2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

HeatNet: Bridging the Day-Night Domain Gap in Semantic Segmentation with Thermal Images

Johan Vertens*, Jannik Zürn*, and Wolfram Burgard

Abstract— **The majority of learning-based semantic segmentation methods are optimized for daytime scenarios and favorable lighting conditions. Real-world driving scenarios, however, entail adverse environmental conditions such as nighttime illumination or glare which remain a challenge for existing approaches. In this work, we propose a multimodal semantic segmentation model that can be applied during daytime and nighttime. To this end, besides RGB images, we leverage thermal images, making our network significantly more robust. We avoid the expensive annotation of nighttime images by leveraging an existing daytime RGB-dataset and propose a teacher-student training approach that transfers the dataset's knowledge to the nighttime domain. We further adopt a domain adaptation method to align the learned feature spaces across the domains and propose a novel two-stage training scheme. Furthermore, due to a lack of thermal data for autonomous driving, we present a new dataset comprising over 20,000 time-synchronized and aligned RGB-thermal image pairs. In this context, we also present a novel target-less calibration method that allows for automatic robust extrinsic and intrinsic thermal camera calibration. Among others, we use our new dataset to show state-of-the-art results for nighttime semantic segmentation.**

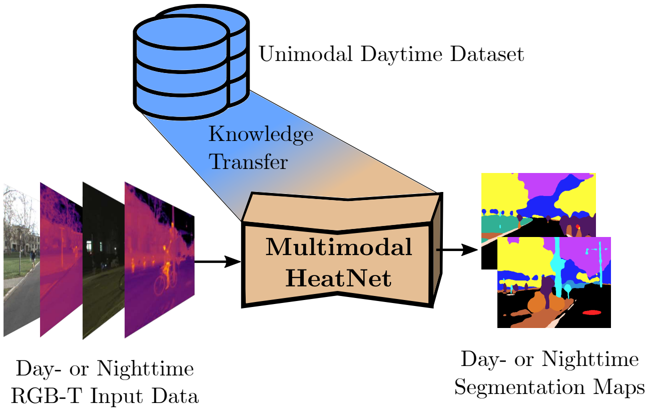
Unimodal Daytime Dataset
Knowledge Transfer
Multimodal HeatNet
Day- or Nighttime RGB-T Input Data
Day- or Nighttime Segmentation Maps

Fig. 1. Our multimodal segmentation network leverages both nighttime and daytime images. We transfer relevant knowledge from a large-scale unimodal daytime dataset for semantic segmentation with a teacher model to our multimodal HeatNet and simultaneously adapt our model to the nighttime domain by unsupervised domain adaptation.

I. INTRODUCTION

Robust and accurate semantic segmentation of urban scenes is one of the enabling technologies for autonomous driving in complex and cluttered driving scenarios. Recent years have shown great progress in RGB image segmentation for autonomous driving [36], [5], which were predominantly demonstrated in favorable daytime illumination conditions. While the reported results demonstrate high accuracies on benchmark datasets [5], [18], these models tend to generalize poorly to adverse weather conditions and low illumination levels present at nighttime as no large-scale nighttime dataset for image-segmentation is publicly available. This constraint becomes especially apparent in rural areas where artificial lighting is weak or scarce. In autonomous driving, to ensure safety and situation awareness, robust perception in these conditions is a vital prerequisite.

Transfer learning and domain adaptation approaches aim at narrowing the domain gap between a source domain, where supervised learning from labelled data is possible, to a target domain, where labelled data is either sparse or not available. Such approaches, as demonstrated in [28] or [35], allow to adapt a given segmentation model to a different domain. These approaches, however, do not leverage a complementary modality such as thermal infrared images that can contain more relevant information to solve a given task in certain environmental conditions than a single modality would provide.

In order to perform similarly well in challenging illumination conditions, it is beneficial for autonomous vehicles to leverage modalities complementary to RGB images [29], [30]. Encouraged by prior work in thermal image processing for object detection [31], object tracking [14], and semantic segmentation [9], [25], we investigate leveraging thermal images for nighttime semantic segmentation of urban scenes. Thermal images contain accurate thermal radiation measurements with a high spatial density. Furthermore, thermal radiation is much less influenced by sunlight illumination changes and is less sensitive to adversary conditions. Existing RGB-thermal datasets for semantic image segmentation such as [9] are not as large-scale as their RGB-only counterparts. Thus, models trained on such datasets generalize poorly to challenging real-world scenarios.

In contrast to previous works, we utilize a semantic segmentation network for RGB daytime images as a teacher model to provide labels for the RGB daytime images in our dataset. We project the thermal images into the viewpoint of the RGB camera images using extrinsic and intrinsic camera parameters that we determine using our novel target-less camera calibration approach. Afterwards, we can reuse labels from this teacher model to train a multimodal semantic segmentation network on our daytime RGB-thermal image pairs. In order to encourage day-night invariant segmentation

*These authors contributed equally. All authors are with the University of Freiburg, Germany. Wolfram Burgard is also with the Toyota Research Institute, Los Altos, USA. Corresponding author: vertensj@informatik.uni-freiburg.de

978-1-7281-6211-9/20/$31.00 ©2020 IEEE

of scenes, we simultaneously train a feature discriminator that aims at classifying features in the semantic segmentation network to belong either to daytime or nighttime images.

Furthermore, we propose a novel training schedule for our multimodal network that helps aligning the feature representations between day and night. Finally, we propose a new way of training a nighttime-daytime RGB-only semantic segmentation network by using thermal images as a bridge modality. In our baseline comparison and our ablation studies, we show that our model achieves comparable performance to fully supervised multimodal models. Additionally, we demonstrate that our first-of-its-kind method significantly reduces the domain gap between daytime and nighttime.

