Incorporating depth information into few-shot semantic segmentation

Yifei Zhang, Désiré Sidibé, Olivier Morel, Fabrice Meriaudeau

Incorporating Depth Information into Few-Shot Semantic Segmentation

Yifei Zhang*‡, Désiré Sidibé†, Olivier Morel*, Fabrice Meriaudeau*
*ERL VIBOT CNRS 6000, ImViA, Université Bourgogne Franche Comté, 71200, Le Creusot, France
†Université Paris-Saclay, Univ Evry, IBISC, 91020, Evry, France
‡Email: yifei.zhang@u-bourgogne.fr

***Abstract*—Few-shot segmentation presents a significant challenge for semantic scene understanding under limited supervision. Namely, this task targets at generalizing the segmentation ability of the model to new categories given a few samples. In order to obtain complete scene information, we extend the RGB-centric methods to take advantage of complementary depth information. In this paper, we propose a two-stream deep neural network based on metric learning. Our method, known as RDNet, learns class-specific prototype representations within RGB and depth embedding spaces, respectively. The learned prototypes provide effective semantic guidance on the corresponding RGB and depth query image, leading to more accurate performance. Moreover, we build a novel outdoor scene dataset, known as Cityscapes-3^i, using labeled RGB images and depth images from the Cityscapes dataset. We also perform ablation studies to explore the effective use of depth information in few-shot segmentation tasks. Experiments on Cityscapes-3^i show that our method achieves excellent results with visual and complementary geometric cues from only a few labeled examples.**

I. Introduction

With the advent of multiple sensory modalities, multimodal data has attracted much attention in the computer vision domain. As one of the most commonly-used modalities, depth-sensing cameras provide rich geometric information of the scenes. Several deep neural networks exploit these depth maps as an addition image channel [1, 2] or point cloud in 3D space [3, 4]. Arguably, the integration of additional depth features in semantic image segmentation leads to significant performance improvement. Different from fully supervised semantic segmentation, few-shot segmentation concentrates on the generalization of segmentation ability to unseen categories given only a few samples. To be specific, some existing few-shot segmentation methods learn the representative features for each target class in the support images, then guide the pixel-level prediction on the query image. However, the generalization and discrimination abilities of these methods still remain to be improved, especially for complex scenes.

For the above reasons, we take inspiration from existing RGB-centric methods for few-shot semantic segmentation and propose a two-stream deep neural network based on metric learning, called RDNet. The original intention of our work is to incorporate supplementary depth information into a few-shot segmentation model. As shown in Figure 1, the proposed RDNet employs both RGB and depth images of the same scene in the support and query set. The abstract foreground and background features of target classes are embedded into the corresponding embedding space. These prototype representations learned from RGB and depth inputs provide further similarity guidance on the query feature. Then our RDNet fuses multiple probability maps generated by the two streams into a joint prediction. In this way, our method outperforms the baseline networks with higher accuracy.

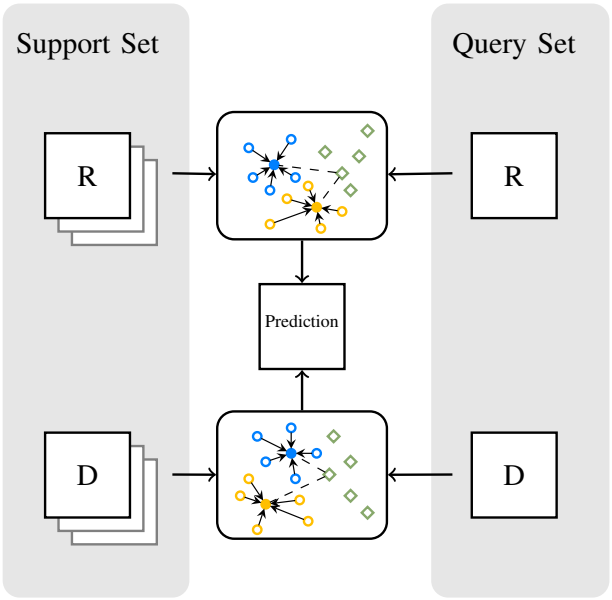


Fig. 1: Overview of the proposed RDNet approach. R and D indicate the RGB and depth image input, respectively. The abstract features of labeled support images are mapped into the corresponding embedding space (circles). Multiple prototypes (blue and yellow solid circles) are generated to perform semantic guidance (dashed lines) on the corresponding query features (rhombus). RDNet further produces the final prediction by combining the probability maps from RGB and depth stream.

