DooDLeNet: Double DeepLab Enhanced Feature Fusion for Thermal-color Semantic Segmentation

Oriel Frigo   Lucien Martin-Gaffe   Catherine Wacongne

AnotherBrain
Paris, France
{oriel, lucien, catherine}@anotherbrain.ai

Abstract

In this paper we present a new approach for feature fusion between RGB and LWIR Thermal images for the task of semantic segmentation for driving perception. We propose DooDLeNet, a double DeepLab architecture with specialized encoder-decoders for thermal and color modalities and a shared decoder for final segmentation. We combine two strategies for feature fusion: confidence weighting and correlation weighting. We report state-of-the-art mean IoU results on the MF dataset [1].

1. Introduction

Thermal images can improve the performance of computer vision algorithms, particularly during night time or in harsh weather conditions, when visibility becomes limited. On the other hand, color images provide a richer source of visual information during day time. Hence it can be highly beneficial for vision models to combine both color and thermal images into a RGBT (red, green, blue, thermal) image pipeline.

Multimodality or the combination of different imaging sources becomes crucial to alleviate perception and ensure robustness under poor visibility conditions, but the combination of different cameras and sensors comes with alignment and synchronization challenges. Calibration of thermal images is a painstaking task where special heated or masked calibration patterns need to be employed. Alignment of color and thermal images is also a challenging problem, not only because of common differences in field of view and resolution between the color and thermal cameras but also because of the huge differences in image appearance between the two imaging modalities.

In this paper we consider the task of semantic segmentation of paired color RGB and thermal LWIR (Long Wave Infra-Red) images. Most previous research in this field has been focused in searching for good fusion strategies between the color and thermal images. However, previous work on semantic color-thermal segmentation has barely considered how to fuse the images taking into account the feature disagreement due to poor alignment.

Inspired by recent advances in semantic matching, here we consider an end-to-end deep learning strategy which indirectly takes into account the alignment of RGB and LWIR thermal images through a matching correlation layer.

We show the benefits of the correlation and confidence aware feature fusion for the task of semantic segmentation on RGB-thermal data.

The main contributions of this paper are the following:

- We propose a new CNN model for color-thermal semantic segmentation. We show the interest of learning deep features that are specialized for both thermal images and color images with specialized encoder-decoder paths.

- We show that semantic segmentation performance on color-thermal image pairs is increased by the aid of two different weighted feature fusion strategies:

    - For matching correlation weighting, we explicitly estimate the feature matching at the last specialized decoder layer by computing feature correlations.
    - For confidence weighting, we estimate the segmentation confidence at both color and thermal features, and this confidence is used to weight differently the color and thermal features.

2. Literature Review

2.1. Semantic segmentation

Autonomous driving requires understanding the scene ahead. The release of datasets BDD100K [2] and

Cityscapes [3] has boosted the interest of using semantic segmentation as a way to process images for driving applications. Semantic segmentation consists in assigning a class to every pixel of an image. Early efforts in deep-learning based semantic segmentation used fully convolutional network (FCN) [4] which allowed for dense pixel classification of arbitrary sized images. However, FCN generated pixel labels from high level feature maps, resulting in relatively imprecise object boundaries. Subsequently, one of the main challenge of this task was to propose methods that extracted global context while providing pixel precise information to recover both the semantic categories present in the image and the pixel level boundaries of each object. Several approaches were developed to do so.

Modifying the original FCN design, encoder/decoder architectures were proposed, that compress information to capture the global context before decompressing it using either deconvolution [5], pooling indexes (SegNet) [6] or skip connections (U-net) [7] to recover the fine spatial information.

Other methods use pyramids of features [8, 9] extracted by an encoder model applied to the same image at different resolutions. The features maps can then be upsampled to match full scale the image size. More recent work proposed to use atrous convolutions [10, 11] to capture features at different spatial scales without the need to upsample the feature maps. Combining atrous convolutions and spatial pyramid pooling, Deeplabv3+ [10] proposes a simple and effective encoder-decoder FCN architecture.

Another approach consists in introducing attention mechanisms [12–14] to combine different levels of feature maps and better capture long-range image dependencies.

2.2. Multimodality and driving perception

The release of the KAIST Multispectral dataset [15] can be seen as an important step in the direction of combining color and thermal images for driving perception. The dataset provides physically aligned color and thermal images with bounding boxes annotations for a few semantic classes representing humans, such as *person*, *people* and *cyclist*.