In summary, the contributions of this work are:

- A novel multimodal approach for daytime and nighttime image segmentation, leveraging both RGB and thermal images while not requiring annotations for nighttime RGB or thermal infrared images.
- The *Freiburg Thermal* dataset containing more than 20,000 time-synchronized RGB and thermal images recorded in urban and rural environments both in daytime and in nighttime conditions. We also provide LiDAR pointclouds, accurate GPS data and IMU readings.
- A novel target-less thermal camera calibration approach.
- Extensive qualitative and quantitative evaluation of our approach, including ablation studies.

II. RELATED WORKS

A. Multimodal RGB-Thermal Datasets and Calibration

While unimodal datasets with images in the visible domain are prevalent in computer vision research, some datasets have been proposed that entail aligned RGB-thermal image pairs. Berg *et al.* [1] propose a dataset that consists of thermal infrared images which is mainly targeted towards object tracking. Similarly, Li *et al.* [14], propose a RGB-thermal dataset for multimodal object tracking in varying outdoor settings and conditions. The authors of CATS [26] present a general outdoor dataset for color and thermal stereo disparity estimation. Besides RGB-thermal image pairs, LiDAR-based ground-truth disparity maps are available. Furthermore, the work of Shivakumar *et al.* [23] targets the scenarios of the DARPA Subterranean Challenge providing 894 RGB-thermal image pairs with pixel-wise semantic annotations for underground rescue scenarios. There exist only a few datasets that contain thermal infrared imagery in the context of autonomous driving. In the work of Hwang *et al.* [11], a dataset is proposed that consists of more than 95k RGB-thermal image pairs. Each pair is annotated with bounding boxes for persons and is hence aimed towards pedestrian detection research. The KAIST multispectral dataset [4] entails multiple modalities such as RGB, thermal infrared, LiDAR, GNSS and IMU for a total of 7512 frames. They also provide annotations/ground-truth for 2D bounding boxes, drivable region, image enhancement, depth, and colorization. The authors of MFNet [9] present the first urban scene dataset for multimodal semantic segmentation, comprising 1569 pixel-wise annotated RGB-thermal image pairs. Approximately half of the recorded images were captured during nighttime. However, many of the most common classes in the context of semantic segmentation for autonomous driving such as road, sidewalk, pole, sign, building or sky are not annotated.

Due to the lack of large-scale RGB-thermal datasets for urban semantic segmentation, we propose the Freiburg Thermal dataset comprising over 20000 high-resolution RGB-thermal image pairs in particularly challenging environments. We additionally provide semantic annotations for a distinct test set.

For most previously proposed datasets, distinct RGB and thermal cameras were used and calibrated leveraging hand-made patterns such as checkerboards [26], [23] or lines on printed circuit boards [4]. A different approach was presented by Lussier *et al.* [17] in which an edge response map between depth and thermal images is minimized using grid search over the calibration parameter space.

In contrast to prior work, we propose a method to calibrate the intrinsic, extrinsic and distortion parameters of the thermal infrared camera in a purely target-less fashion, leveraging spatial transformer networks [12] and stochastic gradient descent over a large number of images.

B. Semantic Segmentation of Thermal Images

Recently, semantic segmentation of thermal images began to attract more attention in the computer vision community. Qiao *et al.* [20] use a level set method to detect pedestrians in thermal images. More recently, Li *et al.* [15] proposed an edge-conditioned segmentation network for thermal images, trained supervised on a dataset containing various indoor and outdoor scenes. The works closest to our work are [9] and [25]. In the work of Ha *et al.* [9], the authors propose a multimodal fusion network architecture for RGB and thermal images. They evaluate their approach on their own dataset MF [9]. Similarly, Sun *et al.* [25] propose an RGB-thermal fusion network and show their results on the MF dataset.

In contrast to the works mentioned above, we train an RGB-thermal semantic segmentation model without requiring any manual labeling efforts. We instead use a teacher model trained solely on RGB images to provide supervision for the daytime image pairs. We further present an extended multimodal domain adaptation method that enables robust nighttime segmentation.

C. Domain Adaptation for Semantic Segmentation

Many works in transfer learning explore unsupervised domain adaptation from synthetic data to real environments [27], [3], [33]. Other recent works explore model adaptation from daytime to nighttime via an intermediate twilight domain [6], [22]. Following a different approach, works were proposed that conduct unpaired image-to-image translation using generative models to create synthetic nighttime training data [19], [24], [21]. Most similar to our work, in [32], the authors investigate adversarial domain adaptation, where they use a binary classifier to discriminate between daytime and nighttime image features produced by an encoder network. A domain confusion loss penalizes features that can easily be classified as originating from the daytime or nighttime domain.

In contrast to the above works, our approach leverages additional modalities such as thermal images that provide complementary inputs for semantic segmentation in challenging illumination conditions, significantly narrowing the daytime-nighttime domain gap.

III. TECHNICAL APPROACH

In the following, we describe our approach to multimodal semantic segmentation for daytime and nighttime scenes, leveraging RGB and thermal images. In our approach, we first train a semantic segmentation teacher model in a supervised fashion on the Mapillary Vistas dataset [18]. Subsequently, we use this teacher network to infer labels of daytime RGB images on our multimodal Freiburg Thermal dataset. We then train a student network supervised on the daytime image annotations provided by the teacher model, using both RGB and thermal infrared images. While the thermal modality is mostly invariant to lighting changes, the RGB modality differs significantly between daytime and nighttime and thus exhibits a significant domain gap. We thus further utilize a domain adaptation technique that aligns the internal feature distributions of the multimodal segmentation network, enabling the network to perform similarly well for nighttime images as for daytime images. Note that we do not use any hand-annotated nighttime image labels for training at any time. As thermal cameras are not yet available in most autonomous platforms, we further propose to distill the knowledge from the domain-adapted multimodal model back into a unimodal segmentation network that exclusively uses RGB images. We distinguish between daytime and nighttime in a binary manner, neglecting images taken in twilight.