Furthermore, we report the experimental results on a new benchmark dataset, Cityscapes-3^i. Different from the frequently-used PASCAL-5^i dataset for object segmentation, Cityscapes-3^i is derived from the large-scale Cityscapes dataset, which consists of diverse urban street scenes at varying times. Complex category information greatly increases the difficulty of scene understanding, especially with limited supervisory samples. To tackle this challenge, we conduct various comparative experiments to exploit the potential of

depth information and effective fusion pattern. To the best of our knowledge, we are the first to facilitate the few-shot segmentation problem with additional depth cues. This work also promotes the use of multimodal data in the few-shot learning field. To sum up, the main contributions are summarized as follows:

- We propose a metric learning-based deep neural network for few-shot semantic segmentation, which processes RGB-D data in two streams.
- We define a new few-shot segmentation benchmark on the Cityscapes dataset, named Cityscapes-3^i.
- Extensive experiments and ablation studies demonstrate the effectiveness of the proposed RDNet, as well as the positive effects of geometric information in limited supervisory scene understanding.

The remainder of this paper is organized as follows. Section II reviews the related work in fully-supervised RGB-D semantic segmentation and state of the art for few-shot segmentation. Section III describes the proposed two-stream architecture in detail. Section IV presents a new few-shot segmentation benchmark called Cityscapes-3^i. Section V reports the extensive experimental results as well as the ablation studies. Conclusions are drawn in Section VI.

II. Related work

a) RGB-D Semantic Segmentation: Recent advances in deep learning enable the fully-supervised semantic segmentation on 2D images to achieve a significant performance enhancement [5, 6, 7]. With the advent of various depth sensors, a growing number of approaches have been proposed which use depth cues for complex scene understanding. To name a few, Qi et al. [8] proposed a 3D graph neural network that builds a k-nearest neighbor graph on top of the 3D point cloud. This method employs both the 2D appearance information and 3D geometric relations to produce excellent results on RGB-D segmentation benchmarks. In [3, 4], a series of PointNet was proposed to take point clouds as input and output point clouds directly. These architectures can effectively learn representative features from informative points of the point cloud.

Otherwise, some works [1, 2, 9, 10, 11] attempt to tackle RGB-D semantic segmentation tasks by processing geometric information as a supplementary image or an additional channel. Namely, multimodal image input was fed into an elaborated neural network for a joint prediction. As an alternative method, RGB and depth images can be separately trained in a two-branch network. Moreover, Gupta et al. [12] presented a geocentric embedding algorithm to generate three channels HHA images, which contain horizontal disparity, surface normal, and height above ground. In our work, we process depth information by combining a complementary depth stream with the RGB one. Then our model maps the support depth data into a depth embedding space, which provides further semantic guidance on the query image.

b) Few-shot Segmentation: Many approaches for few-shot learning are proposed to generalize prior knowledge to new tasks using only a few examples. Some research [13, 14] introduced the metric learning-based matching network for the few-shot classification task. The non-parametric structure facilitates the generalization of models to new training sets. Snell et al. [15] presented a method to represent the prototypes per class in a representation space, known as Prototypical Networks. Moreover, several studies such as [16] have focused on the graph-based methods for few-shot learning.

Furthermore, few-shot semantic segmentation refers to the pixel-level prediction of new categories on the query set, given only a few labeled support images. For example, Shaban et al. [17] first presents a dual branch parallel network for one-shot segmentation, known as OSLSM, including a conditioning branch and a segmentation branch. The conditioning branch extracts representative high-level features from the supporting image-label pair, whilst the segmentation branch integrates the parameters learned from the conditioning branch and performs a segmentation mask on the query image. Other variants of OSLSM include Co-FCN [18], PL+SEG [19] and MDL [20]. All of which extend such dual branch structure to achieve a substantial performance improvement. In the AMP model, Siam et al. [21] replaces the guidance branch with a multi-resolution weight. Moreover, SG-One [22] proposed a Masked Average Pooling block (MAP) to extract the representative vectors of support objects. Then the segmentation mask was predicted via a similarity guidance network. More recently, Wang et al. [23] presents a novel prototype alignment network, called PANet, based on non-parametric metric learning.