Besides the release of the KAIST dataset, other works have considered the problem of multimodality in driving perception. The work of [16] proposed a cycle consistent GAN which generates synthetic color images from LWIR and synthetic LWIR from color images while enforcing that object detection remains accurate in both modalities. The authors obtain the best object detection accuracy when combining synthetic RGB and true LWIR images.

It should be noted that thermal and color images are often not perfectly aligned, due to the parallax effect, poor calibration or registration. This implies that naive input fusion, which consists in simply stacking the color and thermal images together by forming a 4-channel RGBT data input is rarely the most effective fusion strategy.

A number of works investigated better strategies to fuse different image modalities together in one model. Fusenet [17] was one of the first works dealing indirectly with the multimodal fusion problem, by proposing a double encoder architecture for the fusion of depth maps and color images.

The work of [1] proposed the MFNet (Multi-spectral Fusion Networks), a method for semantic segmentation on RGBT data. Their model consists of separate encoders for RGB and thermal data and fusion at the decoder level. The authors also published a semantic segmentation dataset on RGBT data, the MF dataset. The RTFNet method [18] performs semantic segmentation on RGBT data by fusing RGB and thermal features at multiple network depths, the method was also evaluated on the MF dataset [1]. Fusion of features extracted at multiple CNN layers as performed by [18] can be seen a strategy to attenuate feature disagreements, as the misalignment magnitude is decreased in deeper network layers.

FuseSeg-161 [19], a multi-modal fusion model has been able to fuse feature from thermal and RGB images in one end-to-end model. It consists of two encoders to extract features from input, one for RGB and another for thermal images with a two-stage fusion (TSF) strategy. In the first stage, the thermal feature maps are hierarchically added with the RGB feature maps in the RGB encoder. The fused feature maps are then concatenated with the corresponding decoder feature maps in the second fusion stage. TSF method achieved over state-the-art semantic segmentation performance under various lighting conditions.

More recently, FEANet [20] proposed a new way to extract detail spatial information with a two-stage Feature-Enhanced Attention Network. The Feature-Enhanced Attention Module (FEAM) improve excavation feature capacity from both the channel and spatial views, involving a better preservation of spatial information. Thus, the model shift more attention to high-resolution features which outperform other state-of-the-art techniques.

Another approach consists in training a deep neural network that weights differently thermal and color features. A first work in this direction was done by [21], where the model learns to estimate a position shift between pedestrian bounding boxes in color and thermal modalities.

Our work is related to the method proposed by [21] in the sense of performing a sort of confidence feature weighting. However, while [21] is focused on bounding box retrieval and weighting, we compute dense confidence and correlation weights that are better suited for semantic segmentation.
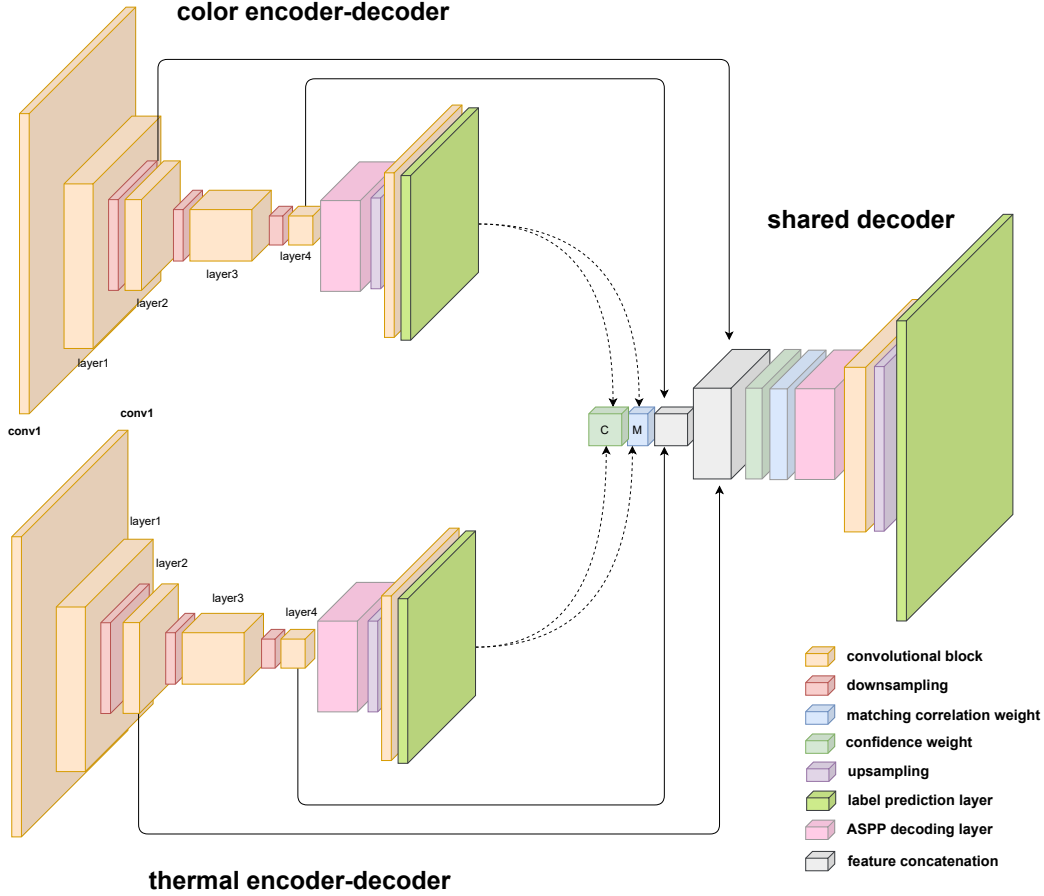
Figure 1. Summarized illustration of model architecture for weighted fusion semantic segmentation. The RGB and thermal encoder-decoder paths are specialized and both provide a rough segmentation used to compute a spatial confidence score. The confidence and matching correlation is used to weight the RGB and thermal features before fusing them in the final shared decoder. Solid arrows represent concatenation-based skip connections, dashed arrows represent data flow. Weighted concatenated fusion is performed with features coming from two different encoder depths (layer 2 and layer 4).

