Semantic Segmentation for Thermal Images: A Comparative Survey

Zülfiye Kütük Görkem Algan
Dept. of Image Proc. & Computer Vis. Technologies, Aselsan Inc., Turkey
{zulfiyekutuk, galgan}@aselsan.com.tr

Abstract

Semantic segmentation is a challenging task since it requires excessively more low-level spatial information of the image compared to other computer vision problems. The accuracy of pixel-level classification can be affected by many factors, such as imaging limitations and the ambiguity of object boundaries in an image. Conventional methods exploit three-channel RGB images captured in the visible spectrum with deep neural networks (DNN). Thermal images can significantly contribute during the segmentation since thermal imaging cameras are capable of capturing details despite the weather and illumination conditions. Using infrared spectrum in semantic segmentation has many real-world use cases, such as autonomous driving, medical imaging, agriculture, defense industry, etc. Due to this wide range of use cases, designing accurate semantic segmentation algorithms with the help of infrared spectrum is an important challenge. One approach is to use both visible and infrared spectrum images as inputs. These methods can accomplish higher accuracy due to enriched input information, with the cost of extra effort for the alignment and processing of multiple inputs. Another approach is to use only thermal images, enabling less hardware cost for smaller use cases. Even though there are multiple surveys on semantic segmentation methods, the literature lacks a comprehensive survey centered explicitly around semantic segmentation using infrared spectrum. This work aims to fill this gap by presenting algorithms in the literature and categorizing them by their input images.

1. Introduction

Semantic segmentation is one of the high-level tasks in computer vision that assigns a label for each pixel of an image. Semantic segmentation plays a significant role in many applications since it is able to provide scene understanding at the pixel level. Some of those applications include pedestrian segmentation, autonomous driving, and medical diagnosis. Semantic segmentation differs from other common computer vision tasks such as image classification and object detection in terms of its output. For instance, image classification provides which object exists in an image, while object detection gives the object labels and locations by a bounding box. Image segmentation is divided into three sub-branches: semantic, instance, and panoptic segmentation. Semantic segmentation provides a class label for each pixel of an image, while instance segmentation identifies and segments each instance of a class separately. Moreover, panoptic segmentation aims to find the class label for every pixel in an image and all the instances.

The interest in semantic segmentation has increased rapidly since the deep learning methods achieved promising results. In other words, deep learning-based semantic segmentation approaches have demonstrated a significant boost in efficiency compared to older methods. Different DNN architectures and mechanisms are proposed to obtain better segmentation results. For instance, fully convolutional networks (FCN) [19] have led to recent advances in deep learning-based semantic segmentation since many novel models use it to get dense predictions. Many semantic segmentation networks also employ the encoder-decoder structure. The encoders extract the features by reducing the resolution, and the decoders restore the resolution. SegNet [1] passes the indices of the max locations during pooling in the encoder to the decoder. Also, Unet [25] employs an encoder-decoder structure with skip connections that pass high-resolution features from the contracting path to the expanding path to guide semantic segmentation. DeepLabV3+ [6] benefits from spatial pyramid pooling and atrous convolution mechanisms. In addition, BiseNetV2 [37] captures high-level semantics and spatial details with two-pathway architecture. An aggregation layer is also used to exploit these extracted features for the semantic segmentation task.

Most of the methods in the literature exploit three-channel RGB images captured by visible cameras. However, due to the visible imaging limitations, these methods cannot provide the desired performance in adverse environmental conditions such as low-illuminated, rainy, foggy. Therefore, thermal images have been utilized for semantic segmentation tasks since thermal cameras capture ther-

mal radiation, which is more stable at any weather and time. However, thermal images usually have low resolution and ambiguous object boundaries caused by thermal crossover, a phenomenon where the thermal radiation coming from two different objects cannot be distinguished. Moreover, thermal crossover and the lack of the thermal dataset cause semantic segmentation in thermal images to be under-explored.

To the best of our knowledge, this work is the first survey of semantic segmentation methods in thermal images. The key contributions of this survey are as follows:

- A broad survey of current datasets, including RGB and thermal (RGB-T) image pairs and solely thermal images.
- A comprehensive review of the deep learning-based thermal image semantic segmentation methods with their architectures and contributions.
- A well-organized comparison of the methods with the announced quantitative measures in the papers.

This paper is organized as follows; Section II overviews the datasets, including thermal and RGB images, deep learning-based semantic segmentation methods for multi-spectral inputs, and a brief discussion on the presented methods. Section III introduces thermal image datasets, semantic segmentation methods using only thermal images, and a comparison of the methods with quantitative measures according to the revealed results in the papers. Finally, Section IV concludes this survey.

2. Combining Infrared and Visible Spectrum for Semantic Segmentation

The fact that infrared and visible spectrum information is in different light spectrums allows them to compensate for each other's deficiencies. While this provides an advantage, it restricts the use cases of the proposed algorithms only on specific hardware with two different sensors for thermal and visible light. Moreover, these methods require additional algorithms to fuse information coming from different spectrums. The fusion methods should avoid information conflicts while incorporating complementary information from different modalities. Besides, few datasets provide RGB-T aligned images with annotations. The number of proposed methods is limited due to the reasons mentioned above.

