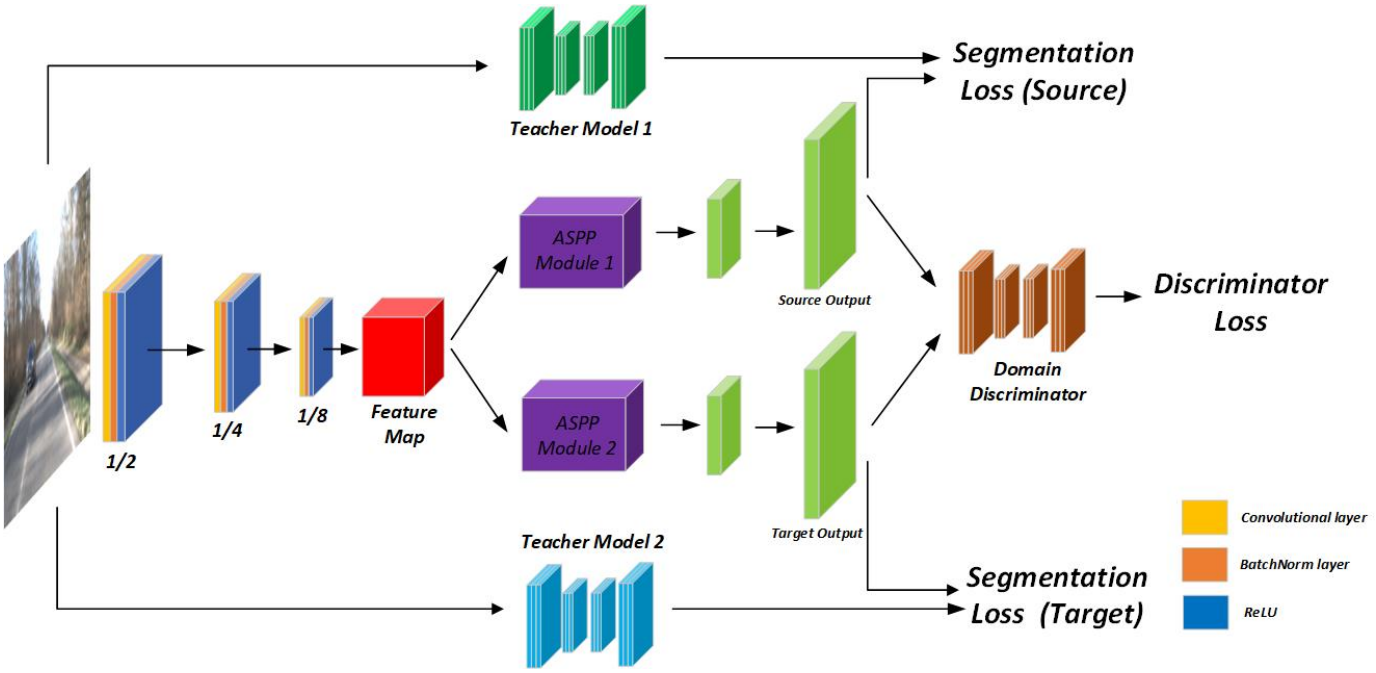

Fig. 4: The overview of the multi-head network. Each time, we sample a batch of images from the source and target domains, respectively. The images pass through a common encoder, and then through their corresponding decoders. We optimize the network with two pre-trained teacher models, and minimize a domain confusion loss from the discriminator to narrow the gap between two domains at the same time.

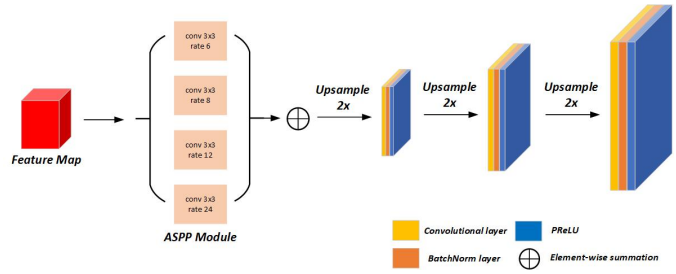

Fig. 5: The modified decoder. After the ASPP module, we insert three convolutional layers in the upsampling operation. We upsample the feature twice for three times, and insert a layer (conv + BN + PReLU) after each upsampling.