3. Proposed method

Given a dataset of paired color and thermal images with pixel-level target labels, we seek to perform semantic segmentation on this data by fusing color and thermal images in a manner that maximizes overall segmentation performance. Here we propose a mid-level feature fusion approach, which aims to limit the feature disagreements caused by misalignment between the thermal and color images.

Differently than the work of [18] which performs fusion of RGB and thermal features directly at the RGB encoder, we employ completely specialized thermal and color encoders.

We rely upon the DeepLabV3+ [10] model backbone, which has dilated (atrous) convolutions, limiting the loss of spatial resolution and better preserving structural details. The main decoder ingredient is the ASPP (Atrous Spatial Pyramid Pooling) which obtains multi-scale features obtained by atrous convolutions with different dilation rates.

We propose a double encoder-decoder variant of DeepLabV3+ where fusion is performed at two different feature resolutions at the decoder side, with a shared decoder for the final segmentation. A motivation for our approach comes from the observation that rich encoder features from pretrained networks such as ResNet [22] could be adapted as seamless as possible for multiple deep learning tasks, in order to minimize training time and maximize prediction performance.

Moreover, we perform a weighted feature fusion which takes into account feature agreement and feature confidence. Our proposed architecture is illustrated in Fig. 1. We detail the two weighting strategies in the following subsections.
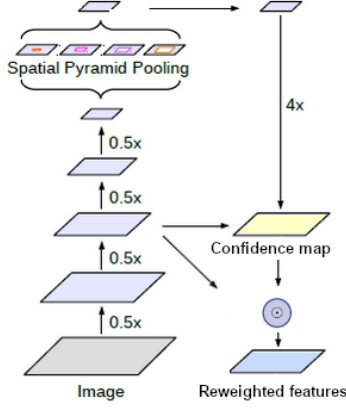
Figure 2. Encoder-decoder path for modality confidence-aware feature weighting based on DeepLabv3+ architecture. The illustrated procedure is computed for both color and thermal images. Resulting reweighted features from color and thermal are concatenated together as input for the shared decoder.

3.1. Confidence weighted fusion

It can be claimed that direct fusion of RGB and thermal features may be problematic depending on the time of day, weather and visibility conditions of the captured scene.

For instance, during day time and favorable weather, RGB features could provide the best features to perform segmentation, as thermal images are only one-channel and texture details from objects are hardly noticeable.

On the other hand, during night time and at low light, thermal images may provide better features to segment hardly visible objects, in particular the ones with distinguishable changes in temperature compared to the background such as pedestrians and cars.

If we always give equal importance for features coming from both the modalities, the model may achieve lower performance in segmenting objects when one of the modalities provide unreliable features.