Utilizing RGB and thermal images simultaneously improves the model's performance. Therefore, this part mentions datasets with RGB-T images and segmentation methods using both visible and infrared spectrums.

2.1. Multi-spectral Datasets

Multi-Spectral Fusion Networks (MFNet) Dataset [10] contains both RGB and IR images captured using an InfRec R500 camera. This camera has different lenses and sensors for visible and infrared spectrum. The spatial resolutions of all images are 480x640. The dataset consists of 820 daytime and 749 nighttime urban scene images annotated with eight classes (car, person, bike, curve, car stop, guardrail, color cone, and bump). Moreover, the training set contains 50% of the daytime images and 50% of the nighttime images, while the remaining images are separated equally for the validation and test sets. Some prediction results of MFNet [10] and SegNet [1] can be seen in Figure 2 which is directly taken from [10]. Also, RGB-T image pairs and ground truth annotations from the dataset are presented in the first three rows of Figure 2.

Shivakumar et al. introduced Penn Subterranean Thermal 900 Dataset (PST900) [26] containing 894 aligned RGB-T image pairs with ground truth annotations. A Stereolabs ZED Mini stereo camera and a FLIRBoson 320 camera are used for data collection. The PST900 aims to meet the needs of the DARPA Subterranean Challenge[1] that requires the identification of four objects (fire extinguisher, backpack, hand drill, and thermal mannequin or person) and robustness in various underground situations. Therefore, images are gathered from diverse environments with varying degrees of lighting. Two RGB-T image pairs and the ground truth annotations from the dataset can be seen in Figure 1. Additionally, the dataset provides 3416 annotated RGB images.

Figure 1. RGB, thermal and annotation images from the PST900 dataset [26]

The Freiburg Thermal Dataset [30] includes 12051 daytime and 8596 nighttime time-synchronized RGB-T image pairs captured in rural and urban environments. A stereo RGB camera rig (FLIR Blackfly 23S3C) and a stereo thermal camera rig (FLIR ADK) are used for data collection. However, only the testing set including 32 daytime and 32 nighttime images annotated with the following classes: road, sidewalk, building, curb, fence, pole/signs, vegetation, terrain, sky, person/rider, car/truck/bus/train, bicycle/motorcycle and background.

[1] https://www.subtchallenge.com/

Multi-model multi-stage network (MMNet) [16] compares its nighttime segmentation performance with other models using its own dataset. The dataset contains 541 urban scenes RGB and thermal images taken only at night. All the images have a resolution of 300x400. Since the dataset is not publicly available, its use is limited to the MMNet.

2.2. Multi-spectral Semantic Segmentation Methods

Ha et al. [10] proposed Multi-Spectral Fusion Networks (MFNet) having two identical encoders for thermal and RGB images and one decoder block. Also, the encoder has a mini-inception block with dilated convolution so that the size of the receptive field is enlarged while the time complexity is the same with a normal 3x3 convolutional layer when the number of input and output channels are the same. MFNet aims to achieve high inference speed for real-time semantic segmentation for autonomous vehicles, and MFNet dataset, including RGB-Thermal (RGB-T) urban scene images, is introduced with pixel-level annotations for the self-driving task. MFNet includes a small decoder designed to reduce the number of parameters, and the decoder makes use of the low-level feature maps extracted in encoders to improve up-sampling efficiency. A concatenation operation fuses the outputs of the RGB and infrared (IR) encoders, and the decoder receives the fused result as input. Some segmentation predictions of MFNet and SegNet [1] can be seen in Figure 2 which is directly taken from [10].

Sun et al. proposed RGB-Thermal Fusion Network (RTFNet) [27] to achieve semantic segmentation of urban scenes for autonomous vehicles. RTFNet adopts an encoder-decoder structure with two encoders for extracting features of RGB and IR inputs and one decoder restoring the resolution of feature maps. The encoders are identical except the first layers' input channel numbers and slightly changed ResNet-50 [12] is employed as feature extractors. The infrared feature maps are fused into the RGB encoder through the element-wise summing. The decoder uses the output of the last fusion layer as input to obtain dense predictions. The encoder and decoder of the model are designed asymmetrically, two large encoders and a small decoder. Each decoder layer has two sub-blocks introduced by RTFNet, namely Upception A and Upception B. Upception A does not change the resolution and channel number, whereas Upception B changes, and the final channel number equals the number of classes. Also, Upception blocks have short-cut connections. In short, the decoder block gradually restores the resolution.