In the following we detail our approach.

A. RGB-T Semantic Segmentation

We initially train a PSPNet model [36] for semantic RGB image segmentation on the Mapillary Vistas dataset [18], which contains 20,000 RGB images and semantic annotations from highly diverse and challenging urban scenes. We use this model as a teacher model M_D for daytime images, providing pixel-wise semantic annotations for all daytime RGB images in our Freiburg Thermal dataset. Since we project each thermal image into the viewpoint of the RGB camera using the extrinsic and intrinsic camera calibration parameters, described in Sec. IV-A, we can use the same annotations for each respective thermal image. Given the labels produced by M_D, we subsequently train our multimodal RGB-T model M_M by minimizing the cross-entropy loss between the network and the teacher model prediction. Note that the teacher model can only provide supervision for the daytime domain since we can assume that M_D does not generalize to nighttime images as it is not trained on data from this domain. We formulate the daytime segmentation loss as:

$$\mathcal{L}_s^D = -\frac{1}{HW}\sum_{h,w} M_D(I_{\text{RGB}}^D)\log M_M(I_{\text{RGB}}^D, I_{\text{T}}^D), \quad (1)$$

where $M_D(I_{\text{RGB}}^D)$ denotes the teacher model prediction and $M_M(I_{\text{RGB}}^D, I_{\text{T}}^D)$ denotes the prediction of our multimodal RGB-T model for daytime thermal images I_{T}^D and RGB images I_{RGB}^D. H and W denote the height and width of the output, respectively. For simplicity, we omit the class index i in Eq. 1. By supervised training using the labels from M_D, the student model M_M does not generalize well to nighttime scenes because of the large domain shift in the RGB domain, in contrast to the thermal domain. In order to adapt the model to the nighttime domain in an unsupervised manner, we utilize a domain adaptation approach similar to [27] and insert a domain discriminator C after the softmax prediction layer of M_M. The domain discriminator has as inputs the softmax activations S_D or S_N of our segmentation model for daytime or nighttime inputs, respectively, and is trained to differentiate between both domains. We thus define the discriminator loss $\mathcal{L}_d$ as

$$\mathcal{L}_d = \frac{1}{HW}\sum_{h,w}\begin{cases}[0-C(S_X)]^2, & \text{if } X = D\\ [1-C(S_X)]^2, & \text{if } X = N\end{cases} \quad (2)$$

In order to adapt our model to the nighttime domain we aim to predict semantic segmentation maps that fool the discriminator model. In other words, we want to output predictions whose origin is classified as the daytime domain. If this confusion of the discriminator model can be achieved, it can be assumed that the distribution of the internal feature representations of our multimodal model are matched and the model is adapted to the nighttime domain. We train our model with an alternating training scheme for the two networks, where we step-wise alternate between adjusting the parameters of the discriminator model while freezing the segmentation model parameters and adjusting the parameters of the segmentation model while freezing the discriminator model parameters. In each iteration we sample an RGB-T image pair from the daytime and nighttime domain. In the first step of an iteration, we train our semantic segmentation network for the daytime domain while adapting the nighttime feature representations to daytime. We minimize an overall loss $\mathcal{L}_{p_1}$:

$$\mathcal{L}_{p_1} = \mathcal{L}_s^D + \lambda[0 - C(S_N)]^2, \quad (3)$$

where λ denotes a constant weighting factor between both losses, which we set to 0.01 during all experiments. In the second step, we exclusively train the discriminator to differentiate between day and night segmentation maps with the overall loss $\mathcal{L}_{p_2}$:

$$\mathcal{L}_{p_2} = \mathcal{L}_d \quad (4)$$

Our model architectures and the overall training scheme are illustrated in Fig. 2. In addition to the described approach, we propose the following extensions:

1) Two-Stage Training: We argue that the domain gap between day and night is much smaller for thermal images than for RGB images. This results in superior nighttime performance if a network is exclusively trained on thermal infrared images without any domain adaptation. As our goal

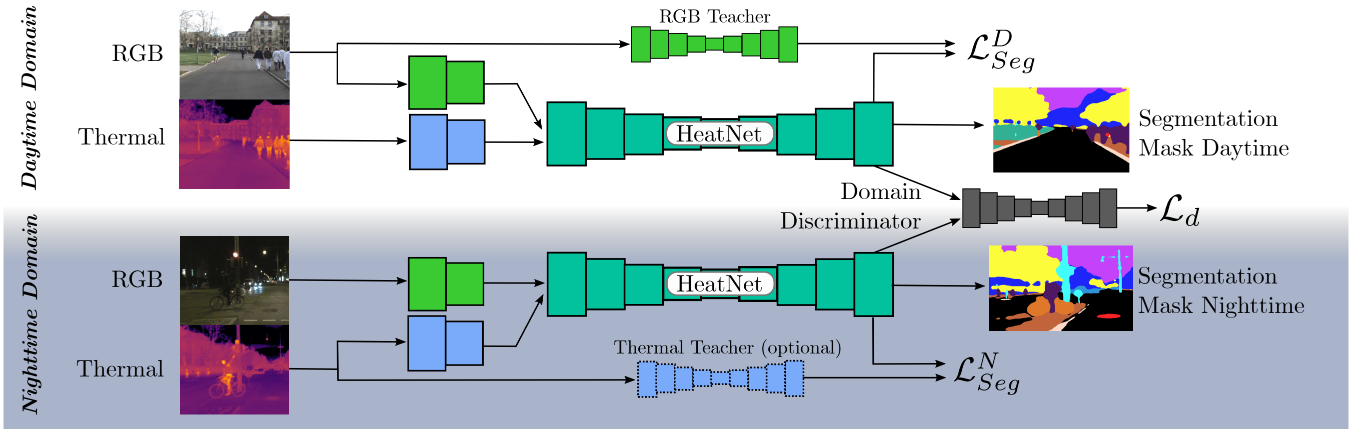

Fig. 2. Our proposed HeatNet architecture uses both RGB and thermal images and is trained to predict segmentation masks in daytime and nighttime domains. We train our model with daytime supervision from a pre-trained RGB teacher model and with optional nighttime supervision from a pre-trained thermal teacher model trained on exclusively thermal daytime images. We simultaneously minimize the cross entropy prediction loss to the teacher model prediction and minimize a domain confusion loss from a domain discriminator to reduce the domain gap between daytime and nighttime images.