III. Methodology

A. Problem setting

Few-shot semantic segmentation involves three datasets: a training set D_{train}, a support set D_s, and a query set D_q. The segmentation model is trained on D_{train}, and evaluated on D_s and D_q. Moreover, we adopt the training and testing protocols in [17]. Suppose the set of semantic classes in D_{train} is C_{seen}. We assume that the set of classes at test time, C_{unseen}, does not overlap with C_{seen}, i.e. $C_{seen} \cap C_{unseen} = \emptyset$. We formally define these datasets in the following lines:

- $D_{train} = (x_i^R, x_i^D, y(l)_i)_{i=1}^N$, where x_i^R is a color image, x_i^D is a depth image of the same scene, $y(l)_i$ denotes the corresponding segmentation mask of class l ($l \in C_{seen}$), and N indicates the number of training examples.
- $D_s = (x_j^R, x_j^D, y(l)_j)_{j=1}^M$, where x_j^R and x_j^D denote the corresponding RGB and depth image, $y(l)_j$ is the mask for the semantic class l ($l \in C_{unseen}$), and M indicates the number of labeled samples given in the test phase.
- $D_q = (x_j^R, x_j^D)_{j=1}^n$ is the query set of n pairs of RGB and depth images. Evaluations on D_q show the relative performance of the models.

Therefore the goal of few-shot segmentation is to train a model f with high discriminative power and generalizability from D_{train}, then produces a segmentation prediction on D_q

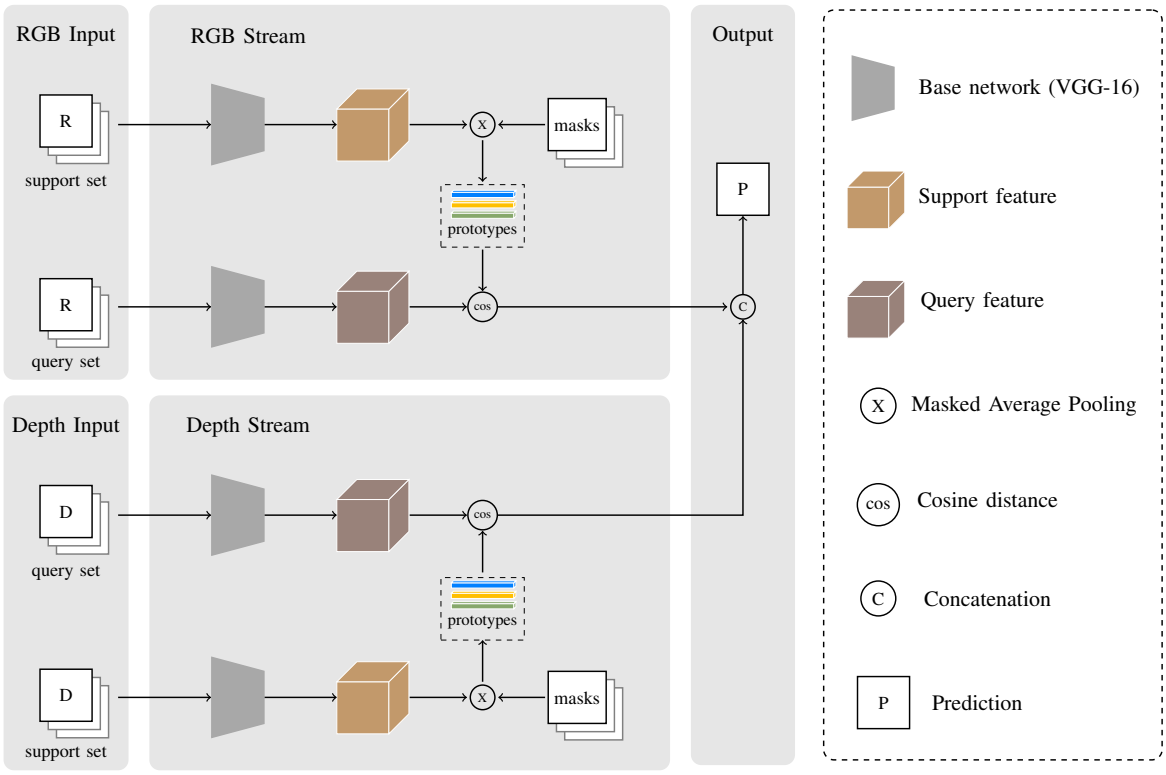


Fig. 2: Details of the proposed RDNet architecture. It includes two mirrored streams: an RGB stream and a depth stream. Each stream processes the corresponding input data, including a support set and a query set. The prototypes of support images are obtained by masked average pooling. Then the semantic guidance is performed on the query feature by computing the relative cosine distance. The results from these two streams are combined at the late stage.