The Penn Subterranean Thermal Network (PSTNet) [26] includes independent RGB and Fusion streams to generate segmentation maps from RGB and thermal images. RGB stream can be trained without thermal data since collecting aligned RGB-T images is challenging. Therefore, the designed model uses thermal images to improve the initial segmentation result in the Fusion stream. The RGB stream is a ResNet-18 [12] architecture with an encoder-decoder and skip-connection scheme similar to U-Net [25]. The RGB stream is trained with the annotated RGB images to get the per-pixel confidence volume for the classes. This volume, thermal, and RGB input images are concatenated, and the result is passed to the Fusion stream, which is essentially an ERFNet-based [24] encoder-decoder architecture.

Sun et al. proposed FuseSeg [28] employing encoder-decoder structure and two-stage fusion strategy to achieve segmentation in urban scenes. There are two encoders taking three-channel RGB and one-channel thermal images as inputs, and DenseNet-161 [13] is employed as the backbone of the encoders. Moreover, FuseSeg introduces a decoder including three modules: a feature extractor with two convolutional layers, an upsampler, and an out block. The upsampler and the out block each have a transposed convolutional layer. The feature extractor is responsible for extracting features from the fused feature maps while keeping the resolution of the feature maps unchanged. The upsampler and the out block increase the resolution by 2. The out block outputs the final segmentation result. Sun et al. also proposed a two-stage fusion strategy to effectively use the multi-spectral inputs and reduce the loss of spatial information due to downsampling. In the first stage of the fusion, feature maps extracted from the inputs in the encoder are fused with element-wise summation in the RGB encoder. The summations are again fused with the corresponding feature maps in the decoder through concatenation.

Vertens et al. [30] proposed HeatNet intending to achieve daytime and nighttime image segmentation tasks without costly annotations of nighttime images. The PSPNet [39] is exploited as a teacher model to get the annotations of daytime images in the Freiburg Thermal dataset [30]. For this purpose, PSPNet is trained on the Mapillary Vistas dataset [22]. Then, a multimodal semantic segmentation network is trained using RGB and thermal daytime images annotated by the teacher model. The multimodel network also exploits PSPNet architecture and the first two block of the corresponding ResNet-50 [12] encoder. Moreover, a domain adaptation method similar to [29] is proposed to obtain nighttime segmentation results; therefore, a domain discriminator is employed after the softmax layer of the multimodal RGB-T model. Besides, the training is conducted using an alternating training scheme.

Graded-Feature Multilabel-Learning Network (GMNet) [40] includes two encoders for feature extraction and three grading decoding stages to restore original resolution. The proposed model employed ResNet-50 [12] as the backbone of the encoders. The fully connected and average pooling

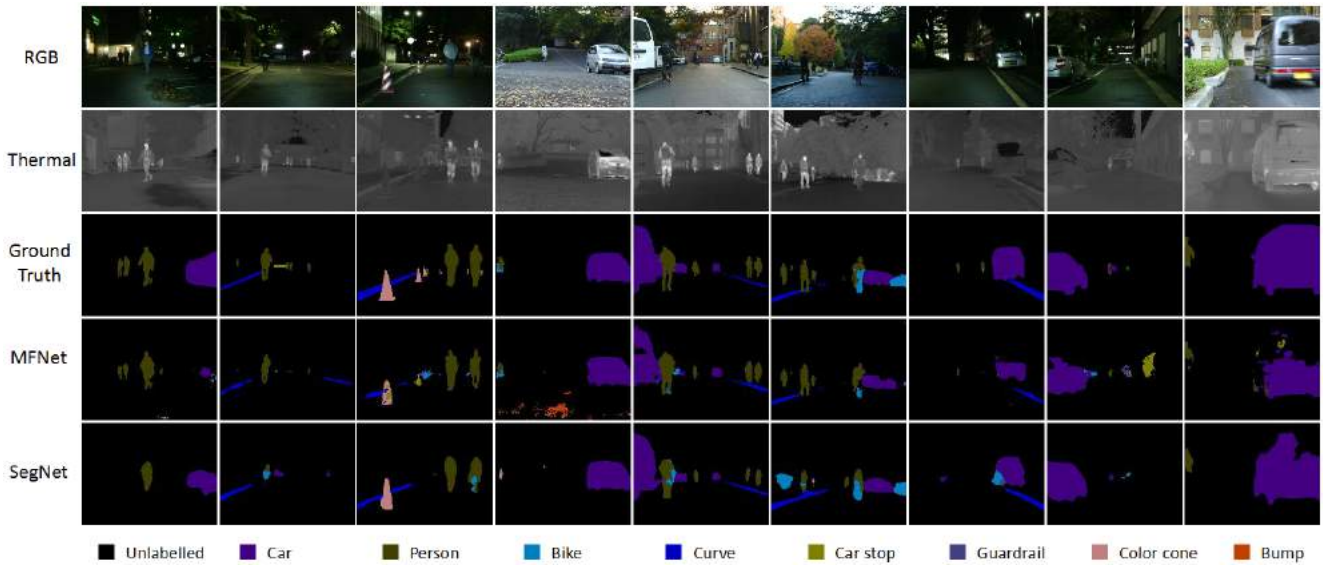

Figure 2. Some prediction results of MFNet [10] and Segnet [1] on the MFNet dataset [10]