C. Application study

1) From RGB to thermal domain: Fig. 6 shows the specific configuration of the multi-head network for RGB-to-Thermal adaptation. The training data are daytime RGB and thermal images of Freiburg dataset. It is worth mentioning that since they are inputted to the same encoder, their channels should be the same. However, thermal images in Freiburg dataset is one-channel. To make them uniform, the channel of each thermal image is duplicated three times before inputting to the network. The pre-trained RGB teacher model provides supervisions for inputs of both domains, since each RGB image and its thermal pair of Freiburg dataset are time-synchronized to the same view.

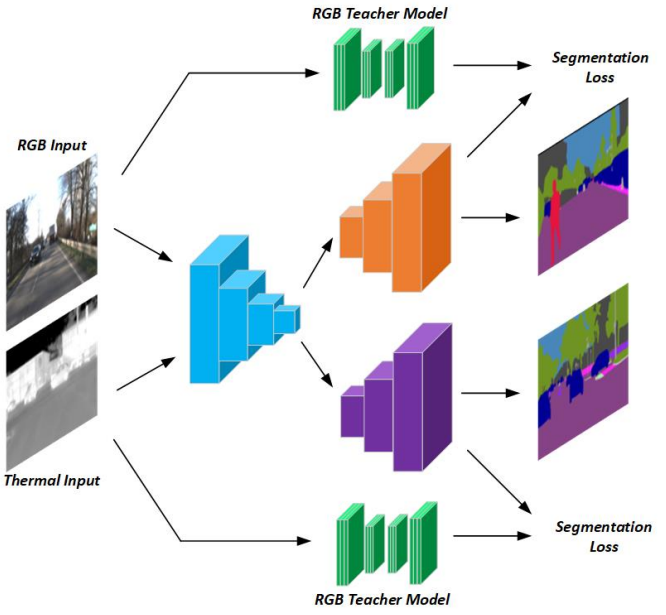

Fig. 6: The configuration of the multi-head network for RGB-to-Thermal adaptation. We only use daytime images here. Since each RGB and its thermal pair are time-synchronized, the RGB teacher model can provide supervisions for them simultaneously.

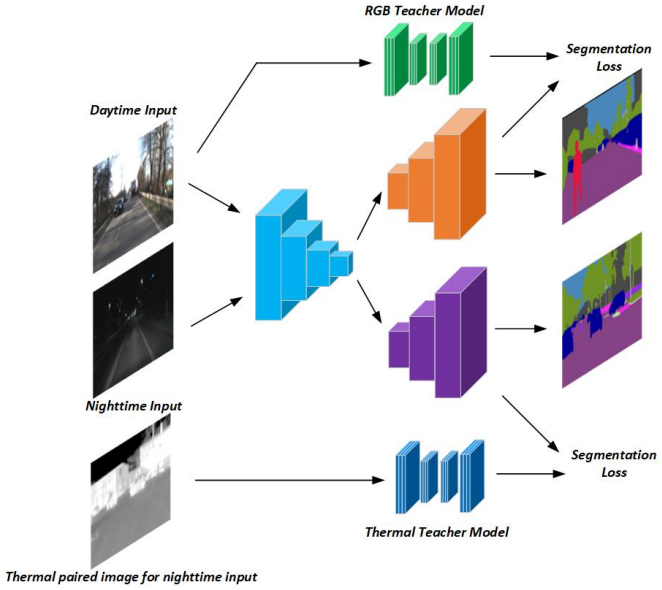

Fig. 7: The configuration of the multi-head network for Day-to-Night adaptation. Since this is a unimodal network, we use daytime and nighttime RGB images as its inputs. This time, the pre-trained RGB teacher can still guide the training of the daytime input, but we need to employ the pre-trained thermal teacher to generate labels for the nighttime images.