given a support set D_s. Usually, if the the support set consists of K labeled samples for each of C semantic classes, we consider such few-shot learning problem as C-way K-shot segmentation task.

B. Proposed model

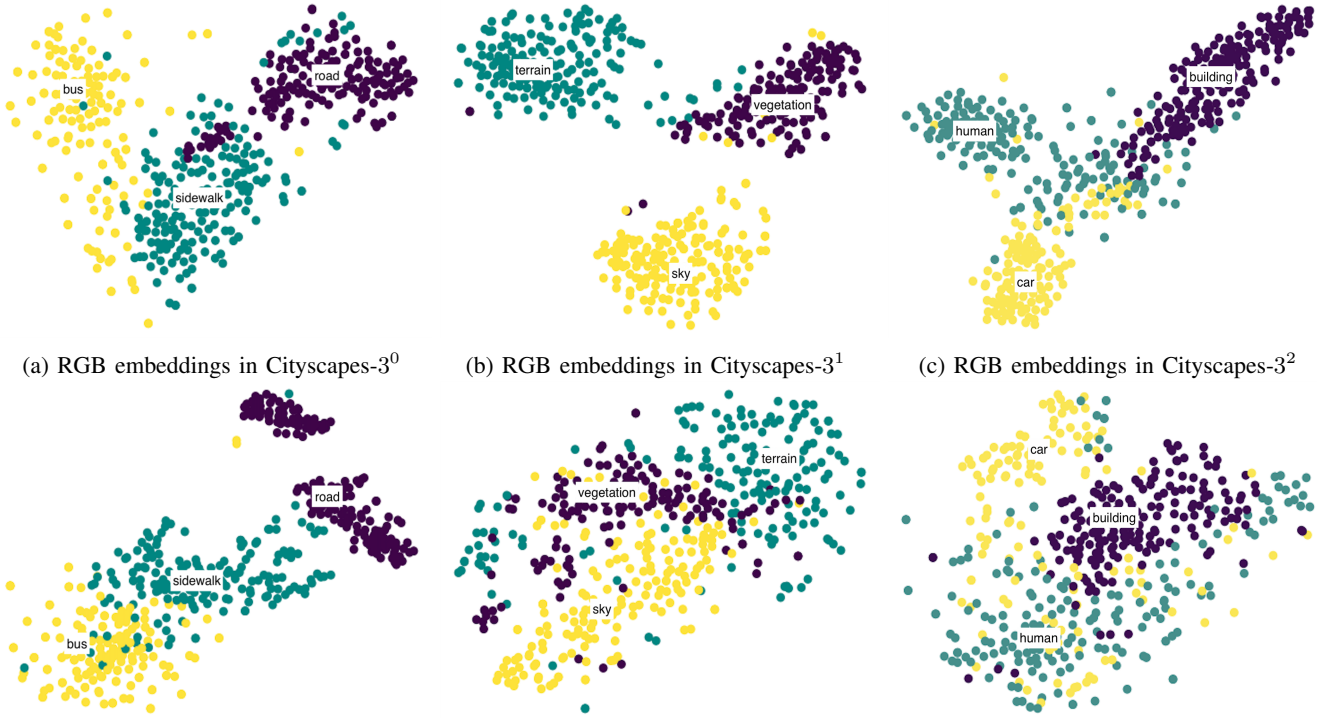
The main motivation of our work is to facilitate the few-shot segmentation task by incorporating complementary depth information. Existing supervised semantic segmentation approaches for RGB-D data do not offer a satisfactory solution to learn new categories rapidly from limited data. For this reason, we employ ideas from previous work of non-parametric metric learning and propose a two-stream deep neural network (RDNet). The main novelty of this study is to separately learn the RGB and depth prototype representations in different embedding spaces. The learned prototypes are applied to the corresponding query features as semantic guidance. Then we integrate the results from these two streams for an improved segmentation performance.