layers of ResNet are removed as they may result in the loss of spatial information and details. GMNet divides multi-level features into senior, intermediate, and junior grades. The features extracted from the ResNet's last three layers, in which the visual receptive fields are enlarged, are selected as senior features. Besides, the features from the first layer, which have more detailed information, are selected as junior features. Moreover, GMNet introduces two fusion modules, the shallow feature fusion module (SFFM) and the deep feature fusion module (DFFM), to use the junior, intermediate, and senior features. SFFM fuses the features from the first two layers of the encoders, whereas DFFM accomplishes the fusion operation for the last three layers. Finally, semantic, binary, and boundary loss functions are used to find the optimum parameters of the model.

Multi-Modal Multi-Stage Network (MMNet) [16] tackles the semantic segmentation problem by employing three encoder-decoder structures. In the first stage of the network, two separate encoder-decoder structures process RGB and thermal images with no information interactions between the modalities. In the second stage, one encoder-decoder fuses and refines the features from the first stage. The proposed model deploys ResNet-18 [12] as encoders in the first stage, whereas Mini Refinement Block (MRB) is proposed as the encoder for the second stage. Each encoder sends information to the corresponding decoder using skip connections. As the direct connection may impact the fusion performance, EFEM (Efficient feature enhancement module) has been proposed to reduce the semantic gap between encoders and decoders.

Zhang et al. [38] proposed Adaptive-weighted Bi-directional Modality Difference Reduction Network (ABMDRNet) containing three parts: Modality Difference Reduction and Fusion (MDRF) subnetwork, Multi-Scale Spatial Context (MSC) module, and Multi-Scale Channel Context(MCC) module. All RGB-T networks strive to use complementary information from RGB and thermal images to their advantage. The integration and utilization of multi-modality complementary information from RGB and thermal images may be hampered by the modality difference generated by distinct imaging mechanisms. Therefore, the MDRF subnetwork uses a bridging-then-fusing strategy to reduce the modality difference and utilize the multi-modality complementary information. An MSC module and an MCC module are designed to exploit multi-scale contextual knowledge of cross-modality features and their long-range relationships along spatial and channel dimensions.

Xu et al. [36] proposed Attention Fusion Network (AFNet) containing an attention fusion module to guide the fusion operation of multi-spectral inputs. Also, AFNet employs two identical encoders for feature extraction and a single decoder for resolution restoration. The encoders are designed based on the ResNet-50 [12] with dilated convolutions. Also, the downsampling operations in the last two residual blocks in ResNet are removed. To make full use of the complementary properties of the RGB and thermal images, AFNet designed an attention fusion module. The attention matrices are obtained considering the cross-spectral and the global contextual relations of the images. The fusion operation takes place under the guidance of attention matrices, and the decoder uses the fused feature map. Moreover, the decoder employs three interpolations and three convolutional layers to obtain the segmentation result.

2.3. Analysis & Results

Exploiting thermal images as well as RGB images may improve the segmentation network in terms of accuracy and robustness. For multi-spectral input semantic segmentation, several methods have been developed, and this review describes the similarities and differences of these methods in many aspects. The proposed architectures and fusion strategies tackle the difficulties of employing two images and producing a precise segmentation result. Two encoders and a single decoder are commonly used in RGB-T methods, such as [10], [27], [28], [40], and [36] to extract features and restore the resolution. Moreover, [26] employs two distinct streams to process RGB images and the fused features, respectively. In this way, it can be trained by using only RGB images, and thermal images may provide further improvement. Also, [16] has three encoder-decoder structures with EFEM connections. [30] achieves nighttime image segmentation as well as daytime image segmentation by using a domain adaptation method. Furthermore, different fusion strategies are proposed to use complementary information from different modalities without information conflicts. [10] concatenates the outputs of the encoders, whereas the [27] fuses the thermal features into the RGB encoder through element-wise summation. Besides, more complex fusion strategies are proposed for the fusion operation, such as two-stage fusion, bridging-then-fusing strategies, attention fusion module, SFFM, and DFFM.

The MFNet dataset [10] includes both day and night RGB images with aligned thermal images. Since the images in the dataset can provide complementary information, RGB-T correlations are essential. The quantitative results of the RGB-T methods on test images from [10] can be found in Table 1. The results are shown using a standard evaluation metric, mean Intersection over Union (mIoU). All the results provided in the table are taken from the original papers of the mentioned methods. On the other hand, the PST900 dataset [26] is more challenging for thermal fusion networks since plenty of information is provided by RGB alone, and the same object images are captured at both above and below the ambient temperature, making learning RGB-T correlations challenging. The PSTNet [26] reports better results on PST900 because the model has a separate RGB stream and employs a late fusion approach.