2) From daytime to nighttime domain: As the main objective of this work is to create a unimodal network to achieve adaptation between daytime and nighttime domains, so multi-head network is expected to play this role. This time, the network architecture and training scheme are exactly the same as those in the RGB-to-Thermal adaptation task except the training data. To make the model unimodal, the input data are daytime and nighttime RGB images of Freiburg dataset.

Fig. 7 demonstrates the configuration of this unimodal network. There is no distinct difference from Fig. 6, but the pre-trained thermal teacher model is used to provide supervisions for nighttime images. It is noted that we do not need a thermal paired image when doing inference for a RGB image, although synchronized thermal inputs are required during training.

Pre-trained weights initialization. Since multi-head network is researched in two aspects: RGB-to-Thermal and daytime-to-nighttime, one can apply a transfer learning technique by using the weights trained on RGB-to-Thermal task to initialize the network when training it for daytime-to-nighttime problem. Initialization with some pre-trained weights is a regular method when training a deep neural network, which can help the training converge more quickly and even boost the final result. As there is a large gap between day and night within RGB modality, thermal image is a middle modality to bridge them. Therefore, we argue that the weights trained on thermal domain can boost the nighttime head.

IV. Dataset

There are already many public datasets available for computer vision tasks, such as Cityscapes [2], Mapilliary Vistas [15], BDD 100K [16] datasets and so on. However, most of them are focusing on RGB images, thermal dataset is rarely seen. In this work, the dataset from HeatNet [3] is employed. This paper released a large-scale urban-scene dataset of Freiburg, Germany. The training set consists of 12170 daytime images and 8683 nighttime images, and the test set has 32 daytime and nighttime images, respectively, with each image's corresponding time-synchronized thermal pair included. Freiburg dataset contains highly diverse driving scenarios including highways, densely populated urban scenes, residential areas and rural districts.

All images' resolution is 1920×650. However, there is some black margin in the two sides of each thermal image, so they are cropped to the size of 1280×640 to guarantee that the visual region is valid and the same process is also applied to RGB images. Each RGB image has three channels and each thermal image has only one channel. During all experiments, they are normalized with the mean 0.5 and standard variance 0.5 for all channels.

As the bit depth of thermal images captured by the camera is so large, their pixel values need to be clamped to an interesting range so that they can look like normal images. The specific range is usually chosen empirically, and here the minimal value is set to be 21800 and the maximum value is 23700. In each thermal image, the pixel values which are larger than the maximum are set to be 23700, and the ones that are smaller than the minimum are set to be 21800. The rest pixels keep unchanged. Moreover, each pixel value is normalized within the range of 0 to 1 using Eq. 7.

$$P_{norm} = \frac{P - P_{min}}{P_{max} - P_{min}}, \tag{7}$$

where P_{norm} denotes the normalized pixel value, and P denotes the original pixel value. P_{max} and P_{min} denote the maximum and minimum values of all pixels, respectively.

V. Experimental results

In the following, we present our experimental results for the reproduction of HeatNet and our multi-head network. We evaluate all models on Freiburg test set.

A. Training details

In this work, we implement our network using the PyTorch toolbox with CUDA 11.0, and all experiments are executed on a single NVIDIA GPU. Most are performed on NVIDIA RTX 3090 with 24 GB memory, and several are on RTX A6000 GPU with 48 GB memory. In the experiment for the reproduction of HeatNet, we adopt RMSProp optimizer to train the network for 100 epochs and the initial learning rate is set to be 10^{-4}. We halve it every 30 epochs, and the batch size is 4. In the remaining experiments, we use the Stochastic Gradient Descent (SGD) optimizer with momentum 0.9 and weight decay 5×10^{-4} to train the segmentation

TABLE I: Semantic segmentation performance mIoU(%) of RGB and the thermal teacher models.