RGB-D input As shown in Figure 2, the proposed RDNet consists of two mirrored prototypical networks, which process RGB and depth input separately. Note that the support and query set through the depth stream provide the same scene information as the RGB Stream. Then the support images are embedded into high-level abstract features via a base network. For efficient implementation, we adopt a VGG-16 as the backbone network following the setup in [23]. In this way, we can map RGB and depth data into different embedding spaces.

Prototype learning for RGB-D data Snell et al. [15] proposed a prototypical network that learns a common metric space. Few-shot classification can be achieved by computing distances to prototype representations of each class. We employ the Masked Average Pooling [22] to build pre-class prototypes from both foreground and background information of the support images. Given a support set $D_s = (x_j^R, x_j^D, y(l)_j)_{j=1}^M$ (see Section III-A), let $F(l)_j'$ be the output feature maps of the base network with support RGB or depth input. Then $F(l)_j$ denotes the resized feature maps, which have the same width w and height h as the semantic mask $y(l)_j \in \{0,1\}^{W\times H}$. The prototype of target class l can then defined via Masked Average Pooling by the following equation:

$$p_c = \frac{1}{M}\sum_j \frac{\sum_{w=0,h=0}^{(w,h)} F(l)_j^{(w,h)} \mathbb{1}[y(l)_j^{(w,h)} = l]}{\sum_{w=0,h=0}^{w,h} \mathbb{1}[y(l)_j^{(w,h)} = l]} \quad (1)$$

where $\mathbb{1}(\cdot)$ is the indicator function that equals to 1 if the argument is true or 0 otherwise. Similarly, the prototypes for the background can be computed with $\mathbb{1}[y(l)_j^{(w,h)} \neq l]$. It is notable that both foreground and background information of RGD and depth images should be considered in this work.



(a) RGB embeddings in Cityscapes-3^0 (b) RGB embeddings in Cityscapes-3^1 (c) RGB embeddings in Cityscapes-3^2

(d) Depth embeddings in Cityscapes-3^0 (e) Depth embeddings in Cityscapes-3^1 (f) Depth embeddings in Cityscapes-3^2

Fig. 3: Visualization using t-SNE [24] for RGB and depth prototype representations in our RDNet.

These representative prototypes are the premise of reliable semantic guidance. To take an example, Figure 3 shows the visualization of RGB and depth prototype representations in our experiments.

Similarity guidance and feature fusion We compare the abstract query feature with expressive prototypes using distance metric learning method. To be specific, we map the query feature vector into the corresponding embedding space. The computed cosine distance indicates the similarity of target class. Besides, according to the previous work in fully-supervised semantic segmentation with RGB-D data [1, 11], there are two main fusion stategies, i.e., early fusion and late fusion [25, 26]. In our work, we employ the late fusion strategy, and concatenate all the probability maps generated from RGB and depth steams for a joint prediction.

IV. Dataset

To fully exploit few-shot semantic segmentation with additional depth information, we create a new dataset, named Cityscapes-3^i. We adopt the annotated RGB images and the depth images of the same scene from the Cityscapes dataset [27]. Cityscapes is a popular benchmark dataset for semantic understanding of outdoor scenes, which consists of thousands of precise depth images and pixel-wise semantic segmentation. Compared with object segmentation datasets such as PASCAL VOC [28] and COCO [29], it is more challenging to predict a pixel-wise mask for semantic classes in the image of Cityscapes. First, Cityscapes contains more complex urban street scenes. Images provide a broader perspective from the ground to the sky, involving a variety of categories. Then, most of the categories in the image have irregular shapes and lack distinct boundaries. Objects may overlap and be arranged randomly. Therefore it is a difficult task for segmentation models to learn characteristic features from only a few labeled samples and generalize to unseen classes.

TABLE I: Training and evaluation on Cityscapes-3^i dataset using 3-fold cross-validation, where i denotes the number of subsets.

