frontiers | Frontiers in Computer Science

TYPE Brief Research Report
PUBLISHED 22 November 2022
DOI 10.3389/fcomp.2022.1062792

OPEN ACCESS

EDITED BY
Mingming Gong,
The University of Melbourne, Australia

REVIEWED BY
Haopeng Li,
University of Melbourne, Australia
Zhaoqing Wang,
The University of Sydney, Australia
Dongting Hu,
The University of Melbourne, Australia

*CORRESPONDENCE
Mehdi Khoshboresh-Masouleh
m.khoshboresh@ut.ac.ir
Reza Shah-Hosseini
rshahosseini@ut.ac.ir

SPECIALTY SECTION
This article was submitted to Computer Vision, a section of the journal Frontiers in Computer Science

RECEIVED 06 October 2022
ACCEPTED 09 November 2022
PUBLISHED 22 November 2022

CITATION
Khoshboresh-Masouleh M and Shah-Hosseini R (2022) Real-time multiple target segmentation with multimodal few-shot learning. *Front. Comput. Sci.* 4:1062792. doi: 10.3389/fcomp.2022.1062792

Real-time multiple target segmentation with multimodal few-shot learning

Mehdi Khoshboresh-Masouleh* and Reza Shah-Hosseini*

School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran

Deep learning-based target segmentation requires a big training dataset to achieve good results. In this regard, few-shot learning a model that quickly adapts to new targets with a few labeled support samples is proposed to tackle this issue. In this study, we introduce a new multimodal few-shot learning [e.g., red-green-blue (RGB), thermal, and depth] for real-time multiple target segmentation in a real-world application with a few examples based on a new squeeze-and-attentions mechanism for multiscale and multiple target segmentation. Compared to the state-of-the-art methods (HSNet, CANet, and PFENet), the proposed method demonstrates significantly better performance on the PST900 dataset with 32 time-series sets in both Hand-Drill, and Survivor classes.

KEYWORDS

few-shot learning, multimodal images, target detection, real-time processing, squeeze-and-attention CNN

Introduction

Real-time multiple target segmentation is an important task for real-world applications (Morelande et al., 2007; Wagner et al., 2009), such as search and rescue robots. In this regard, the quadruped mobile robot based on multimodal images can provide more comprehensive spatial and spectral information for the development of real-time multiple target segmentation (Rahman et al., 2021). Multimodal target segmentation is challenging due to the various background, shadows, and occluded areas, and multiscale targets. The goal of multimodal imaging is to improve detection and localization of objects in complex scenes (Martí-Bonmatí et al., 2010). In multimodality imaging, the need to combine different information can be approached by either acquiring images at different times. In this regard, the image fusion is the process of merging data from multiple imaging modalities [e.g., red-green-blue (RGB), thermal, and depth] to obtain a fused image with a large amount of information for increasing the scene understanding applicability. Multiple target segmentation with the use of multimodal data for time-series images potentially improves scene understanding with a limited amount of labeled training data, while many target segmentation methods appear to understand single-time localization with a big training dataset. Multimodal few-shot learning can perform on unseen tasks after training a few annotated data and considers several tasks to produce a predictive function, and is an inductive transfer system whose main goal is to improve generalization ability for multiple targets. These approaches excel at learning complicated features from small set using weakly-supervised learning.

Deep learning has been successful in target segmentation (Dimou et al., 2016). But the major bottleneck of deep learning in target segmentation is the need for large-scale labeled datasets for training, particularly in multimodal data (Yao et al., 2017). Owing to the advances of deep learning networks, new insights have been presented in the field of multimodal processing for time-series images. In Shivakumar et al. (2020), a camera calibration method and a dual-stream CNN architecture was applied to multimodal image segmentation that is able to fuse RGB with thermal information. The quantitative assessments of this study for PST900 dataset show that the mean intersection over union (mIoU) is about 68% for RGB-Thermal mode. A squeeze-and-attention network (Zhong et al., 2020) is proposed for RGB image segmentation based on pixel-group attention, and pixel-wise prediction, which achieves 83.2% mIoU. Moreover, an Edge-Aware Guidance Fusion Network (EGFNet) was introduces in Zhou et al. (2022) for multimodal scene parsing and segmentation. This study used only RGB and thermal modality for scene understanding. The quantitative results of this work show that the mean accuracies for FuseSeg-161 (Sun et al., 2021), and the proposed method are about 62.1%, and 74.4%, respectively. FuseSeg-161 is a multimodal data fusion with RGB and thermal images to achieve superior performance of semantic segmentation in urban scenes.