According to the revealed results in [30], on the MFNet dataset with three classes (person, car, bicycle), [30] and [10] has comparable results, while [27] outperforms with 0.707 mIoU. Also, the paper indicates that [30] achieves mIoU score of 0.597 on the Freiburg Thermal dataset while [10] and [27] perform with only 0.314 mIoU and 0.586 mIoU, respectively.

The inference speed of [10], RTFNet-50, RTFNet-152 [27] and [28] are declared as 229.86, 88.87, 34.07 and 30.01 FPS according to the results announced in [28]. Also, [16] and [36] reports their inference speeds slightly higher than [27].

Table 1: The Results of the RGB-T Methods on MFNet Dataset [10]

Method	**mean IoU**
MFNet [10]	0.649
RTFNet-50 [27]	0.517
RTFNet-152 [27]	0.532
PSTNet [26]	0.484
FuseSeg [28]	0.545
GMNet [40]	0.573
MMNet [16]	0.528
ABMDRNet [38]	0.548
AFNet [36]	0.546

3. Semantic Segmentation Using Only Infrared Spectrum

In terms of capturing details under adverse environmental conditions, thermal imaging cameras outperform visual imaging cameras. Thermal imaging cameras are widely used in the defense industry, and since they have become more affordable, they have gained popularity in various other applications. Therefore, a couple of approaches have been developed to achieve high accuracy under a wide range of conditions by using only infrared images. Unlike methods using RGB-T image pairs, aligned images are not required, making data collection easier. But still, the lack of a thermal image dataset limits the number of works conducted in the area. Also, extracting features might be challenging due to thermal crossover, low resolution, and contrast of the infrared images.

3.1. Thermal Datasets

Li et al. introduced Segment Objects in Day and Night (SODA) dataset [17]. The SODA consists of 2168 real and 5000 synthetically generated thermal images. The real subset is captured by a FLIR camera (SC260). The thermal images generated from annotated RGB images are included in the synthetic subset. An image-to-image translation method, pix2pixHD [32], is trained with KAIST Multispectral Pedestrian Dataset [14]. After training the model, the synthetic subset is generated from Cityscapes [8]. Figure 3 shows some synthetically generated thermal images and ground truth annotations. Labels of the generated thermal images can be obtained directly from RGB annota-

tions. Besides, the real subset images are manually annotated. Three thermal images and annotations from the real subset can be seen in Figure 4.

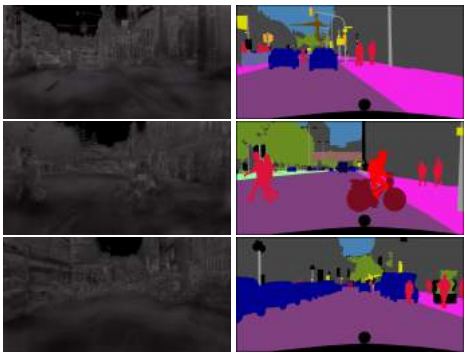

Figure 3. Synthetically generated thermal images and ground truth annotations in the SODA dataset [17]

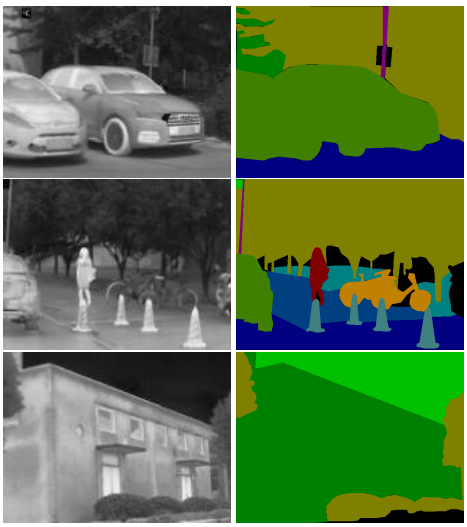

Figure 4. Thermal images and ground truth annotations from the real subset of the SODA dataset [17]

For pedestrian detection from thermal images, there are a few well-known datasets such as OSU Thermal Pedestrian Database (OSUT) [9], Terravic Motion IR Database (TMID) [2], and Pedestrian Infrared/Visible Stereo Video Dataset (PISVD) [3]. However, these datasets are not suited for the segmentation task due to the lack of annotations. Wang et al. [31] introduced a new dataset including thermal pedestrian images from the driver's perspective for autonomous driving applications. The dataset consists of 1031 thermal images at a resolution of 720x480 sampled from 25 scene videos. The dataset is also split into two equal parts for train and test sets. However, the dataset is not publicly available.

[2]http://vcipl-okstate.org/pbvs/bench/

Another application of the thermal semantic segmentation might be the ground vehicle segmentation from aerial images. In this context, NPU_CS_UAV_IR_DATA [18] dataset includes UAV-based infrared vehicle images. The dataset also provides four groups of road images for testing. Flying altitude, resolution of the images, and ambient temperature differ in these groups. Also, the captured images differ in terms of the number of vehicles and surroundings.

