GMFNet: Gated Multimodal Fusion Network for Visible-Thermal Semantic Segmentation

Etienne Balit[1] and Amine Chadli[1]

Neovision
{etienne.balit, amine.chadli}@neovision.fr

Abstract. Robust perception in challenging illuminations is an essential component for autonomous driving. Visible and thermal cameras have been shown to be complementary for the task of semantic segmentation under day and night settings. Most previous methods combine features from the thermal and visible modalities by feeding a combination of both as input to a multimodal decoder. In this paper, we propose a new architecture for visible and thermal fusion for semantic segmentation. It's composed of three parallel U-Net, one per modality and a central one for the multimodal fusion. A gating mechanism in the multimodal encoder is used to actively weight each modality per pixel. Moreover, we use multi-task learning and a specific data augmentation scheme to train these models to be robust to disturbances affecting a single modality. Experiments show that the proposed approach outperforms the current state-of-the-art on the MFNet dataset for a comparable inference-speed.

Keywords: Semantic Segmentation, Autonomous Driving, Gated Multimodal Fusion, Multitask Learning, Deep Neural Networks

1 Introduction

Semantic segmentation is essential for many autonomous driving applications such as scene understanding, environment modeling, and path planning. Most approaches in the literature use Convolutional Neural Networks (CNN) for semantic segmentation, applied to visible images from RGB cameras. However, using only the visible information has some limitations, such as insufficient lighting sensitivity. Thermal cameras are more robust to these conditions since they are sensitive to far infrared radiations emitted by objects.

Previous work have investigated how to design a CNN for the task of multimodal segmentation [1,6,5]. The same problematic of multimodal fusion has also been studied for other tasks and domains. CentralNet [7] has been proposed as a general architecture for fusing features by building a central pathway that draws input from each modality at multiple level. In the domain of object detection, GIF networks [3] proposed a mechanism to modulate the fusion of the modalities based on the input data.

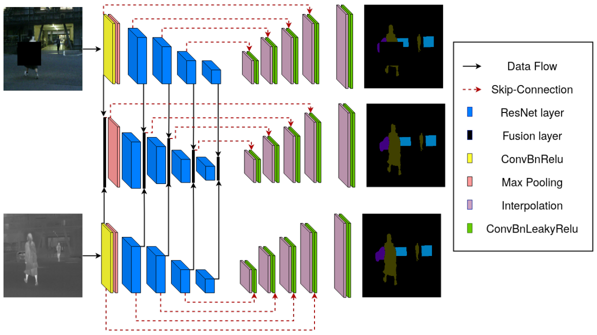

Fig. 1: The overall architecture of our proposed GMFNet model.

We propose a new architecture for multimodal semantic segmentation based on the U-Net [4] architecture. Our model leverages the benefits from both visible and thermal images by fusing their intermediate features using a gating mechanism. We train our model using multi-task learning and a specific data augmentation scheme to improve its robustness to quality degradation that might occur in one of the modalities.

2 Gated Multimodal Fusion Network

The overall architecture of our proposed GMFNet (Gated Multimodal Fusion Network) is shown in figure 1. The model is composed of three U-Nets: a central and two lateral U-Nets. The lateral U-Nets are dedicated to learning modality-specific features while the central U-Net combine the features from different modalities using Gated Fusion Modules. At the inference time, only the central decoder along with the three encoders are used.

Lateral and Central U-Nets Each U-Net is composed of an encoder and a decoder. In the case of the lateral U-Net, encoders use a ResNet network [2] pretrained on ImageNet, excluding the average pooling and classifier layers. The central U-Net encoder uses a similar architecture with the addition of fusion layers to replace the first ConvBNRelu block as well as between each ResNet block. This fusion layer combines features from the two lateral U-Net encoders using a gated fusion mechanism. We use the lightweight detector architecture from MFNet[1] for the three independent decoders and skip connections from the encoder to the decoder layers. Each decoder is responsible for constructing the semantic segmentation mask corresponding to its input.

Gated Fusion Module The fusion module, presented in figure 2, takes inspiration from [3]. It computes a per pixel weighted sum of the unimodal feature maps. The weight map for each modality is computed by a weight generator network taking as input a concatenation of the feature maps from both modalities. The objective is to allow the network to selectively aggregate the information from both modalities depending on the input.