is to train a multimodal network that performs best in both domains, day and night, we conduct domain adaptation to compensate for the illumination changes in the RGB images. We argue, however, that during domain adaptation the training could converge into local minima due to the large domain gap within the RGB modality and insufficient feature distribution overlap. We thus propose to first train our multimodal network M_M with the daytime teacher model M_D and an additional nighttime teacher model M_N. This additional teacher network is trained exclusively on daytime thermal infrared images and therefore predicts reasonable nighttime segmentation maps without domain adaptation due to the small domain gap. However, we argue that the semantic predictions provided by M_N can still be improved as M_N does not use complementary RGB data. Therefore, after the training with both teacher networks, we continue with the normal training procedure including domain adaptation, but without the nighttime teacher model, which we explained in the previous section. Following this training scheme, the feature representations align reasonably in the first training stage and the domain adaption in the second training stage does not need to bridge the full domain gap anymore.

2) RGB-T to RGB Model Distillation: Since thermal infrared cameras are not always installed on mobile robots, we propose a simple, yet effective strategy to enable RGB-only nighttime semantic segmentation using our approach. As previously mentioned, due to the domain gap in the visible spectrum, it is challenging to adapt an RGB-only model to the nighttime domain. Meanwhile, with our previous multimodal adaption approach we are capable of training a multimodal network that leverages RGB information jointly with thermal infrared information which exhibits a significantly smaller domain gap. We thus propose to first train a multimodal RGB-T network following the previously described method. We afterwards distill the knowledge of the RGB-T network to an RGB-only network. To this end we use the previously described RGB-only daytime teacher model to provide supervision in daytime and our best-performing RGB-T network to provide supervision in nighttime and train this RGB-only model fully supervised in both domains.

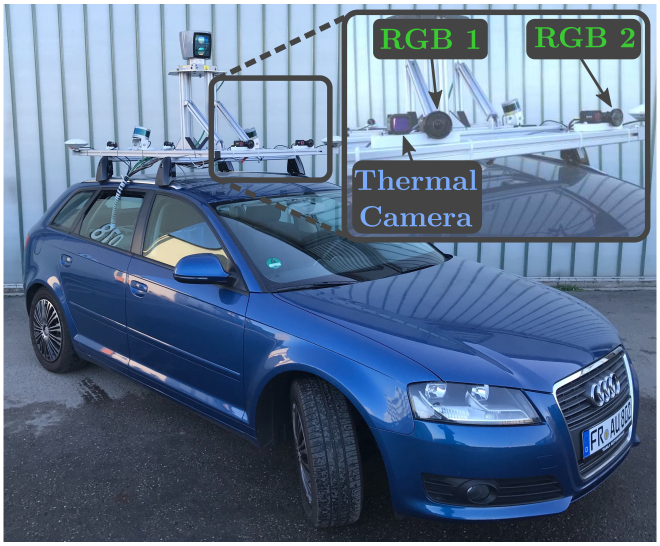

Fig. 3. Our stereo RGB and thermal camera rig mounted on our data collection vehicle.

IV. DATASET

To kindle research in the area of thermal image segmentation and to allow for credible quantitative evaluation, we create the large-scale dataset *Freiburg Thermal*. We provide the dataset and the code publicly available at *http://thermal.cs.uni-freiburg.de/*. The Freiburg Thermal dataset was collected during 5 daytime and 3 nighttime data collection runs, spanning the seasons summer through winter. Overall, the dataset contains 12051 daytime and 8596 nighttime time-synchronized images using a stereo RGB camera rig (FLIR Blackfly 23S3C) and a stereo thermal camera rig (FLIR ADK) mounted on the roof of our data collection vehicle. In addition to images, we recorded the GPS/IMU data and LiDAR point clouds. The Freiburg Thermal dataset contains highly diverse driving scenarios including highways, densely populated urban areas, residential areas, and rural districts. We also provide a testing set comprising 32 daytime and 32 nighttime annotated images. Each image has pixel-wise semantic labels for 13 different object classes. Annotations are provided for the following classes: *Road, Sidewalk, Building, Curb, Fence, Pole/Signs, Vegetation, Terrain, Sky, Person/Rider, Car/Truck/Bus/Train, Bicycle/Motorcycle,* and *Background*. We deliberately selected extremely challenging urban and rural scenes with many traffic participants and changing illumination conditions.

A. Camera calibration

For our segmentation approach it is important to perfectly align RGB and thermal images as otherwise the RGB teacher model predictions would not be valid as labels for the thermal modality. Thus, in order to accurately carry out the camera calibration for the thermal camera, we propose a novel target-less calibration procedure. While in previous works [23], [16] different kinds of checkerboards or circleboards have been leveraged, our method does not require any pattern. Although, for RGB cameras, these patterns can be produced and utilized easily, it still remains a challenge to create patterns that are robustly visible both in RGB and thermal images. In general, the used modalities infrared and RGB entail different information. However, we note that the edges of most common objects in urban scenes are easily observable in both modalities. Thus, in our approach we minimize the pixel-wise distance between such edges. In the case of aligning two monocular cameras, targetless calibration without any prior information results in ambiguities for the estimation of the intrinsic camera parameters. We therefore utilize our pre-calibrated RGB stereo rig in order to provide the missing sense of scale. Due to the target-less nature of our approach, our thermal camera calibration method can be easily deployed in an online calibration scenario.