For the networks aiming for good segmentation results despite illumination and noise, the Low Illumination Image dataset (LII) [7] includes manually labeled thermal, motion blur, night, and weak lighting images. The images' average SNR (signal-to-noise ratio) is 25.5 dB.

Xiong et al. introduced SCUT-Seg dataset [35] which includes nighttime driving scenes from different environments. The dataset includes 2010 thermal images with semantic-wise annotations for ten classes (background, road, person, rider, car, truck, fence, tree, bus, and pole). Also, instance-wise annotations are provided for future works. The training and testing sets consist of 1365 and 665 images, respectively. Four example images and their ground truth annotations from the training set are presented in Figure 5.

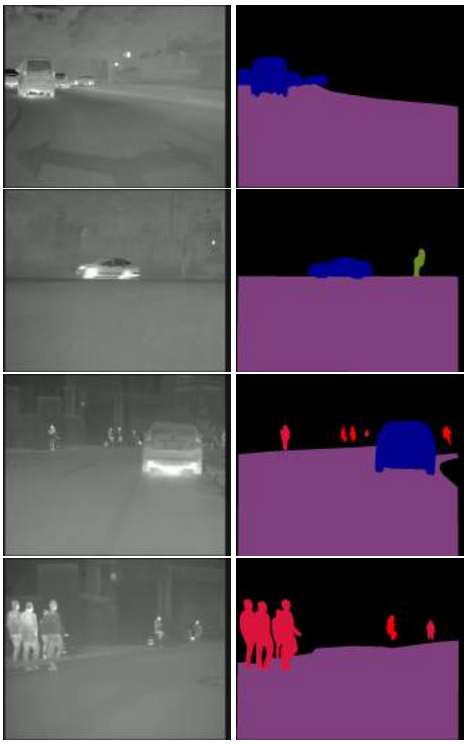

Figure 5. Thermal images and corresponding ground truth images from the SCUT-Seg dataset [35]

3.2. Thermal Semantic Segmentation Methods

Edge-Conditioned CNN (EC-CNN) [17] exploits edge prior information to increase the quality of segmentation output since thermal crossover and thermal sensors cause ambiguous object boundaries and imaging noise, respectively. Some gated feature-wise transform (GFT) layers are inserted into the model to embed edge information properly. The proposed model consists of an edge extractor (EdgeNet), EC-CNN blocks, and a DeepLabV3 [5] based semantic segmentation network. As an edge extractor, HED (Holistically-nested Edge Detection) [34] is employed to obtain high-quality edge information. However, there is no thermal dataset with edge annotations; the RGB dataset was used for HED training. Even though HED is trained on an RGB dataset with ground truth edge annotations, the edge results of thermal images are quite successful. EC-CNN blocks consist of convolutional layers and GFT layers to guide the segmentation of the input image by using the output of the EdgeNet. Also, the DeepLabV3 model employs ResNet as feature extractor and atrous convolutions, whereas some ResNet blocks are replaced with EC-CNN block to embed edge prior.

Wang et al. [31] proposed a thermal infrared pedestrian segmentation algorithm including a conditional generative adversarial network (IPS-cGAN). The generator of the IPS-cGAN is based on Unet [25] with two modifications so that a more suitable network for thermal infrared pedestrian segmentation is obtained. Firstly, to have more efficient connections, original convolutional blocks are replaced by residual blocks. Secondly, 0.5 rate dropout has been deployed so that the network becomes more robust. Moreover, SandwichNet is designed with a symmetrical structure as the discriminator of the proposed network. SandwichNet takes original image and segmentation results as inputs. The SandwichNet is designed based on multi-channel input PatchGAN [15]. The difference is that SandwichNet needs symmetrical three-channel result-image-result with segmentation result from the generator and thermal image, and three-channel truth-image-truth with segmentation ground truth. Moreover, the designed generator and the discriminator are trained as an end-to-end GAN algorithm with cross-entropy loss. The modifications on Unet and the design of the discriminator provide a more robust model against noises for thermal infrared pedestrian segmentation.

The combination of the Gaussian-Bernoulli Restricted Boltzmann Machine (GB-RBM) and convolutional neural network is proposed in RT-SegRBM-Net [20] to segment the vehicles from the UAV-based thermal images in real-time. The deep learning algorithm is designed based on SegNet [1] architecture, and GB-RBM is embedded into the overall structure to make use of the geometry information of the vehicles.

Nightvision-Net (NvNet) [7] is proposed for semantic segmentation of low-resolution infrared images in weak illumination environmental conditions. Nv-Net suggests the network architecture of the FCN-8S [2] with a contracting and an expanding path and a weighting loss. Also, transfer learning is utilized to increase the performance of semantic segmentation. NvNet architecture consists of four parts, data refinement (DR), data normalization (DN), the contracting path, and the expanding path. The contracting path has several convolution layers and average pooling operations and outputs down-sampled feature maps. Therefore, the aim of the expanding path is to enhance the output feature map's resolution. The contracting path is complemented by the expanding path that applies consecutive layers with upsampling operations instead of pooling operations. Also, the expanding path uses the feature maps from the corresponding layers of the contracting path to achieve better localization of the objects. Moreover, data normalization is performed, which accelerates the convergence of the training. Besides, NvNet introduces weighted-sigmoid-cross-entropy loss to calculate the error between the prediction and ground truth.