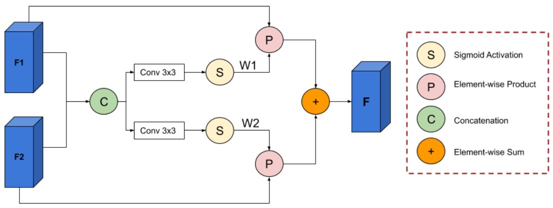

Fig. 2: Gated Fusion Module

Multi-Task Learning The three U-Net are trained jointly in a multi-task setting to force each lateral U-Net to exploit all the available information in its respective modality. Moreover, each U-Net loss is taken as a linear combination of cross entropy (CE) and weighted cross entropy since the classes in the dataset are dramatically imbalanced.

Data Augmentation Besides the usual data augmentation techniques as horizontal flip, Gaussian noise, and random brightness, we also use a special data augmentation strategy inspired by [3]. It consists of blanking a random rectangular area in one of the modalities at a time. We found that combining this strategy with the multi-task learning and the weighted sum fusion makes the model more robust by learning to extract pertinent information from each modality.

3 Experimental Evaluation

In this section, we evaluate the performances of our approach on the MFNet dataset [1]. The MFNet dataset is a visible and thermal dataset for semantic segmentation of 9 different classes. It consists of 1569 pairs of RGB and thermal images from urban scenes with an image resolution of 480x640. Images are split into day and night settings. We follow the same dataset split scheme proposed in the MFNet paper for training and evaluation.

Training The ResNet encoders are initialized with pretrained weights on ImageNet, whereas the other layers are initialized using the Xavier scheme. We combine the CE loss and the weighted CE with a ratio of 0.8 to 0.2 respectively. The class weights are inversely proportional to their frequency in the training data. We train the model with Stochastic Gradient Descent optimizer (SGD) with a batch size of 16, a momentum of 0.9 and a weight decay of 0.0005. We adopt the exponential decay scheme to gradually decrease the learning rate with an initial learning rate of 0.01 and rate decay of 0.975.

Results In table 1, we compare our GMFNet model with MFNet [1], PSTNet [5] and RTFNet [6]. MFNet and RTFNet results are computed using the pretrained models provided by the authors whereas PSTNet results are reprinted from the paper. We report on the mean accuracy (mAcc) and the mean Intersection over Union (mIoU). The inference speed (FPS) is measured on a NVIDIA GTX 1070

Methods	Day		Night		All		FPS
	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	
MFNet [1]	39.9	34.5	36.0	32.1	41.1	36.5	**138.5**
PSTNet [5]	—	—	—	—	—	48.4	52.6
RTFNet-50 [6]	57.7	44.3	60.3	51.5	62.9	51.2	50.8
RTFNet-152 [6]	60.3	45.7	61.6	54.1	64.1	52.7	18.2
GMFNet-18 (ours)	56.0	44.7	62.6	52.9	64.7	52.8	85.5
GMFNet-34 (ours)	58.3	45.7	66.3	54.4	68.7	53.8	53.4
GMFNet-50 (ours)	**61.6**	**46.9**	**67.9**	**55.3**	**70.5**	**54.8**	36.7

Table 1: Our model compared to competitive methods on MFNet test set [1].

GPU. Our model outperforms all other models on all reported metrics while preserving a competitive inference time.

4 Conclusion

We proposed an efficient CNN architecture to leverage visible and thermal information for semantic segmentation of urban scenes. Our model uses a gated fusion mechanism to make use of informative data in each modality. We use multi-task learning as a regularization technique to make our model more robust and generalizable. Presented experiments show that our model outperforms the state of the art on the MFNet test set [1] and allows real-time inference. A perspective that we would like to explore in the future involves making the fusion mechanism more robust to data misalignment.

References

1. Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., Harada, T.: 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). pp. 5108–5115 (2017)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. arXiv e-prints arXiv:1512.03385 (Dec 2015)
3. Kim, J., Koh, J., Kim, Y., Choi, J., Hwang, Y., Choi, J.W.: Robust Deep Multi-modal Learning Based on Gated Information Fusion Network. arXiv e-prints arXiv:1807.06233 (Jul 2018)
4. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
5. Shivakumar, S.S., Rodrigues, N., Zhou, A., Miller, I.D., Kumar, V., Taylor, C.J.: PST900: RGB-Thermal Calibration, Dataset and Segmentation Network. arXiv e-prints arXiv:1909.10980 (Sep 2019)
6. Sun, Y., Zuo, W., Liu, M.: RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes. IEEE Robotics and Automation Letters **4**(3), 2576–2583 (July 2019). https://doi.org/10.1109/LRA.2019.2904733
7. Vielzeuf, V., Lechervy, A., Pateux, S., Jurie, F.: Centralnet: a multilayer approach for multimodal fusion. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018)