	RGB teacher	Thermal teacher
Ours	70.1	56.3
Paper	69.4	57.0

network and Adam optimizer with momentum 0.9 and 0.99 to train the discriminator network. The initial learning rate for segmentation network is 2.5×10^{-4}, and it is 10^{-5} for the discriminator network. During the process, the learning rate is decreased using the polynomial decay with power of 0.9. When the network is trained for two stages, we set 50 epochs for the first stage and 100 epochs for the second one.

B. Evaluation metric

In this work, we evaluate all the experimental results using mIoU on the following 12 object classes: road/parking, sidewalk, building, curb, fence, pole/signs, vegetation, terrain, sky, person, car, bicycle. The mIoU is defined as:

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{8}$$

In the equations above, it is assumed that there are $k+1$ classes in total, including k object classes and one void/ignored class. p_{ij} denotes the amount of pixels of class i inferred to belong to class j. Hence, p_{ii} represents the total number of true positives (TP), while p_{ij} and p_{ji} denote false positives (FP) and false negatives (FN), respectively.

C. Teacher model

The teacher model is used to generate labels for images in Freiburg dataset. Both RGB and thermal teacher models are based on PSPNet with ResNet as backbone. The two teacher models are based on the open-sourced GitHub project of [1].

The RGB teacher model is trained on Mapiliary Vistas, and then it is used to generate labels for each RGB daytime image. The network is trained for 400 epochs. During the training process, SGD with momentum and polynomial learning rate scheduler are adopted. With the pre-trained RGB teacher model, each RGB daytime image has its label. As its thermal pair is time-synchronized, the label can be also used for thermal supervision. In this way, a thermal teacher model is trained in a supervised manner as well. The training network and scheme are the same as training the RGB teacher. The evaluation result is listed in Tab. I.

From the table, it can be found that the mean result of RGB teacher slightly surpasses the result reported in [3], and this can ensure the supervision in the work is at the same level with the original work. As for the thermal teacher, the overall result is worse than that of the RGB one because the supervision produced by the RGB teacher is not 100% accurate. Besides, although the mean mIoU is a bit lower than that in the original paper, it is a very comparable result, which is only 0.7% lower. Therefore, this model can be used in the following task.

D. Overall results

In the following part, we report results of experiments we mentioned above, and they can be divided into three parts.

1) Reproduction of HeatNet: Tab. II shows the comparison between our reproduction result for HeatNet and the paper's original one. Without the thermal teacher, the reproduced result is around 7% lower than the paper reported on daytime domain while almost achieving the same effect on nighttime domain. We note that not all details of HeatNet architecture and training scheme were described in the original paper. As consequence, some variations and minor differences between our implemented version and the one presented by paper can occur. This leads to the difference in performance and our reproduced mIoU results are lower than of the reported ones. When the thermal teacher is applied, the results on both domains improve, but they are still around $4-5\%$ lower than the original HeatNet result. In this experiment, the architecture of the network was corrected, and the parameters' setting was the same as the original paper. The reasons why it performs worse are as follows. Firstly, the batch size 4 is different from that in the paper 8 due to the memory limitation of GPU. Secondly, the method of data argumentation is also different. In our execution, only gaussian blur and random horizontal flip are used; however, in the paper, the authors applied additional random rotation and block drop to images. Besides, because the RGB teacher model is trained on Mapilliary Vistas dataset, and its annotation of object classes is different from that of Freiburg dataset, a label mapping is applied to making the annotation unified. This mapping process was not described in the original paper, either, so it may lead to the worse performance.

2) From RGB to thermal domain: Tab. III presents our results of all experiments on RGB-to-Thermal adaptation task. Although single-head network using maximum squares loss can achieve more than 60% on RGB domain, it can only achieve 31.3% on thermal images, which is a relatively low performance. By contrast, single-head network with adversarial learning boosts the performance on both domains, especially on thermal images by almost 9.3%. When we also apply the thermal teacher, the result on thermal domain improves further by 10% while it decreases by only 1.3% on RGB. This proves that compared to a single target loss minimization, adversarial learning strategy is more effective to domain adaptation on thermal images. Furthermore, our multi-head network boosts the mIoU on thermal domain by 3.2%, while impairing it on RGB by only 1.2%. With the modified decoder, the performance on the thermal domain further increases to 57.3%, and it keeps almost the same on RGB images, which also achieves the highest mean mIoU among all methods.