Xiong et al. proposed a Multi-level correction network (MCNet) [35] to achieve thermal images segmentation for nighttime driving scenes. Thermal images have low resolution and blurred edges caused by the thermal crossover; therefore, MCNet proposes the multi-level attention module (MAM) to solve this problem. The MAM includes two sub-modules, the context aggregation module (CAM) and the correlation matrix correction module (CMCM). CAM is chosen to model the spatial correlations within pixels' position, and the correlation matrix learns the dependency between any two pixels. The correlation matrix has significant importance since the properties of the thermal images, such as low resolution and ambiguous object boundaries, may cause misleading results about the related contextual information. So, to prevent this misleading information and suppress the noisy information, the CMCM module is also included in the proposed method. If the correlation values between the intra-class pixels are lower than that of inter-class pixels, the CMCM module corrects these wrong values. Furthermore, thermal images are more dependent on edge information due to the lack of color information. Hence, a multi-level edge enhancement module (MEEM) is designed to enhance the edge information and improve the final feature representation in multiple iterations.

Feature Transverse Network (FTNet) [23] is an end-to-end trainable convolutional neural network architecture. FTNet employs an encoder-decoder structure and an edge guidance part to conduct reliable pixel-wise classification. FTNet introduces a feature transverse network (decoder) exploiting a set of residual units [12]. Moreover, ResNeXt [33] based encoder network provides thermal image features at different resolutions by subsampling at several

stages. These feature maps are passed through the aforementioned residual units. Also, FTNet employs a fully connected layer to combine the outputs of the residual units. Edge information is also exploited to reduce the effects of thermal crossover and noise created by the sensors on the segmentation map. Moreover, the edges are extracted from the feature map obtained in the third layer of the encoder and passed through the combination of layers convolution, batch normalization, ReLU, respectively. Then, the edge map is upsampled to the input image resolution before passing through another convolutional layer. Finally, the edge map is also fused with the feature maps obtained in the decoder, and the result is passed through the final block, including convolutional, batch normalization, and ReLU layers. In addition, an edge-based loss function is adapted with the semantic loss while training FTNet to increase the segmentation accuracy. Edge ground truths are calculated from the semantic label gradients.

3.3. Analysis & Results

Thermal images alone can provide sufficient information in adverse environmental conditions, so a few segmentation methods have been developed using only thermal images. Although thermal crossover and noise introduced by the thermal imaging sensors make the segmentation task more difficult, the thermal segmentation methods propose different approaches such as employing edge information and correlation matrix. [17], [35] and [23] propose mechanisms to extract the edge information from the thermal image and use it to guide segmentation. [20], [7], and [31] employ encoder-decoder structures, whereas [17] exploits atrous convolutions to obtain the output segmentation map. Moreover, [35] creates the correlation matrix, which models the dependency between any two pixels, to focus on the same classes and avoid noisy information. Also, [7] exploits weighted-sigmoid-cross-entropy loss for images collected in weak illumination environmental conditions to discriminate important pixels while calculating the loss. [20] attempts to segment vehicles from UAV-based thermal images, so a Boltzmann machine is employed for geometry extraction from vehicles up view to increase the segmentation accuracy. [35] is proposed for nighttime driving scenes, so the model exploits inherent aspects of driving scene images, such as the fact that object instances show only in narrow bands that cross horizontally through the image's center.

SODA dataset [17] includes day and night thermal images for generic purposes and commonly used for testing thermal segmentation methods. On the SODA testing set, [17] reported the performance of the proposed method and [5] as 0.619 and 0.571 mIoU, respectively. Moreover, [35] and [23] reaches 0.503 and 0.600 mIoU on the same dataset, as reported in [23]. It can be noted that [17] and [23] have comparable results on SODA. In addition, [31] is designed to overcome regional intensity inhomogeneity and be more robust against various noises for the infrared pedestrian segmentation. According to the revealed results in [31], the proposed method performs with 0.939 mIoU on its own dataset, and outperforms [5], [21] and [15]. In terms of the average precision results, [20], [11], [5], and [1] achieves the similar performance on the NPU_CS_UAV_IR_DATA [18] test sets, whereas [20] achieves slightly better average processing time, as reported in the [20]. In addition, [7] reported that the proposed model performs with 0.912 mIoU, which is better than [4] with 0.469 mIoU on the LII dataset. On the SCUT-Seg nighttime driving dataset, [35] reported 0.676 mIoU and 32.52 FPS with a single NVIDIA GTX 1080 Ti. Moreover, [23] announces its accuracy as 0.667 mIoU on the same dataset. Similar to SCUT-Seg, MFNet dataset [10] also contains driving scenes, and [35] reaches 0.519 mIoU on the thermal images in MFNet dataset, while [6] only achieves 0.504 mIoU and RTFNet-50 [27] using both RGB and thermal images achieves 0.503 mIoU according to the revealed results in [35]. Using the thermal images in the [10] dataset, [23] reported its accuracy as 0.471 mIoU which is also comparable with the results of [35] and [27].