Researchers have studied multimodal data from time-series images, with deep learning approaches as the preferred choice. Although some efforts have been devoted to the development of scene understanding with a large training dataset from thermal, and RGB images, little attention has been devoted to multimodal target detection for different specific target of interest with a small training dataset based on depth, thermal, and RGB images. Most of the current few-shot learning methods use single-modal sensory data, which are usually the RGB images produced by visible cameras. However, the target segmentation performance of these networks is prone to be degraded when lighting conditions are not satisfied, such as dim light or darkness. Moreover, we can improve the accuracy by overcoming the segmentation challenges such as dim light or darkness by thermal information. Thermal image is invariant to lighting variations and affords the ability to take advantage of spectral separation between objects. To the best of the authors' knowledge, although the related deep learning methods are fairly powerful for image segmentation from multimodal image segmentation with a large training dataset, there is not still outstanding performance for multiple target segmentation with a small training dataset. A robust target segmentation method not only needs a strong model architecture and learning algorithms but also relies on a comprehensive large-scale training set (Zheng, 2022). Generating and annotating such multimodal datasets can be labor-intensive and costly (Bauer et al., 2021). In real-world applications, only a few labeled datasets may be available at model training time. As a solution, few-shot learning aims to build accurate trained models with less training data, contrary to the general experiment of using a large amount of data (Wang et al., 2019; Feyjie et al., 2020).

In this study, we propose a new multimodal few-shot learning for real-time multiple target segmentation from a small labeled multimodal dataset, including RGB, thermal, and depth images for a search and rescue scenario.

Proposed method

For multimodal few-shot learning, we define three datasets where each set contains N multimodal images, including: a training dataset S_{train} with semantic classes y_c, for training step $S_{train} = \{x_i, y_i\}_{i=1}^{N_{train}}$, with $D \subset \mathbb{R}^3$ a composite RGB, depth, and thermal image (RGBDT) space, $x_i : D \rightarrow \mathbb{R}^3$ an input RGBDT image, and $y_i : D \rightarrow \{0, 1\}^{|y_c|}$ its corresponding binary mask, a support dataset $S_{support} = \{x_i, y_i\}_{i=1}^{N_{support}}$, and a test dataset $S_{test} = \{x_i\}_{i=\ 1}^{N_{test}}$.

The proposed multimodal few-shot learning method aims at training a new squeeze-and attention CNN $\varphi(\varepsilon, \theta)$ on the time-series training set to have the capability to extract a new target tS_{train} on the time-series test set based on t references from $S_{support}$. The proposed squeeze-and-attention mechanism Y_{sa} is defined as follows:

$$\begin{aligned} A_i &= f_{attn}(P\,(x_i)\,;\Theta_{attn}, \Omega_{attn}) \\ &= \sum\nolimits_{i=1}^{5} \sum\nolimits_{j=1}^{H-1} \sum\nolimits_{k=1}^{W-1} (P\,(x_i) \times Conv1_j\,(\Theta_{attn}, \Omega_{attn})) \\ &\quad \times Conv1_k(\Theta_{attn}^{'}, \Omega_{attn}^{'}) \qquad (1) \\ Y_{sa} &= \mho(\sigma\,(A_i)) \bigotimes (x_{res} + 1) \qquad (2) \end{aligned}$$

where A_i is a function to calculate the attention maps given the input feature maps $P\,(x_i)$. $P\,(x_i)$ is a median pooling layer for input feature map. f_{attn} is an attention function emphasizes the attention of pixel groups that belong to the same classes at different spatial scales (Zhong et al., 2020), which is parameterized by Θ_{attn} and Ω_{attn}. Θ_{attn} and Ω_{attn} represent the weights and biases from two stacked convolutional layers is added to output map. Moreover, $\mho(.)$ is an upsampling function for expanding the result of the attention channel, σ is a relu function, and x_{res} is a residual feature map with an element-wise multiplication $\bigotimes$ (Khoshboresh-Masouleh and Shah-Hosseini, 2021).