To alleviate this issue, we propose a confidence-based weighting before performing feature fusion, such that RGB and the thermal features receive difference importance based on their reliability. The decoder and the segmentation heads of each modality are able to perform a preliminary segmentation, from which a confidence can be estimated.

Based on the model confidence of the preliminary segmentation, we may be able to give more weight to the more reliable features.

Figure 2 illustrates how confidence is computed for each imaging modality.

Formally, we denote the color image as I_c and thermal image as I_t. Each image modality I_m, with $m \in \{c, t\}$ has a set of features computed by a specialized encoder $\mathbf{f}_m = e_m(I_m)$ and a modality prediction computed by specialized decoder $\mathbf{y}_m = d_m(\mathbf{f}_m)$.
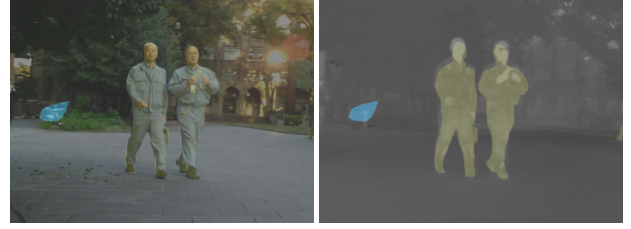

Figure 3. Example of misalignment between color and thermal images. It can be seen that the label for the pedestrian class (overlayed in yellow) is not perfectly aligned with the pedestrians on the thermal image. Such misalignment can cause the RGB and the thermal encoder features to disagree. We propose a correlation feature weighting to alleviate this issue.

The confidence weight matrix C_m for modality m is given by an estimation of the confidence on the most probable prediction:

$$C_{m_i} = \max\left(\frac{\exp(\mathbf{y}_{m_i})}{\sum_j \exp(\mathbf{y}_{m_j})}\right) \quad (1)$$

where $\mathbf{y}_{m_i}$ denotes the one-hot logits for modality m at spatial position i.

The confidence map C_m is used to re-weight the features. Color and thermal features receive different weights, based on the confidence of the specialized color and thermal decoders. Weighting is performed by point-wise hadamard product, and resulting weighted color and thermal features are concatenated together:

$$\mathbf{f}'_c = C_c \odot \mathbf{f}_c \quad (2)$$
$$\mathbf{f}'_t = C_t \odot \mathbf{f}_t \quad (3)$$
$$\mathbf{f}'_{ct} = \mathbf{f}'_c c \oplus \mathbf{f}'_t. \quad (4)$$

where $\odot$ denotes the hadamard product and $\oplus$ denotes concatenation.

3.2. Matching correlation weighted fusion

Another possible issue of directly fusing color and thermal features is the spatial misalignment between both images. Figure 3 illustrates a sample image from MF dataset [1] where feature disagreement can be caused due to image misalignment.

As it may be hard to achieve perfect pixel by pixel alignment between the different imaging modalities, we propose to evaluate the alignment through a matching layer.

Inspired by the work of [23] on deep image alignment, we propose to employ a correlation layer which accounts to evaluate the matching between features coming from the two different encoders.

The correlation layer computes pair-wise feature similarities between the color and thermal decoder and the ob-

tained scores are normalized such that weak matched features are suppressed by down weighting.

In order to deal with strong appearance differences between color and thermal features, we estimate semantic matching between the last specialized decoder, such that correlation is invariant to appearance changes.

Formally, given $\mathbf{y}_c = d_c(\mathbf{f}_c)$ corresponding to the color decoder predictions and $\mathbf{y}_t = d_t(\mathbf{f}_t)$ corresponding to the thermal decoder predictions, the correlation weight matrix $M_{ct} \in \mathbb{R}^{h \times w}$ is obtained by computing modality-wise prediction correlations (normalized scalar product of a pair of predictions), followed by a channel compressing convolutional block:

$$M_{ct} = c(\|\sigma(\overline{\mathbf{y}}_t^T \overline{\mathbf{y}_c})\|_2) \tag{5}$$

where σ denotes ReLU activation, $\overline{\cdot}$ denotes spatially flattening ($y_m \in \mathbb{R}^{k \times h \times w}$, $\overline{y_m} \in \mathbb{R}^{k \times N}$, $N = h \times w$, k being the number of predicted classes) , $\|.\|_2$ denotes L_2 vector normalization and c is a convolutional block performing channel compression ($N \times h \times w$ input dimensions, $h \times w$ output dimensions). Note that c acts in analogy to a feature regressor estimating parametric alignment [23], but here we instead estimate alignment-aware spatial weights.