4. Conclusion

This survey reviews recent progress in deep learning-based semantic segmentation methods using thermal images and compares them in terms of their architectures, performance, applications, and the proposed approaches to improve the models. Also, this survey provides thermal image datasets descriptions.

In conclusion, using thermal images in semantic segmentation tasks helps to increase the robustness and success of the systems. Also, the proposed methods can be used in a wide range of applications. Due to the limited number of available thermal image datasets and characteristics of the images, only a few methods have been developed. The semantic segmentation of thermal images is very promising, and further research can be advanced in several directions, such as creating synthetic data, data augmentation, and fusion strategies.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 1, 2, 3, 4, 7, 8

[2] Ashish Kumar Bhandari and Kusuma Rahul. A context sensitive masi entropy for multilevel image segmentation using moth swarm algorithm. *Infrared Physics & Technology*, 98:132–154, 2019. 7

[3] Guillaume-Alexandre Bilodeau, Atousa Torabi, Pierre-Luc St-Charles, and Dorra Riahi. Thermal–visible registration of

human silhouettes: A similarity measure performance evaluation. *Infrared Physics & Technology*, 64:79–86, 2014. 6

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 8

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 7, 8

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1, 8

[7] Shaohui Chen, Zengzhao Chen, Xiaogang Xu, Ningyu Yang, and Xiuling He. Nv-net: Efficient infrared image segmentation with convolutional neural networks in the low illumination environment. *Infrared Physics & Technology*, 105:103184, 2020. 6, 7, 8

[8] 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. 5

[9] James W Davis and Mark A Keck. A two-stage template approach to person detection in thermal imagery. In *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)-Volume 1*, volume 1, pages 364–369. IEEE, 2005. 6

[10] 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. 2, 3, 4, 5, 8

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 8

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 4, 7

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 3

[14] 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. 5

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 7, 8

[16] Xin Lan, Xiaojing Gu, and Xingsheng Gu. Mmnet: Multimodal multi-stage network for rgb-t image semantic segmentation. *Applied Intelligence*, pages 1–13, 2021. 3, 4, 5

[17] 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. *IEEE Transactions on Neural Networks and Learning Systems*, 2020. 5, 6, 7, 8

[18] Xiaofei Liu, Tao Yang, and Jing Li. Real-time ground vehicle detection in aerial infrared imagery based on convolutional neural network. *Electronics*, 7(6):78, 2018. 6, 8

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[20] Mehdi Khoshboresh Masouleh and Reza Shah-Hosseini. Development and evaluation of a deep learning model for real-time ground vehicle semantic segmentation from uav-based thermal infrared imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 155:172–186, 2019. 7, 8

[21] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 8

[22] 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. 3

[23] Karen Panetta, KM Shreyas Kamath, Srijith Rajeev, and Sos S Agaian. Ftnet: Feature transverse network for thermal image semantic segmentation. *IEEE Access*, 9:145212–145227, 2021. 7, 8

[24] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, 2017. 3

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 3, 7

[26] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9441–9447. IEEE, 2020. 2, 3, 5

[27] 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. 3, 5, 8

[28] Yuxiang Sun, Weixun Zuo, Peng Yun, Hengli Wang, and Ming Liu. Fuseseg: semantic segmentation of urban scenes based on rgb and thermal data fusion. *IEEE Transactions on Automation Science and Engineering*, 2020. 3, 5

[29] 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. 3

[30] Johan Vertens, Jannik Zürn, and Wolfram Burgard. Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8461–8468. IEEE, 2020. 2, 3, 5

[31] Peng Wang and Xiangzhi Bai. Thermal infrared pedestrian segmentation based on conditional gan. *IEEE transactions on image processing*, 28(12):6007–6021, 2019. 6, 7, 8

[32] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 5

[33] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 7

[34] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 7

[35] Haitao Xiong, Wenjie Cai, and Qiong Liu. Mcnet: Multi-level correction network for thermal image semantic segmentation of nighttime driving scene. *Infrared Physics & Technology*, 113:103628, 2021. 6, 7, 8

[36] Jiangtao Xu, Kaige Lu, and Han Wang. Attention fusion network for multi-spectral semantic segmentation. *Pattern Recognition Letters*, 146:179–184, 2021. 4, 5

[37] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129(11):3051–3068, 2021. 1

[38] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2633–2642, 2021. 4, 5

[39] 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. 3

[40] Wujie Zhou, Jinfu Liu, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Gmnet: graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation. *IEEE Transactions on Image Processing*, 30:7790–7802, 2021. 3, 5