3) From daytime to nighttime domain: Tab. IV illustrates our results of all experiments on day-to-night adaptation task. Our basic multi-head network achieves 66.9% and 42.7% on day and night, respectively, with the mean value of 54.8%. When the modified decoder is adopted, both results increase

TABLE II: Semantic segmentation performance mIoU(%) of our reproduced HeatNet and of the original paper. The former one represents the reported result, and the latter one represents ours. Missing results are marked with a dash(-).

Method	mIoU-Day	mIoU-Night	Mean mIoU
HeatNet w/o thermal teacher	70.5/63.1	43.2/43.0	56.9/53.1
HeatNet	70.8/65.7	59.0/54.8	64.9/60.3
HeatNet RGB-only	-/65.6	-/46.0	58.0/55.8

TABLE III: Semantic segmentation performance mIoU(%) on RGB-to-Thermal adaptation. The abbreviations "SH", "Adv", "Seg." and "IR" stand for single-head, adversarial learning, segmentation loss and infrared thermal images. The "Multi-head*" denotes the multi-head network with the modified decoder. The "HeatNet*" represents our reproduced result of HeatNet.

Method	mIoU-RGB	mIoU-Thermal	Mean mIoU
SH-Maximum squares loss	61.5	31.3	46.4
SH-Adv	66.4	40.6	53.5
SH-Adv(Seg. on IR)	65.1	50.5	57.8
Multi-head	63.9	53.7	58.8
Multi-head*	63.5	57.3	60.4
HeatNet	70.8	-	-
HeatNet*	65.7	-	-

by $1-3\%$; when we use the pre-trained weights from RGB-to-Thermal task to initialize the network further, the mIoU on nighttime domain increases by 2% while decreasing by 0.6% on daytime domain. Moreover, it is worth mentioning that the mean value is 58.0% reported, and it is exactly the same as the result of HeatNet RGB-only model. This is a very comparable result, and it proves the effectiveness of our multi-head network. Compared to the training approach of HeatNet RGB-only, we just need to train the multi-head network for 150 epochs, but it requires 200 epochs to train a multimodal network at first, and then even 300 epochs to train a unimodal network.

E. Parameters comparison

Tab. V shows the comparison of parameters' numbers between HeatNet and a multi-head network. It can be found that a multi-head network contains less parameters and its total number is around 18% less than that of HeatNet. Under the circumstance, our multi-head network is more efficient to train.

TABLE IV: Semantic segmentation performance mIoU(%) on daytime-to-nighttime adaptation. The "Multi-head*" denotes the multi-head network with the modified decoder, and the "Multi-head**" denotes "Multi-head*" with the pre-trained weights initialization. The "HeatNet RGB-only*" represents our reproduced result.

Method	mIoU-Day	mIoU-Night	Mean mIoU
Multi-head	66.9	42.7	54.8
Multi-head*	68.5	46.0	57.3
Multi-head**	67.9	48.0	58.0
HeatNet RGB-only*	65.6	46.0	55.8
HeatNet RGB-only	-	-	58.0

TABLE V: The comparison of parameters' numbers between HeatNet and a multi-head network.

	HeatNet	Multi-head network
Number of parameters	54.4M	44.6M

VI. Conclusions and discussion

In this work, we present a novel multi-head network to perform semantic segmentation on both daytime and nighttime urban RGB images. In this process, we adopt adversarial learning strategy to help narrow the gap between different domains, and prove that this method is more effective than merely using a target loss minimization. By using our multi-head network, it can achieve the same performance as HeatNet RGB-only model. Compared with the approach to creating this unimodal network, our multi-head network contains less parameters and we do not have to train a multimodal network to produce supervisions for nighttime data at first. In this case, our model is simpler and more efficient to implement and train. In experiments, we also demonstrate that the weights pre-trained on RGB-to-Thermal adaptation task is beneficial to the training of Day-to-Night, and the thermal head can boost the nighttime head especially.