Our aim is to overlay the RGB and thermal images as best as possible, solving both for the extrinsic and intrinsic parameters. If this alignment can be achieved, our cameras are assumed to be fully calibrated. In the following we assume the RGB image I_{RGB} to be undistorted and rectified. We formulate the misalignment E as the difference between the gradients of the calibrated RGB image and the transformed thermal image as:

$$E = \sum_{u,v} [\nabla I_{\mathrm{RGB}} - \nabla \mathcal{S}(I_T, F)] \tag{5}$$

Here, $\mathcal{S}(I_{\mathrm{T}}, F)$ denotes a function that transforms a source thermal image I_{T} to a target RGB image I_{RGB} while using a pixel displacement map F that maps from I_{T} to I_{RGB}. A successful calibration would result in the minimum value of E and would therefore align the thermal image with the RGB image. We follow [12] in order to implement $\mathcal{S}$, using differentiable spatial transformer networks.

We compute F by projecting the pixel coordinates of the RGB images to 3D, transforming them into the thermal camera coordinate system and projecting them back to the thermal image plane. Thus, the displacement map $F = p_{\mathrm{RGB}} - p_{\mathrm{T}}$ between the RGB pixel coordinates p_{RGB} and the thermal image pixel coordinates p_{T} can be found with:

$$p_{RGB} = \phi\big(K_{\mathrm{T}}\, T_{\mathrm{RGB}\to\mathrm{T}} \gamma(p_{\mathrm{RGB}} \mid K_{\mathrm{RGB}}, D_{\mathrm{RGB}})\big) \tag{6}$$

where the function $\gamma(p \mid K, D) = D(p)K^{-1}h(p)$ back-projects the RGB pixel coordinate into the 3D camera coordinate system while $h(p)$ transforms p in the homogeneous vector form. The intrinsic calibration of the RGB camera is denoted as K_{RGB} and D_{RGB} refers to the depth corresponding to the RGB image I_{RGB}. Further, $T_{\mathrm{RGB}\to\mathrm{T}}$ and K_{T} refer to the sought extrinsic and intrinsic thermal camera calibration values, respectively. The function $\phi(x)$ simply divides the vector x by its last element. We infer D_{RGB} by leveraging a dense stereo depth estimation method based on a convolutional neural network [2]. Due to the locality of the edges within the RGB and the thermal image, the direct minimization of the misalignment E would lead to vanishing gradients and would prevent fast convergence on the global minimum. In order to cope with this problem, we convolve the difference of gradients $\big(\nabla I_{\mathrm{RGB}} - \nabla \mathcal{S}(I_{\mathrm{T}}, F)\big)$ with a large Gaussian kernel $G(\sigma)$ which we empirically parameterize with zero mean, standard deviation $\sigma = 3$, and 51 pixel aperture size, resulting in our loss function:

$$\mathcal{L}_c = \sum_{u,v} \big[G(\sigma) * \big(\nabla I_{\mathrm{RGB}} - \nabla U(\mathcal{S}(I_{\mathrm{T}}, F), v)\big)\big]^2 \tag{7}$$

We follow [10] to model the distortion of the thermal image by the function U and optimize its parameters $v = [k_1,\, k_2,\, p_1,\, p_2]$, referring to radial and tangential distortion respectively, while optimizing the objective function.

We define the extrinsic calibration $T_{\mathrm{RGB}\to\mathrm{T}}$ as a rigid body transformation $T_{\mathrm{RGB}\to\mathrm{T}} = \begin{pmatrix} R & t \\ 0 & 1 \end{pmatrix} \in SE(3)$ where $R \in SO(3)$ and $t \in \mathbb{R}^3$. In order to ease the optimization we optimize the transformation in Lie-algebraic exponential coordinates $\xi = (v^T\ \omega^T) \in \mathfrak{se}(3)$ and use the exponential map with small-angle approximations [7] to map from $\mathfrak{se}(3)$ to $SE(3)$.

In our implementation we use Adam [13] for stochastic gradient descent to minimize Eq. 7 which yields the optimal extrinsic calibration $T^{*}_{\mathrm{RGB}\to\mathrm{T}}$, thermal camera intrinsic matrix K^{*}_{T}, and undistortion parameters v^{*}.

We take 600 random image-pairs for the optimization process and set a batch size of 10. Furthermore, we set the number of iterations to 8000 and halve the step size every 500 steps. We initialize K_{T} as:

$$K_{\mathrm{T}} = \begin{pmatrix} f_m/l & 0 & r_w/2 \\ 0 & f_m/l & r_h/2 \\ 0 & 0 & 1 \end{pmatrix}, \tag{8}$$

where f_m denotes the ideal manufactured focal-length of the lens, l the size of a single square pixel in mm, r_w the horizontal resolution and r_h the vertical resolution. All other parameters such as extrinsic calibration and distortion parameters are set to 10^{-4} to prevent vanishing gradients.

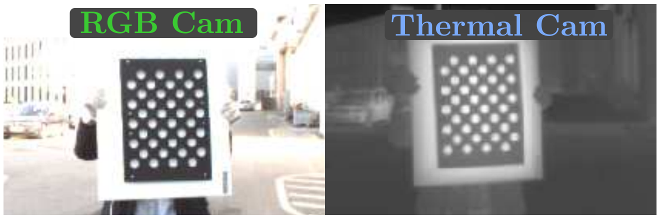

Fig. 4. Our calibration board placed in front of a heating panel is visible in the RGB and thermal domain. We record multiple image-pairs covering the whole camera frustum and obtain the calibration parameters with Kalibr [8]. This method is used as a baseline for our target-less calibration approach.