The pipeline illustration of the proposed method is shown in Figure 1. The proposed method has three components, including (a) learns the representation from time-series images pair based on Equation (1) from training dataset with two encoding blocks for support and query images, (b) for getting prior knowledge from a few support samples, a new class registration (CR) network for few support samples with new target class is designed based on dilation convolution layers, and (c) a multiscale decoder network by a transposed convolution

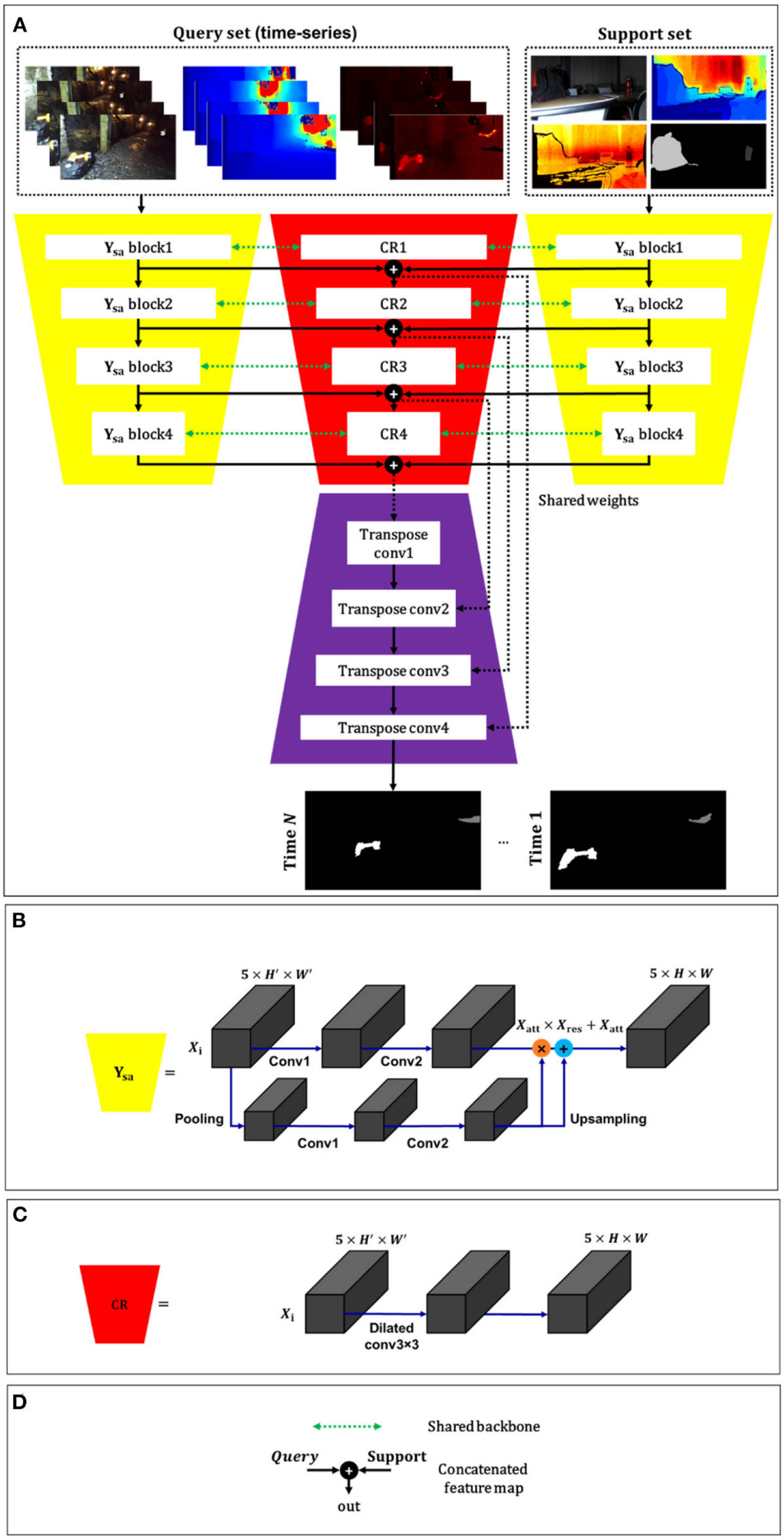

FIGURE 1
Overview of our method for multimodal few-shot learning. **(A)** Proposed architecture for feature extraction and real-time multiple target segmentation, **(B)** proposed squeeze-and-attention mechanism, **(C)** proposed class registration (CR) network, and **(D)** legend.