Dataset	Test classes
Cityscapes-3^0	road, sidewalk, bus
Cityscapes-3^1	vegetation, terrain, sky
Cityscapes-3^2	human, car, building

We adopt all the RGB-D image pairs as well as the corresponding segmentation masks from Cityscapes training set for training, referred to as D_{train}. The test set D_{test} is formed by including all the samples in Cityscapes validation set. Then we choose 9 typical categories out of 30 as our target classes, containing *road*, *sidewalk*, *bus*, *vegetation*, *terrain*, *sky*, *human*, *car*, *building*. Following the setup of few-shot segmentation dataset PASCAL-5^i [17], we sample 3 classes out of all 9 categories as test label-set $L_{test} = \{3i+1, 3i+2, 3i+3\}$ where $i \in [1, 3]$ denotes the number of subsets, and the remaining 6 classes form the train label-set L_{train} (see Table I). Namely, $L_{train} \cap L_{test} = \emptyset$. The images in D_{train} and D_{test} contain at least one pixel in the semantic mask from the label-set

TABLE II: Results of 1-way 1-shot and 1-way 2-shot semantic segmentation on Cityscapes-5^i using mean-IoU(%) metric.

Methods	Modality	1-way 1-shot				1-way 2-shot			
		Cityscapes-3^0	Cityscapes-3^1	Cityscapes-3^2	Mean	Cityscapes-3^0	Cityscapes-3^1	Cityscapes-3^2	Mean
PANet	RGB	35.2	19.7	32.1	29.0	37.2	23.2	36.7	32.4
RDNet-R		35.7	22.3	32.6	30.2	36.7	24.1	37.5	32.8
PANet	Depth	32.6	14.5	19.3	22.1	34.2	15.8	22.5	24.2
RDNet-D		35.1	15.8	21.0	24.0	33.7	17.3	25.3	25.4
RDNet-concat	RGB-D	33.8	15.7	20.7	23.4	34.3	17.9	26.9	26.4
RDNet (ours)		**36.8**	**23.5**	**33.3**	**31.2**	**37.3**	**26.1**	**37.6**	**33.7**

TABLE III: Per-class mean-IoU(%) comparison of ablation studies for 1-way 1-shot semantic segmentation

Class	RDNet	RDNet-R	RDNet-D
Mean	**31.2**	30.2	24.0
Road	83.0	80.9	**84.4**
Sidewalk	**17.8**	15.7	15.7
Bus	9.5	**10.6**	5.3
Vegetation	**43.1**	40.2	26.9
Terrain	8.3	**10.1**	6.8
Sky	**19.1**	16.7	13.7
Human	**47.8**	46.6	36.9
Car	**12.1**	12.1	5.0
Building	**39.9**	39.2	21.1

TABLE IV: Results of 1-way 1-shot semantic segmentation using binary IoU and the runtime.

Mehtods	Modality	binary IoU	Runtime
PANet	RGB	55.0	71ms
RDNet-R		56.5	65ms
RDNet-concat	RGB-D	51.9	67ms
RDNet (ours)		57.9	135ms

L_{train} and L_{test}, respectively. Moreover, we reset the pixels in segmentation masks that not belong to the corresponding label-sets as the background. In our experiments, we train and evaluate the proposed model on 3 folders in a cross-validation manner. For each folder, we take a random 500 samples and average the results from 5 runs to evaluate the performance of the models.

V. Experiments

A. Setup

a) Implementation details: We conduct the experiments with implementations in PyTorch [30]. The backbone network (i.e., VGG-16) was initialized with pre-trained weights on ImageNet [31]. We resized the input images to 768×384 and trained on a single Nvidia TITAN Xp GPU with 12GB memory. All the few-shot segmentation models were trained using stochastic gradient descent (SGD) with a batch size of 1, a momentum of 0.9, and weight decay of 0.0005 for a maximum of 30,000 iterations. The initial learning rate was set to 0.0001 and reduced by 0.1 every 10,000 iterations.

b) Evaluation metrics: Following the previous works on few-shot segmentation [17][22][23], we apply two standard metrics to evaluate the performance of learning models: mean-IoU and binary-IoU. Generally, the mean Intersection-over-Union (mean-IoU) is used to measure the accuracy of each foreground class and average over all the classes. Binary-IoU deals uniformly with all object categories as one foreground class and averages the IoU of both foreground and background. Based on these two metrics, we can fairly compare the accuracy and efficiency of baselines in terms of 1-way N-shot semantic segmentation.

c) Baselines: First we employ PANet [23] as the unimodal baseline model. Arguably, this baseline network shows significant performances on PASCAL-5^i dataset. We report its evaluations on both RGB and depth data. Moreover, we present the performance of RGB stream and depth stream of our model separately, performing a series of ablation studies. Furthermore, we set a multimodal baseline with simple concatenation, referred to as RDNet-concat. In this baseline, we concatenate the RGB and depth images to 4 channels at the early stage. An extra fusion layer (3×3 convolutional filters and ReLU activation) was added to adapt the concatenated inputs, while the rest of the network is exactly the same as the RGB branch in our model.