Then, the concatenated confidence-weighted features obtained by Eq. 4 are re-weighted by the correlation map:

$$\mathbf{f}_{ct}'' = M_{ct} \odot \mathbf{f}_{ct}'. \tag{6}$$

Finally, the obtained feature map $\mathbf{f}_{ct}''$ is used as input to the shared decoder head to produce the final semantic segmentation.

4. Experiments

In this section we perform semantic segmentation experiments that shows the advantages of the proposed method.

4.1. Dataset

We perform semantic segmentation experiments on the MF dataset released by the authors of the method MFNet [1]. To the best of our knowledge, this is the only public dataset with paired color and thermal images and dense annotated masks in the context of urban street scenes. The dataset has eight hand-labeled object classes (car, person, bike, curve, car stop, guardrail, color cone, bump) and one unlabelled background class. This dataset contains 1569 pairs of RGB and thermal images, in which 820 taken at daytime and 749 taken at nighttime. We follow the dataset splitting scheme proposed in [1], with the training set consisting of 784 pairs of images. The validation set consists of 392 pairs of images, and the test set contains 392 image pairs.

4.2. Quantitative comparison

In this section, we compare our model with previous works on color-thermal semantic segmentation: FRRN [24], BiSeNet [25], DFN [26], HRNetV2 [27], MFNet [1], FuseNet [17], DACNN [28], RTFNet(50-152) [18], FuseSeg-161 [19] and FEANet [20].

For comparison with previous work, we compute the mean IoU (Intersection Over Union) and the mean accuracy. The IoU (also called Jaccard index) is arguably the most adapted metric to evaluate semantic segmentation, it measures the overlap between predicted and ground truth mask, while the accuracy simply measures the total percentage of correct predictions.

In Table 1, it can be seen that our model improves mean IoU over all classes in the MF dataset. It improves IoU on three classes and notably on the person class, which is arguably one of the most important category for driving perception.

It should be noted that similarly to other datasets for driving perception, MF dataset is extremely class unbalanced. This implies that the IoU metric is a better indicator of segmentation performance than accuracy for this dataset.

4.3. Qualitative comparison

In Fig. 4, we show a number of semantic segmentation results obtained with our method. It can be seen that segmentation of important details are improved compared to previous work. In the zoomed in parts of columns 4 and 5, it can be seen that our method segments better the pedestrian legs. In the last two columns, it can also be seen that our method is the only one able to retrieve the person and the bike objects, which are hardly visible in the color image due the small size of the object and a saturated highlight.

4.4. Ablation Study

In Table 2, we consider a number of variants of our model, all of them also based on the DeepLabv3+ backbone. We evaluate the individual contributions of the confidence weighting, the concatenated decoder-level fusion and correlation weighting.

We also evaluate the model performance in different lighting conditions, only day, and only night. It can be seen that our proposed model obtains the highest mIoU in all lighting conditions, compared to its simpler variants.

4.5. Learned features visualization

In Fig. 5, we visualize the mean output feature responses of our proposed model before and after semantic segmentation training on RGB-T data. It can be seen that for the given image sample, the shallow second encoder layer gives somewhat similar responses before and after training. On the other hand, the deeper feature response at deeper layer

Table 1. Comparison with previous work on the MF test set for Acc (%) and IoU (%). We denote 3-channel RGB and 4-channel RGB-thermal data respectively as 3c and 4c. Note that mAcc and mIoU are calculated as a mean over all classes, including unlabelled. Bold font highlights the highest score in each column, underlined font highlights the second highest score. We report an improvement of 2% in overall mIoU compared to previous work.