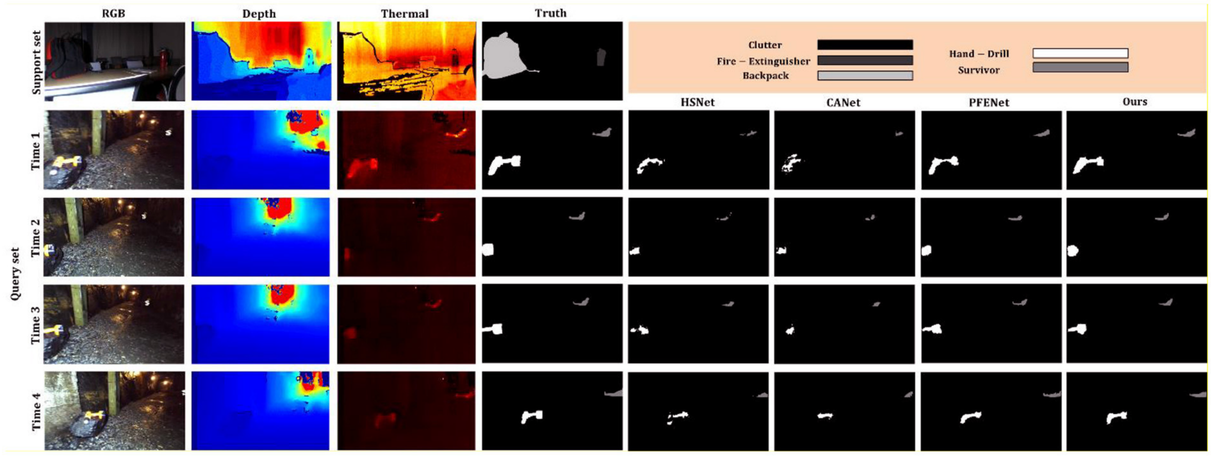

FIGURE 2
Qualitative results (one-shot) of HSNet, CANet, PFENet, and our method on PST900's time-series images.

TABLE 1 Hand-Drill class mIoU and inference time (in ms) results on PST900 for multimodal few-shot learning.

Methods	Backbone	1 shot	5 shot	10 shot	ms[b]
HSNet[a]	ResNet50	48.6	53.2	59.4	40
CANet[a]		38.5	43.3	51.6	51
PFENet[a]		**58.9**	62.1	67.3	64
Ours	ResNet50	58.1	**65.4**	**78.7**	**35**

Best results in bold and the underlined font denotes the second-best result.

[a] This model was revised for multimodal training based on the proposed method for multiple target detection.

[b] On an NVIDIA Tesla K80.

TABLE 2 Survivor class mIoU and inference time (in ms) results on PST900 for multimodal few-shot learning.

Methods	Backbone	1 shot	5 shot	10 shot	ms[b]
HSNet[a]	ResNet50	44.8	47.1	51.4	42
CANet[a]		32.2	33.6	37.5	54
PFENet[a]		59.4	52.3	55.7	65
Ours	ResNet50	**60.3**	**61.1**	**68.4**	**38**

Best results in bold and the underlined font denotes the second-best result.

[a] This model was revised for multimodal training based on the proposed method for multiple target detection.

[b] On an NVIDIA Tesla K80.

with the stride of two for generating the final target detection mask from the test dataset.

As depicted in Figure 1, the convolutional blocks include 3 × 3 kernels which are applied to the input feature maps using stride one. The proposed architecture is composed of eight squeeze-and-attention layers followed by batch normalization, and rectified linear unit functions to generate feature maps, as well as max-pooling blocks to reduce the size of feature maps. We use multiscale fusion to take advantage of its good ability for multimodal data fusion. In the proposed squeeze-and-attention mechanism (Figure 1B), the multiscale multimodal fusion is performed by three multiscale dilated convolution layers and an element-wise summation of feature layers from the direct path for RGB, thermal, and depth data, resulting in better target boundaries.

Dataset

We evaluate our model on the PST900 dataset (Shivakumar et al., 2020) includes 256 images for test set, 128 images for support dataset, and 352 for training set. PST900 is built from synchronized and calibrated RGBDT time-series images with a size of 1,280 × 720 pixels for real-time target segmentation, and contains five target categories. In action, three classes, including Fire-Extinguisher, Backpack, and clutter are used for training, and the remaining two categories, including Hand-Drill, and Survivor for testing. In this study, the PST900 dataset is single-fold with three training classes and two test classes. It consists of 32 time-series sets, where each class contains about 30–80 images with their corresponding pixel-level ground truth annotations.