B. Experimental results

In Table II, we illustrate the performance of our proposed RDNet and other baseline methods on Cityscapes-3^i, including 1-way 1-shot and 1-way 2-shot semantic segmentation. First, we observe that using RGB data provides better segmentation results than using depth data as input. Moreover, one can also notice that a simple concatenation of RGB and depth features, RDNet-concat, does not provide satisfactory results. Indeed, RDNet-concat achieves a mIoU score of 26.4%, which is higher than the score obtained with RDNet-D (25.4%) but much lower than the score obtained by RDNet-R (32.8%) for 1-way 2-shot semantic segmentation. Our method, RDNet, outperforms other unimodal networks and concatenated approach overall. RDNet achieves a mIoU score of 31.2% for 1-way 1-shot and 33.7% for 1-way 2-shot, which represents an increase of +7% compared to RDNet-concat.

We further conduct ablation studies to investigate the validity of RDNet. The results are shown in Table III. We can observe a satisfactory performance enhancement of our method for most of the classes. In particular, the *vegetation*, *sidewalk* and *sky* classes. These experimental results illustrate the effectiveness of our method and the potential of depth information in scene understanding with limited supervision.

Compared with RDNet-concat, our proposed method provides an improvement of +7.8% and +7.3% in terms of mIoU and binary IoU for 1-way 1-shot segmentation (see Table IV). The results also show that simple concatenation has no significant improvement in the segmenatation prediction. Besides, Figure 4 shows the qualitative results of our method,

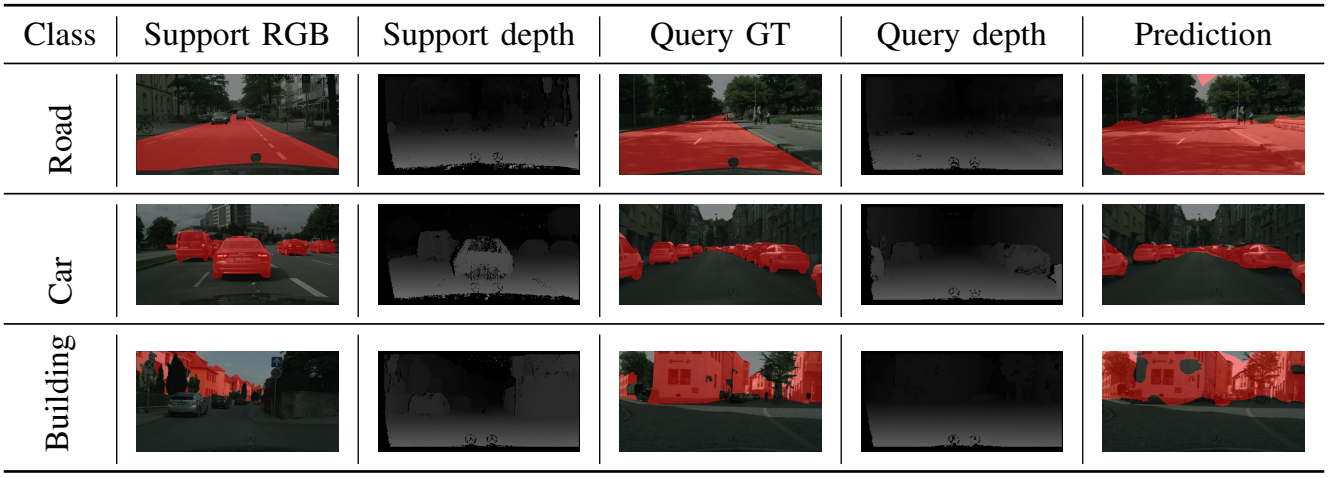


Fig. 4: Qualitative results of our method for 1-way 1-shot semantic segmentation on Cityscapes-3^i.

including multimodal input and the segmentation prediction. Our model yields promising segmentation results in 1-shot settings. However, it is still challenging to distinguish the irregular objects and categories with similar characteristics in the complex scenes, such as *car* and *bus*.