Methods	Car		Person		Bike		Curve		Carstop		Guardrail		Cone		Bump		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
FRRN-4c [24]	81.9	74.7	66.2	60.8	62.8	50.3	41.2	35.0	12.5	11.5	0.0	0.0	37.2	34.0	35.2	34.6	48.5	44.2
FRRN-3c [24]	80.0	71.2	53.0	46.1	65.1	53.0	34.0	27.1	21.6	19.1	0.0	0.0	34.7	32.5	36.2	30.5	47.1	41.8
BiSeNet-4c [25]	89.7	84.1	72.0	63.2	74.1	60.1	45.1	36.7	34.2	25.3	18.2	5.0	47.4	42.2	39.8	35.9	57.7	50.0
BiSeNet-3c [25]	90.0	84.5	65.0	54.3	75.0	61.4	32.1	25.7	32.3	26.2	3.2	0.9	49.6	43.3	48.1	40.5	54.9	48.2
DFN-4c [26]	90.0	84.4	73.2	65.0	75.5	60.9	54.0	40.4	**38.9**	25.7	10.2	2.7	48.3	42.5	55.8	47.4	60.5	52.0
DFN-3c [26]	90.7	81.4	67.7	52.8	71.5	57.5	49.2	34.9	35.1	23.8	4.1	1.4	44.2	31.0	54.6	47.5	57.3	47.5
HRNet2-4c [27]	92.8	87.6	79.3	71.0	78.3	63.4	59.8	42.5	25.7	19.1	18.8	0.0	56.5	49.8	63.5	44.5	63.7	53.2
HRNet2-3c [27]	92.2	86.6	73.1	59.8	74.9	61.3	47.0	33.2	23.8	28.7	7.3	0.0	54.6	47.2	61.5	46.2	60.9	51.3
MFNet [1]	77.2	65.9	67.0	58.9	53.9	42.9	36.2	29.9	19.1	9.9	0.1	**8.5**	30.3	25.2	30.0	27.7	45.1	39.7
FuseNet [17]	81.0	75.6	75.2	66.3	64.5	51.9	51.0	37.8	28.7	15.0	0.0	0.0	31.1	21.4	51.9	45.0	52.4	45.6
DACNN [28]	85.2	77.0	61.7	53.4	76.0	56.5	40.2	30.9	9.9	29.3	22.8	6.4	32.9	30.1	36.5	32.3	55.1	46.1
RTFNet-50 [18]	91.3	86.3	78.2	67.8	71.5	58.2	69.8	43.7	32.1	24.3	13.4	3.6	40.4	26.0	73.5	57.2	62.2	51.7
RTFNet-152 [18]	93.0	87.4	79.3	70.3	76.8	62.7	60.7	45.3	38.5	29.8	0.0	0.0	45.5	29.1	74.7	55.7	63.1	53.2
FuseSeg-161 [19]	93.1	87.9	81.4	71.7	78.5	64.6	68.4	44.8	29.1	22.7	63.7	6.4	55.8	46.9	66.4	47.9	70.6	54.5
FEANet [20]	**93.3**	87.8	**82.7**	71.1	76.7	61.1	65.5	46.5	26.6	22.1	**70.8**	6.6	**66.6**	55.3	77.3	48.9	**73.2**	55.3
Ours	91.7	86.7	81.3	**72.2**	76.6	62.5	58.9	**46.7**	36.2	28.0	35.2	5.1	56.9	50.7	74.8	**65.8**	67.9	**57.3**

seems to vary more in its response after training, by learning to spot at the two pedestrians.

4.6. Implementation and setup details

Our method is implemented using the pytorch library [29] and the torch segmentation models library [30]. For all experiences with our method, we use ResNet-101 model [22] pretrained on Imagenet as backbone. We perform concatenation and confidence weighting with features from layer 2 and layer 4 of DeepLabv3+ encoder. Bilinear interpolation is used to downsample or upsample confidence map C_m and matching map M_{ct} whenever needed to fit the encoder feature spatial resolutions from layer 2 and layer 4.

For a fair comparison, we perform similar training setup as compared to previous methods. We adopt mostly the same hyperparameters as employed by the authors of the RTFNet method [18]. We use random cropping and random flipping as image augmentation during training.

For optimization, we use SGD with momentum set to 0.9, weight decay set to 0.0005, initial learning rate set to 0.01 and a maximum of 50 training epochs. We employ exponential decaying learning rate scheduler with gamma set to 0.95. As commonly done in the field, we minimize the cross entropy loss between the predicted and the target labels.