TABLE 3 Ablation study on using different modalities, loss functions, and backbones.

Feature extractor	Backbone	Loss	Modalities	mIoU[a]
Proposed layer $\mathbf{Y}_{\mathbf{sa}}$ (10 shot)	ResNet50	D	RGB	49.83
	HRNet			49.01
	ResNet50	BC		50.46
	HRNet			48.64
	ResNet50	M2CE		**52.34**
	HRNet			49.72
	ResNet50	D	RGB+Thermal	59.63
	HRNet			57.24
	ResNet50	BC		59.43
	HRNet			58.90
	ResNet50	M2CE		**60.04**
	HRNet			59.74
	ResNet50	D	RGB+Depth+ Thermal	69.17
	HRNet			67.12
	ResNet50	BC		70.20
	HRNet			68.34
	ResNet50	M2CE		**73.55**
	HRNet			68.46

Best results in bold.

[a] Mean intersection over union for Hand-Drill, and Survivor classes.

Implementation details

We train the proposed network in PyTorch with multi-class cross-entropy over the training class during 120 epochs on PST900 with batch size set to 5, and use Adam as optimizer with the initial learning rate set to 10^{-4}. β_1, and β_2 are set to 0.9, and 0.999 and weight decay to $10^{-\ 8}$.

Evaluation protocol

In our work, the evaluation protocol used in most works in few-shot semantic segmentation is employed (Wang H. et al., 2020). The intersection-over-union (IoU) is the standard metric used in evaluating pixel-wise target segmentation. Given two ground truth (g) and predicted segment (p) masks, the IoU can be defined as $IoU = \frac{|p \cap g|}{|p \cup g\ |}$.

Results

To investigate the behavior of the multimodal few-shot learning, we investigate the components of the proposed method PST900 where the training and test classes are required to be simultaneously identified. Although the related models are applied to few-shot learning, there is not still a good method for multimodal few-shot learning for real-time multiple target segmentation. For a fair comparison, all state-of-the-art models were trained from the beginning, using the same training set and feature extractor that was applied for the training of the proposed model. We compare our model against relevant methods HSNet (Min et al., 2021), CANet (Zhang et al., 2019), and PFENet (Tian et al., 2022). Qualitative results of one-shot multiple target detection for HSNet, CANet, PFENet, and our method are shown in Figure 2. We provide some qualitative results on multimodal time-series images that show how our model helps refine the real-time multiple target detection. Note that the proposed method and PFENet can effectively remove irrelevant boundaries and fill the target region. Tables 1, 2 show the performance of the proposed method in comparison to the other state-of-the-art methods on the PST900 for Hand-Drill and Survivor classes.

According to the results, the function of PFENet was found to be better than HSNet and CANet. But, many of the target pixels were not detected in both 5-shot and 10-shot scenarios. We observe that PFENet outperforms the proposed method in the 1-shot scenario for Hand-Drill class. The proposed method achieves a mIoU for the Hand-Drill class of 58.1, 65.4, and 78.7, while the mIoUs for the Survivor class are 60.3, 61.1, and 68.4 in 1-shot, 5-shot, and 10-shot scenarios.

Ablation study

In this section, we present an ablation study to compare a number of different model variants, such as different modalities (e.g., RGB, thermal, and depth), loss functions [Bootstrapped Cross-entropy (BC), Dice, and multi-class cross-entropy], and backbones (e.g., ResNet50, and HRNet), and justify our design choices. Table 3 shows some ablation study results to investigate the behavior of the proposed model.

Visualization of ablation study results of the example set for the highest accuracy for each modality among the compared methods is shown in Figure 3. For each test result with RGB, RGB-Thermal, and RGB+Depth+Thermal, the proposed method is effective in target segmentation with RGB+Depth+Thermal modality. Moreover, the proposed method obtained better target segmentation results than other models with RGB or RGB-Thermal modalities.

Modalities

The proposed model for real-time multiple target segmentation was tested with different data modalities, such as RGB, thermal, and depth images.