C. Comparison of visualized features

To clearly demonstrate the generalization and discrimination of the proposed model, we visualize the prototype representations of target classes in the RGB and depth embedding space using t-SNE (see Figure 3). Each figure was generated using 500 samples of test classes in Cityscapes-3^i. On the whole, the prototypes generated from support RGB input can be well separated, especially for *vegetation*, *terrain*, *sky* in Figure 3b. Although it is challenging to produce distinctive prototypes in the depth embedding space, these prototype representations provide complementary cues regarding depth information. For example, the depth embeddings in Cityscapes-3^0 clearly show the discrimination on the classes *vegetation*, *terrain* and *sky* (see Figure 3d). Consequently, the generalizability of our few-shot segmentation network gets improved by incorporating supplementary depth information, leading to more promising prediction results.

VI. Conclusion

We proposed a few-shot semantic segmentation model with complementary depth information, which consists of two mirrored streams based on metric learning. To fully take advantage of color and geometric information of the scenes, we mapped the representative features of target classes into different embedding spaces. The learned prototype represenations provide effective semantic guidance on the corresponding query feature. Then we integrated the generated probability maps at a late stage. Comprehensive experiments and ablation studies on Cityscapes-3^i dataset demonstrate the improved generalizability and discriminating ability of our method. The proposed RDNet is simple yet effective, and explore the use of depth information in few-shot segmentation task. Our future work will focus on the impact of multimodal data in few-shot learning tasks and how to fuse these data for optimal performance.

Acknowledgment

The authors would like to acknowledge the French ANR project ICUB (ANR-17-CE22-0011) for its financial support as well as a hardware grant from NVIDIA.

References

[1] 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*. Springer, 2016, pp. 213–228.

[2] L. Ma, J. Stückler, C. Kerl, and D. Cremers, "Multi-view deep learning for consistent semantic mapping with rgb-d cameras," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 598–605.

[3] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[4] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in neural information processing systems*, 2017, pp. 5099–5108.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[6] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[7] 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.

[8] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3d graph neural networks for rgbd semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5199–5208.

[9] Y. Li, J. Zhang, Y. Cheng, K. Huang, and T. Tan, "Semantics-guided multi-level rgb-d feature fusion for indoor semantic segmentation," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 1262–1266.

[10] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3029–3037.

[11] L. Deng, M. Yang, T. Li, Y. He, and C. Wang, "Rf-bnet: deep multimodal networks with residual fusion blocks for rgb-d semantic segmentation," *arXiv preprint arXiv:1907.00135*, 2019.

[12] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *European conference on computer vision*. Springer, 2014, pp. 345–360.

[13] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *Advances in neural information processing systems*, 2016, pp. 3630–3638.

[14] X. Li, L. Yu, C.-W. Fu, M. Fang, and P.-A. Heng, "Revisiting metric learning for few-shot image classification," *ArXiv*, vol. abs/1907.03123, 2019.

[15] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in neural information processing systems*, 2017, pp. 4077–4087.

[16] V. Garcia and J. Bruna, "Few-shot learning with graph neural networks," *arXiv preprint arXiv:1711.04043*, 2017.

[17] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," *arXiv preprint arXiv:1709.03410*, 2017.

[18] K. Rakelly, E. Shelhamer, T. Darrell, A. Efros, and S. Levine, "Conditional networks for few-shot semantic segmentation," 2018.

[19] N. Dong and E. Xing, "Few-shot semantic segmentation with prototype learning." in *BMVC*, vol. 3, no. 4, 2018.

[20] Z. Dong, R. Zhang, X. Shao, and H. Zhou, "Multi-scale discriminative location-aware network for few-shot semantic segmentation," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 2. IEEE, 2019, pp. 42–47.

[21] M. Siam, B. Oreshkin, and M. Jagersand, "Adaptive masked proxies for few-shot segmentation," *arXiv preprint arXiv:1902.11123*, 2019.

[22] X. Zhang, Y. Wei, Y. Yang, and T. Huang, "Sg-one: Similarity guidance network for one-shot semantic segmentation," *arXiv preprint arXiv:1810.09091*, 2018.

[23] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9197–9206.

[24] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html

[25] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[26] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[27] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[28] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.