5. Conclusion

In this paper we have presented a new feature fusion approach for the task of semantic segmentation of paired color RGB and thermal LWIR images. We proposed Doo-DLeNet, a double DeepLab net which fuses multimodal images taking into account the feature disagreements and feature confidences. We have shown that this approach improves semantic segmentation performance compared to previous works, in particular for the correct segmentation of pedestrians. Future work may be focused on evaluating different model backbones and to refine further segmentation details.
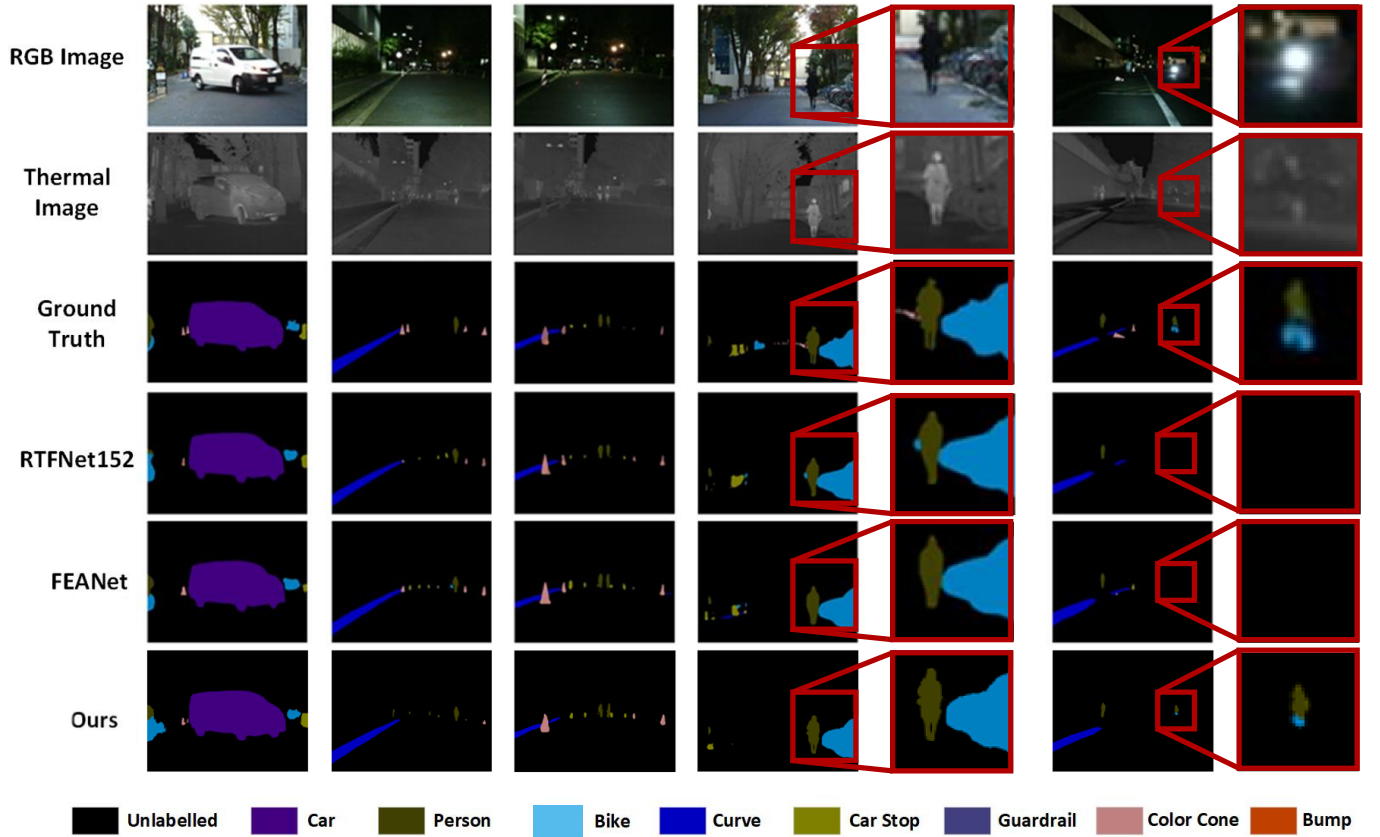
6. Acknowledgements

Figure 4. Qualitative results where the robustness of our method can be appreciated for daytime and nighttime scenes of MF test set split. In the last two columns (original and zoomed in images), it can be seen that our method is the only segmenting the person and the bike objects.
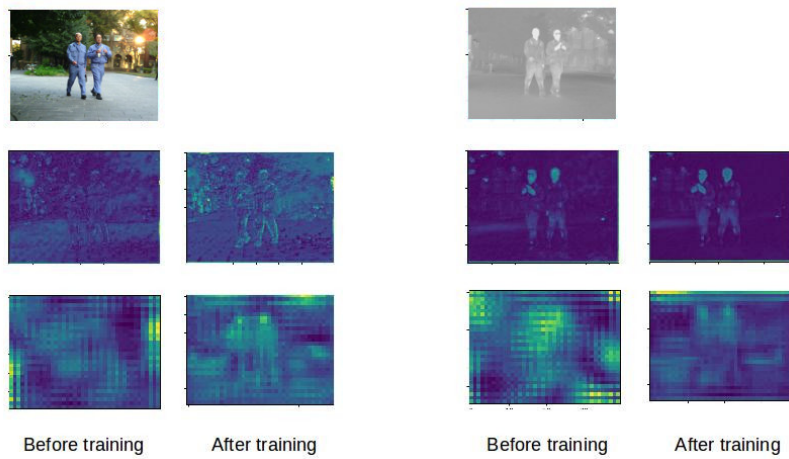