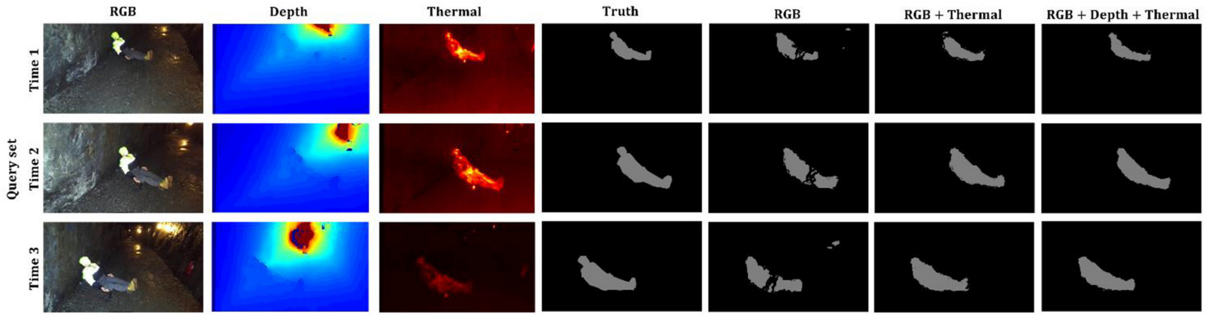

FIGURE 3
The visualization of real-time target segmentation output through each modality with ResNet50 backbone and M2CE function.

Loss functions

All networks were additionally evaluated with different loss functions. Although the proposed model with multi-class cross-entropy (M2CE) function delivers good results, the proposed model was tested with another two loss functions, consist of Dice (D) (Sudre et al., 2017) and Bootstrapped Cross-entropy (BC) (Gaj et al., 2021). In this regard, we train our model using the absolute error between the ground truth map and the model's predicted.

Backbones

The proposed model can be set up with different backbones for real-time multiple target segmentation. We selected two backbones for ablation study. The experiments were carried out with the ResNet50, and HRNet (Wang J. et al., 2020) backbones.

Conclusion

We have presented a new multimodal few-shot learning for real-time multiple target detection in a real-world application. Different from the previous few-shot learning methods, the proposed method aims at accurate and fast identifying new targets from time-series images. Our method has outstanding results on a challenging dataset PST900 with 32 time-series sets, compared with recent prominent models in few-shot learning.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

MK-M and RS-H conceived the original idea for this manuscript and discussed the results. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bauer, D. F., Russ, T., Waldkirch, B. I., Tönnes, C., Segars, W. P., Schad, L. R., et al. (2021). Generation of annotated multimodal ground truth datasets for abdominal medical image registration. *Int. J. Comput. Assist. Radiol. Surg.* 16, 1277–1285. doi: 10.1007/s11548-021-02372-7

Dimou, A., Medentzidou, P., García, F. Á., and Daras, P. (2016). "Multi-target detection in CCTV footage for tracking applications using deep learning techniques," in *2016 IEEE International Conference on Image Processing (ICIP)* (Phoenix, AZ), 928–932. doi: 10.1109/ICIP.2016.7532493

Feyjie, A. R., Azad, R., Pedersoli, M., Kauffman, C., Ayed, I. B., and Dolz, J. (2020). Semi-supervised few-shot learning for medical image segmentation. *ArXiv200308462 Cs*. Available online at: http://arxiv.org/abs/2003.08462 (accessed March 24, 2022).

Gaj, S., Ontaneda, D., and Nakamura, K. (2021). Automatic segmentation of gadolinium-enhancing lesions in multiple sclerosis using deep learning from clinical MRI. *PLoS ONE* 16:e0255939. doi: 10.1371/journal.pone.0255939

Khoshboresh-Masouleh, M., and Shah-Hosseini, R. (2021). Building panoptic change segmentation with the use of uncertainty estimation in squeeze-and-attention CNN and remote sensing observations. *Int. J. Remote Sens.* 42, 7798–7820. doi: 10.1080/01431161.2021.1966853

Martí-Bonmatí, L., Sopena, R., Bartumeus, P., and Sopena, P. (2010). Multimodality imaging techniques. *Contrast Media Mol. Imaging* 5, 180–189. doi: 10.1002/cmmi.393

Min, J., Kang, D., and Cho, M. (2021). "Hypercorrelation squeeze for few-shot segmentation," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC), 6941–6952. Available online at: https://openaccess.thecvf.com/content/ICCV2021/html/Min_Hypercorrelation_Squeeze_for_Few-Shot_Segmentation_ICCV_2021_paper.html (accessed March 24, 2022).