In the future, we could explore the following directions. (i) We can consider use of multi-level strategy, like the work of [6], to train our model because the training on the lower-level feature can also enhance the adaptation. (ii) We can incorporate the target loss for the target domain in our training strategy to see if we can improve the segmentation results further.

References

[1] H. S. Zhao, J. P. Shi, X. J. Qi, X. G. Wang, and J. Y. Jia. Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2881–2890, 2017.

[2] 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, pages 3213–3223, 2016.

[3] J. Vertens, J. Zurn, and W. Burgard. HeatNet: Bridging the Day-Night Domain Gap in Semantic Segmentation with Thermal Images. In Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2020.

[4] S. Masaki1, T. Hirakawa, T. Yamashita and H. Fujiyoshi. Multi-Domain Semantic-Segmentation using Multi-Head Model. In Proc. of the IEEE Intelligent Transportation Systems Conference (ITSC), 2021.

[5] M. Wulfmeier, A. Bewley and I. Posner. Addressing Appearance Change in Outdoor Robotics with Adversarial Domain Adaptation. In Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2017.

[6] Y. Tsai, W. Hung, S. Schulter, K. Sohn, M. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7472–7481, 2018.

[7] H. Chen, C. Wu, B. Du, and Liangpei Zhang. Deep Siamese domain adaptation convolutional neural network for cross-domain change detection in multispectral images. 2020.

[8] T. Vu, H. Jain, M. Bucher, M. Cord, P. Perez. ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019.

[9] M. Chen, H. Xue, and Deng Cai.Domain Adaptation for Semantic Segmentation with Maximum Squares Loss. In Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV), 2019.

[10] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada. MFNet: Towards Real-Time Semantic Segmentation for Autonomous Vehicles with Multi-Spectral Scene. In Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2017.

[11] Y. Sun, W. Zuo, and M. Liu. RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes. IEEE Robot. Autom. Lett. 4(3), 2576–2583 (2019) 25.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016.

[13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In Advances in neural information processing systems, 2014.

[14] LC. 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. TPAMI 2017.

[15] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder. The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV), 2017.

[16] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2020.

[17] X. L. Chen, and K. M. He. Exploring Simple Siamese Representation Learning. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2021.

[18] K. C. You, M. S. Long, Z. J. Cao, J. M. Wang, and M. I Jordan. Universal domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2720–2729, 2019.

[19] A. Gretton, AJ. Smola, J. Huang, M. Schmittfull, KM. Borgwardt, and B. Scholkopf. Covariate shift and local learning by distribution matching, pages 131–160. MIT Press, Cambridge, MA, USA, 2009.

[20] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Scholkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. In Bioinformatics, 2006.

[21] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474, 2014.

[22] Y. Zhang, P. David, and B. Q. Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In Proceedings of the IEEE International Conference on Computer Vision, pages 2020–2030, 2017.

[23] Y. H. Chen, W. Li, and L. V. Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7892–7901, 2018.

[24] J. H. Yang, R. J. Xu, R. Y. Li, X. J. Qi, X. Y. Shen, G. B. Li, and L. Lin. An adversarial perturbation oriented domain adaptation approach for semantic segmentation. arXiv preprint arXiv:1912.08954, 2019.

[25] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In ICLR, 2016.

[26] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In NIPS, 2016.

[27] L. Sun, K. L. Yang, X. X. Hu, W. J. Hu, and K. W. Wang. Real-time Fusion Network for RGB-D Semantic Segmentation Incorporating Unexpected Obstacle Detection for Road-driving Images. IEEE Robotics and Automation Letters (RA-L), 2020.

[28] L. Y. Deng, M. Yang, T. Y. Li, Y. S. He, and C. X. Wang. RFBNet: Deep Multimodal Networks with Residual Fusion Blocks for RGB-D Semantic Segmentation. arXiv:1907.00135, 2019.