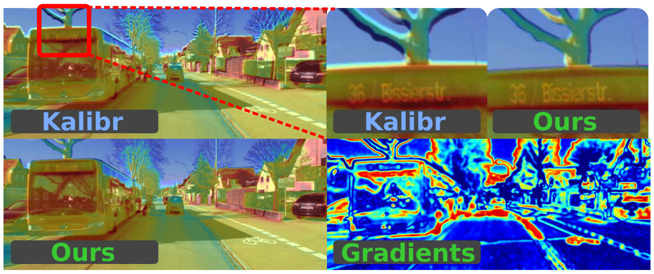

Fig. 5. Qualitative result of our target-less RGB-T calibration approach. In the left column, we show the RGB and thermal image alignment overlay with calibration parameters as obtained with Kalibr and our approach, respectively. The magnified view in the top-right corner demonstrates that our approach yields superior alignment of object edges. The bottom-right corner illustrates the magnitude of gradient difference between RGB and thermal image after the optimization process with our approach.

We qualitatively compare the RGB-thermal image alignment obtained with our target-less approach to a circleboard-based calibration procedure carried out using the publicly available tool Kalibr [8]. We manufactured a circleboard and placed it in front of a heating panel. Fig. 4 shows our calibration board as recorded by the RGB and thermal camera. The recorded image-pairs were used to obtain the extrinsic and intrinsic calibration parameters with Kalibr.

Fig. 5 qualitatively compares the RGB-thermal alignment obtained with our target-less approach to the alignment obtained with calibration parameters produced by Kalibr. Despite our approach not requiring any calibration targets, we observe that our approach yields qualitatively better alignment of RGB and thermal images.

V. Experimental Results

In the following we present the experimental results of our proposed multimodal semantic segmentation method. We evaluate our model on our proposed *Freiburg Thermal* dataset and on MF [9]. Furthermore, we present results on the 30 nighttime images of the Berkeley Deep Drive dataset [34], using the unimodal RGB version of HeatNet, leveraging our proposed knowledge distillation approach described in Sec. III-A.2. We also present various ablation studies and provide a discussion of all results.

A. Network Architecture

As our unimodal architecture for the teacher networks M_D and M_N we use the PSPNet architecture [36]. For our multimodal network we again adopt the PSPNet architecture but replicate the first two blocks of the corresponding ResNet-50 encoder. After passing the individual modalities through the replicated blocks we concatenate the feature maps and proceed with the remaining blocks of the encoder. For the discriminator architecture we follow the described architecture in [27].

B. Training Details

We train our HeatNet segmentation model for 100 epochs with the RMSprop optimizer and with an initial learning rate of 10^{-4}. We use learning rate halving every 30 epochs. In each training batch, using our alternating training scheme, we forward the RGB-T image pair and minimize Eq. 3. We set the batch size to 8 for all our experiments.

C. Baseline Comparison

We report the performance of HeatNet trained on Freiburg Thermal and tested on Freiburg Thermal, MF, and on the BDD night test split. All results are listed in Tab. I. We observe that our RGB Teacher model M_D, which is trained on the Vistas dataset [18], has a high mIoU score of 69.4 in the day domain and an expected low score of 35.7, as the network is neither trained nor adapted to the night domain. Our thermal teacher model M_N achieves a mIoU score of 57.0, which shows that the domain gap is much smaller for this domain as for RGB. Our final RGB-T HeatNet model achieves with 64.9 the overall best score on our test set. Furthermore the RGB-only HeatNet reaches a comparable score to our RGB-T variant, proving the efficiency of our distillation approach which leverages the thermal images as a bridge modality.

We deploy the same distilled RGB network to publish results on the night BDD split. It can be observed that our method boosts the mIoU by 50%.

In order to compare the performance of our network with the recent RGB-T semantic segmentation approaches MFNet [9] and RTFNet-50 [25], we also fine-tune our model on the 784-image MF [9] training set and report scores on the corresponding test set. We select all classes that are compatible between MF and Freiburg Thermal for evaluation which are the classes *Car*, *Person*, and *Bike*. We train our method only with labels provided by the teacher model M_D, while not requiring any nighttime labels or labels from MF in general. Thus, it is expected that MFNet and RTFNet outperform HeatNet as they are trained supervisedly. However, it can be observed that HeatNet achieves comparable numbers to MFNet.

We further evaluate the generalization properties of the models trained on MF and tested on our FR-T dataset. We observe that the model performance deteriorates when evaluating MFNet or RTFNet on our FR-T dataset. We conclude that the diversity and complexity of the MF dataset does not suffice to train robust and accurate models for daytime or nighttime semantic segmentation of urban scenes.

D. Ablation Studies

In order to evaluate the various components of our HeatNet approach, we perform ablation studies with different variants of our model. All ablation studies presented in this section