Morelande, M. R., Kreucher, C. M., and Kastella, K. (2007). A Bayesian approach to multiple target detection and tracking. *IEEE Trans. Signal Process.* 55, 1589–1604. doi: 10.1109/TSP.2006.889470

Rahman, M. M., Rahman, T., Kim, D., and Alam, M. A. U. (2021). "Knowledge transfer across imaging modalities via simultaneous learning of adaptive autoencoders for high-fidelity mobile robot vision," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Prague), 1267–1273. doi: 10.1109/IROS51168.2021.9636360

Shivakumar, S. S., Rodrigues, N., Zhou, A., Miller, I. D., Kumar, V., and Taylor, C. J. (2020). "PST900: RGB-thermal calibration, dataset and segmentation network," in *2020 IEEE International Conference on Robotics and Automation (ICRA)* (Paris), 9441–9447. doi: 10.1109/ICRA40945.2020.9196831

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Cardoso, M. J. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *ArXiv:170703237* Cs 10553, 240–248. doi: 10.1007/978-3-319-67558-9_28

Sun, Y., Zuo, W., Yun, P., Wang, H., and Liu, M. (2021). FuseSeg: semantic segmentation of urban scenes based on RGB and thermal data fusion. *IEEE Trans. Autom. Sci. Eng.* 18, 1000–1011. doi: 10.1109/TASE.2020.2993143

Tian, Z., Zhao, H., Shu, M., Yang, Z., Li, R., and Jia, J. (2022). Prior guided feature enrichment network for few-shot segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 1050–1065. doi: 10.1109/TPAMI.2020.3013717

Wagner, D., Schmalstieg, D., and Bischof, H. (2009). "Multiple target detection and tracking with guaranteed framerates on mobile phones," in *2009 8th IEEE International Symposium on Mixed and Augmented Reality* (Washington, DC), 57–64. doi: 10.1109/ISMAR.2009.5336497

Wang, H., Zhang, X., Hu, Y., Yang, Y., Cao, X., and Zhen, X. (2020). "Few-shot semantic segmentation with democratic attention networks," in *Computer Vision – ECCV 2020 Lecture Notes in Computer Science*, eds A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm (Cham: Springer International Publishing), 730–746. doi: 10.1007/978-3-030-58601-0_43

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., et al. (2020). Deep high-resolution representation learning for visual recognition. *ArXiv190807919 Cs*. Available online at: http://arxiv.org/abs/1908.07919 (accessed May 3, 2022).

Wang, K., Liew, J. H., Zou, Y., Zhou, D., and Feng, J. (2019). "PANet: few-shot image semantic segmentation with prototype alignment," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul), 9197–9206. Available online at: https://openaccess.thecvf.com/content_ICCV_2019/html/Wang_PANet_Few-Shot_Image_Semantic_Segmentation_With_Prototype_Alignment_ICCV_2019_paper.html [Accessed March 24, 2022].

Yao, H., Yu, Q., Xing, X., He, F., and Ma, J. (2017). "Deep-learning-based moving target detection for unmanned air vehicles," in *2017 36th Chinese Control Conference (CCC)* (Dalian), 11459–11463. doi: 10.23919/ChiCC.2017.8029186

Zhang, C., Lin, G., Liu, F., Yao, R., and Shen, C. (2019). "CANet: class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA), 5217–5226. Available at: https://openaccess.thecvf.com/content_CVPR_2019/html/Zhang_CANet_Class-Agnostic_Segmentation_Networks_With_Iterative_Refinement_and_Attentive_Few-Shot_CVPR_2019_paper.html (accessed March 24, 2022).

Zheng, L. (2022). *The 1st Workshop on Vision Datasets Understanding - CVPR 2022*. Available online at: https://sites.google.com/view/vdu-cvpr22 (accessed March 24, 2022).

Zhong, Z., Lin, Z. Q., Bidart, R., Hu, X., Daya, I. B., Li, Z., et al. (2020). "Squeeze-and-attention networks for semantic segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA), 13062–13071. doi: 10.1109/CVPR42600.2020.01308

Zhou, W., Dong, S., Xu, C., and Qian, Y. (2022). Edge-aware guidance fusion network for RGB–thermal scene parsing. *Proc. AAAI Conf. Artif. Intell.* 36, 3571–3579. doi: 10.1609/aaai.v36i3.20269