Figure 5. Illustration of features learned by color encoder (left two columns) and thermal encoder (right two columns) at encoder layer 2 and layer 4 (second and third rows) before (pretrained on Imagenet) and after training for semantic segmentation. Note that at the deeper level the feature responses seem less invariant to the imaging modality after training.

Table 2. Ablation study: comparison of mAcc (%) and mIoU (%) on the day-only, night-only and regular test set of MF dataset for different model variants. For a fair comparison, all of the variants are trained with the same optimizer parameters and are based on DeepLabv3+ with Resnet-101 pretrained backbone. It can be noted the advantage of our approach - simpler model variants based on DeepLabv3+ result in lower mIoUs for all lighting conditions.

Variants	Daytime		Nighttime		Both	
	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU
Only RGB	54.3	44.4	56.5	48.3	61.2	50.7
Only thermal	48.2	39.0	64.0	51.9	65.6	50.1
Input fusion - stacked RGBT	48.7	42.8	56.0	51.0	57.1	51.1
Unweighted fusion	60.7	49.0	61.4	53.6	64.9	54.7
Only-confidence weighting	**63.3**	49.6	61.2	54.1	65.4	55.2
Ours	58.8	**50.1**	**64.1**	**55.7**	**67.9**	**57.3**

References

[1] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," *IEEE International Conference on Intelligent Robots and Systems*, vol. 2017-Septe, pp. 5108–5115, 2017. 1, 2, 4, 5, 6

[2] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 2633–2642, Computer Vision Foundation / IEEE, 2020. 1

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[4] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017. 2

[5] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, pp. 1520–1528, 2015. 2

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation.," *CoRR*, vol. abs/1511.00561, 2015. 2

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015. 2

[8] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1915–1929, 2012. 2

[9] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017. 2

[10] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2, 3

[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017. 2

[12] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on*

computer vision and pattern recognition, pp. 3146–3154, 2019. 2

[13] C. Yu, J. Wang, C. Gao, G. Yu, C. Shen, and N. Sang, "Context prior for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12416–12425, 2020. 2

[14] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, 2021. 2

[15] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baselines," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[16] R. Abbott, N. M. Robertson, J. M. del Rincón, and B. Connor, "Unsupervised object detection via LWIR/RGB translation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pp. 407–415, IEEE, 2020. 2

[17] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Asian conference on computer vision*, pp. 213–228, Springer, 2016. 2, 5, 6

[18] Y. Sun, W. Zuo, and M. Liu, "RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019. 2, 3, 5, 6

[19] Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, "Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion," *IEEE Transactions on Automation Science and Engineering*, vol. PP, pp. 1–12, 06 2020. 2, 5, 6

[20] F. Deng, H. Feng, M. Liang, H. Wang, Y. Yang, Y. Gao, J. Chen, J. Hu, X. Guo, and T. L. Lam, "Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation," *CoRR*, vol. abs/2110.08988, 2021. 2, 5, 6

[21] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 5126–5136, IEEE, 2019. 2

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778, IEEE Computer Society, 2016. 3, 6

[23] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2553–2567, 2019. 4, 5

[24] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," *CoRR*, vol. abs/1611.08323, 2016. 5, 6

[25] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," *CoRR*, vol. abs/1808.00897, 2018. 5, 6

[26] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," *CoRR*, vol. abs/1804.09337, 2018. 5, 6

[27] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions," *CoRR*, vol. abs/1904.04514, 2019. 5, 6

[28] W. Wang and U. Neumann, "Depth-aware CNN for RGB-D segmentation," *CoRR*, vol. abs/1803.06791, 2018. 5, 6

[29] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019. 6

[30] P. Yakubovskiy, "Segmentation models." https://github.com/qubvel/segmentation_models, 2019. 6