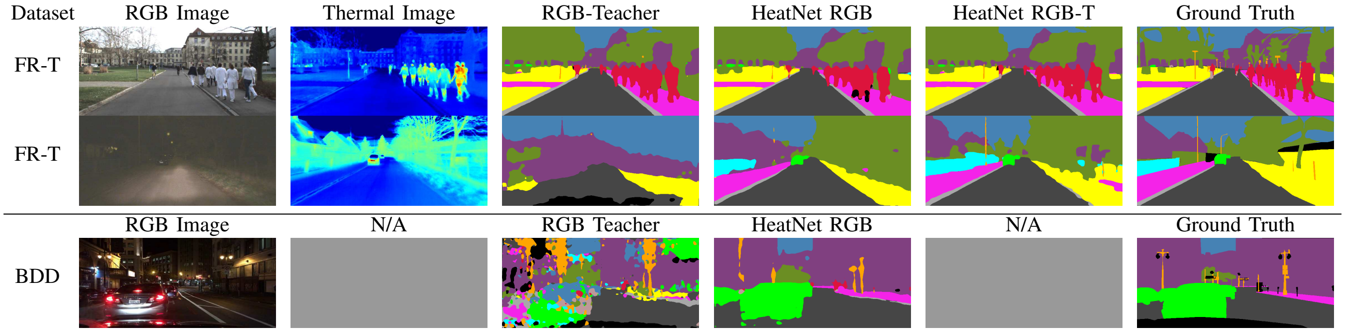

Fig. 6. Qualitative semantic segmentation results of our model variants. We compare segmentation masks of our RGB-only teacher model, HeatNet RGB-only, and HeatNet RGB-T to ground truth. In the first two rows, we show segmentation masks obtained on the Freiburg Thermal dataset. The bottom row illustrates results obtained on the RGB-only BDD dataset. The multimodal approaches cannot be evaluated on BDD and the corresponding images are left blank.

TABLE I

COMPARISON OF RGB-THERMAL SEMANTIC SEGMENTATION PERFORMANCE WITH STATE-OF-THE-ART APPROACHES ON THE MF DATASET AND ON THE FREIBURG THERMAL (FR-T) DATASET. WE MARK RESULTS OBTAINED USING FULLY SUPERVISED METHODS WITH A GRAY BACKGROUND. CLASSES AVAILABLE FOR EVALUATION DUE TO INCOMPATIBLE OR MISSING ANNOTATIONS ARE MARKED WITH A DASH (-).

Train On	Test On	Model	RGB	T	Road ⬛	Sidewalk 🟪	Building 🟪	Curb ⬜	Fence 🟦	Pole 🟧	Vegetation 🟩	Terrain 🟨	Sky 🟦	Person 🟥	Car 🟩	Bicycle 🟫	Mean
MF	MF	MFNet [9]	✓	✓	-	-	-	-	-	-	-	-	-	58.9	65.9	42.9	55.9
		RTFNet-50 [25]	✓	✓	-	-	-	-	-	-	-	-	-	**67.8**	**86.3**	**58.2**	**70.7**
		HeatNet	✓	✓	-	-	-	-	-	-	-	-	-	56.4	68.8	33.9	53.0
	FR-T Day/Night	MFNet [9]	✓	✓	-	-	-	-	-	-	-	-	-	42.8	27.0	24.5	31.4
		RTFNet-50 [25]	✓	✓	-	-	-	-	-	-	-	-	-	**63.2**	61.5	51.3	58.6
		HeatNet	✓	✓	86.7	57.5	67.7	46.4	41.5	43.8	57.9	44.1	63.7	63.1	**85.6**	**58.2**	59.7
FR-T	MF	HeatNet	✓	✓	-	-	-	-	-	-	-	-	-	51.6	61.8	30.2	47.9
(Vistas)	FR-T Day	RGB Teacher	✓	✗	**89.7**	**67.0**	73.8	56.9	48.8	53.8	73.8	62.8	84.3	72.0	90.1	60.4	69.4
FR-T		HeatNet	✓	✓	89.4	65.6	**74.8**	**59.7**	**52.9**	**54.3**	**74.1**	**65.1**	**84.5**	**74.0**	**91.2**	**64.1**	**70.8**
FR-T	FR-T Night	Thermal Teacher	✗	✓	84.9	60.5	**65.5**	43.1	31.8	38.1	51.8	40.1	72.6	49.6	**87.1**	**56.9**	57.0
(Vistas)		RGB Teacher	✓	✗	76.3	22.6	53.4	10.8	14.1	31.6	10.4	13.5	47.7	28.0	74.3	45.2	35.7
FR-T		HeatNet	✓	✓	**86.4**	**60.9**	65.4	**45.5**	**35.5**	**42.0**	**52.5**	**52.3**	**73.9**	**54.9**	85.7	53.3	**59.0**
FR-T	FR-T Day/Night	HeatNet	✓	✓	**87.9**	**63.3**	**70.1**	**52.6**	**44.2**	**48.2**	**63.3**	**58.9**	**79.2**	**64.5**	**88.5**	**58.7**	**64.9**
FR-T		HeatNet RGB-only	✓	✗	82.7	56.0	66.0	45.3	34.0	37.8	58.4	49.5	71.0	54.4	84.2	57.4	58.0
(Vistas)	BDD Night [34]	RGB Teacher	✓	✗	68.8	21.5	32.9	-	0.0	12.3	11.5	6.6	**27.2**	24.5	40.4	-	24.6
FR-T		HeatNet RGB-only	✓	✗	**87.1**	**40.0**	**50.2**	-	**25.9**	**22.9**	**12.8**	**8.5**	25.0	**27.4**	**68.3**	-	**36.8**

TABLE II

ABLATION STUDIES FOR VARIANTS OF OUR HEATNET MODEL ON THE FREIBURG THERMAL DATASET.

Variant	RGB	T	Domain Discrim-inator	Two-Stage Training	mIoU Day	mIoU Night	mIoU Both
V1	✗	✓	✗	✗	68.1	57.0	62.6
V2	✓	✗	✗	✗	68.3	25.1	46.7
V3	✓	✓	✗	✗	67.9	33.7	50.8
V4	✓	✗	✓	✗	70.5	43.2	56.9
V5	✓	✓	✓	✗	70.6	56.3	63.5
V6	✓	✓	✓	✓	**70.8**	**59.0**	**64.9**

were performed on our Freiburg Thermal dataset and are listed in Tab. II.

We first study the impact of the image modalities on the model performance without using domain adaptation or two-stage training. We compare a unimodal RGB-only model (V1) with a unimodal thermal-only model (V2) and the multimodal variant trained both on RGB and on thermal images (V3). All variants are trained exclusively on daytime annotations provided by the RGB daytime teacher model. We observe that the daytime-nighttime domain gap is the smallest for V1, while V2 and V3 suffer from a larger domain gap of the RGB modality, but achieve a higher daytime performance.

Variants V4 and V5, are similar to variants V2 and V3, but with an additional domain discriminator, indicating that adding a domain discriminator loss to the overall training as described in Sec. III-A greatly helps shrinking the domain gap within the RGB image modality. Variant V6, with active domain adaptation and two-stage training procedure as described in Sec. III-A.1 shows the best performance in both the daytime and the nighttime domain. We conclude that our proposed two-stage training scheme by first carrying out supervised training with two teachers and later fine-tuning with domain adaptation leads to the best results and helps aligning the feature representations between day and night as best as possible.

VI. CONCLUSION

In this work, we presented a novel and robust approach for daytime and nighttime semantic segmentation of urban scenes by leveraging both RGB and thermal images. We showed that our HeatNet approach avoids expensive and cumbersome annotation of nighttime images by learning from a pre-trained RGB-only teacher model and by adapting to the nighttime domain. We further proposed a novel training initialization scheme by first pre-training our model with a daytime RGB-only teacher model and a nighttime thermal-only teacher model and subsequently fine-tuning the model with a domain confusion loss. We furthermore introduced a first-of-its-kind large-scale RGB-T semantic segmentation dataset, including a novel target-less thermal camera calibration method based on image gradient alignment maximization. We presented comprehensive evaluations on multiple datasets and demonstrated the benefit of the complementary thermal modality for semantic segmentation and for learning more robust RGB-only nighttime models.

REFERENCES

[1] Amanda Berg, Jörgen Ahlberg, and Michael Felsberg. A thermal object tracking benchmark. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2015.

[2] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018.

[3] Yuhua Chen, Wen Li, and Luc Van 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.

[4] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948, 2018.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt 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.

[6] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3819–3824. IEEE, 2018.

[7] Ethan Eade. Lie groups for 2d and 3d transformations. http://www.ethaneade.com/lie.pdf, 2017.

[8] Paul Furgale, Joern Rehder, and Roland Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1280–1286. IEEE, 2013.

[9] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya 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)*, pages 5108–5115. IEEE, 2017.

[10] Janne Heikkila and Olli Silven. A four-step camera calibration procedure with implicit image correction. In *Proceedings of IEEE computer society conference on computer vision and pattern recognition*, pages 1106–1112. IEEE, 1997.

[11] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015.

[12] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: benchmark and baseline. *Pattern Recognition*, 96:106977, 2019.

[15] Chenglong Li, Wei Xia, Yan Yan, Bin Luo, and Jin Tang. Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation. *arXiv preprint arXiv:1907.10303*, 2019.

[16] Thomas Luhmann, Johannes Piechel, and Thorsten Roelfs. Geometric calibration of thermographic cameras. In *Thermal Infrared Remote Sensing*, pages 27–42. Springer, 2013.

[17] Jake T Lussier and Sebastian Thrun. Automatic calibration of rgbd and thermal cameras. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 451–458. IEEE, 2014.

[18] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4990–4999, 2017.

[19] Horia Porav, Tom Bruls, and Paul Newman. Don't worry about the weather: Unsupervised condition-dependent domain adaptation. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 33–40. IEEE, 2019.

[20] Yulong Qiao, Ziwei Wei, and Yan Zhao. Thermal infrared pedestrian image segmentation using level set method. *Sensors*, 17(8):1811, 2017.

[21] Eduardo Romera, Luis M Bergasa, Kailun Yang, Jose M Alvarez, and Rafael Barea. Bridging the day and night domain gap for semantic segmentation. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1312–1318. IEEE, 2019.

[22] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7374–7383, 2019.

[23] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. *arXiv preprint arXiv:1909.10980*, 2019.

[24] Lei Sun, Kaiwei Wang, Kailun Yang, and Kaite Xiang. See clearer at night: towards robust nighttime semantic segmentation through day-night image conversion. In *Artificial Intelligence and Machine Learning in Defense Applications*, volume 11169, page 111690A. International Society for Optics and Photonics, 2019.

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

[26] Wayne Treible, Philip Saponaro, Scott Sorensen, Abhishek Kolagunda, Michael O'Neal, Brian Phelan, Kelly Sherbondy, and Chandra Kambhamettu. Cats: A color and thermal stereo benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2961–2969, 2017.

[27] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan 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.

[28] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[29] Abhinav Valada, Gabriel L Oliveira, Thomas Brox, and Wolfram Burgard. Deep multispectral semantic scene understanding of forested environments using multimodal fusion. In *International Symposium on Experimental Robotics*, pages 465–477. Springer, 2016.

[30] Abhinav Valada, Johan Vertens, Ankit Dhall, and Wolfram Burgard. Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4644–4651. IEEE, 2017.

[31] Weihong Wang, Jian Zhang, and Chunhua Shen. Improved human detection and classification in thermal images. In *2010 IEEE International Conference on Image Processing*, pages 2313–2316. IEEE, 2010.

[32] Markus Wulfmeier, Alex Bewley, and Ingmar Posner. Addressing appearance change in outdoor robotics with adversarial domain adaptation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1551–1558. IEEE, 2017.

[33] Jihan Yang, Ruijia Xu, Ruiyu Li, Xiaojuan Qi, Xiaoyong Shen, Guanbin Li, and Liang Lin. An adversarial perturbation oriented domain adaptation approach for semantic segmentation. *arXiv preprint arXiv:1912.08954*, 2019.

[34] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018.

[35] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2020–2030, 2017.

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