

Reducing Annotation Burden: Exploiting Image Knowledge for Few-Shot Medical Video Object Segmentation via Spatiotemporal Consistency Relearning

Zixuan Zheng^{1*}, Yilei Shi^{1*}, Chunlei Li¹, Jingliang Hu¹, Xiao Xiang Zhu², and Lichao Mou¹✉

¹ MedAI Technology (Wuxi) Co. Ltd., Wuxi, China
lichao.mou@medimagingai.com

² Technical University of Munich, Munich, Germany

Abstract. Few-shot video object segmentation aims to reduce annotation costs; however, existing methods still require abundant dense frame annotations for training, which are scarce in the medical domain. We investigate an extremely low-data regime that utilizes annotations from only a few video frames and leverages existing labeled images to minimize costly video annotations. Specifically, we propose a two-phase framework. First, we learn a few-shot segmentation model using labeled images. Subsequently, to improve performance without full supervision, we introduce a spatiotemporal consistency relearning approach on medical videos that enforces consistency between consecutive frames. Constraints are also enforced between the image model and relearning model at both feature and prediction levels. Experiments demonstrate the superiority of our approach over state-of-the-art few-shot segmentation methods. Our model bridges the gap between abundant annotated medical images and scarce, sparsely labeled medical videos to achieve strong video segmentation performance in this low data regime. Code is available at <https://github.com/MedAITech/RAB>.

Keywords: medical video segmentation · few-shot learning · relearning.

1 Introduction

The emergence of deep learning has propelled the advancement of numerous medical data analysis tasks, particularly achieving significant accomplishments in segmentation tasks. The automatic segmentation of medical data plays a pivotal role in various clinical applications such as computer-aided diagnosis and disease progression monitoring [1,2]. However, in many practical scenarios, segmentation models often face challenges due to the prohibitive cost of pixel-level annotation from clinical experts and the limited availability of samples for rare anomalies and unusual pathological conditions. To circumvent these constraints,

* Equal contribution.

there has been a growing interest in few-shot segmentation techniques that can learn to segment new classes with only a few annotated examples [3].

Prior few-shot semantic segmentation methods in computer vision can be categorized into three main approaches: conditional networks, latent space optimization, and prototypical learning. Conditional networks commonly comprise two modules—a conditioning branch that ingests support samples to produce model parameters, and a segmentation branch that predicts masks for query images based on these dynamic parameters [4,18]. Latent space optimization methods instead leverage generative models to synthesize images by optimizing latent vectors to match the distribution of queries. Segmentation is then performed by transferring representations from the synthesized examples [5]. Finally, prototypical networks derive class prototypes from support images which are then compared against features of query images to predict segmentation masks [6,27]. In the medical domain, prototypical networks have garnered substantial traction for few-shot medical image segmentation [7,8].

From image to video, [9] and [10] explore segmenting objects of novel categories in query videos using only a few annotated support frames. However, these approaches still require densely annotated frames for training and do not fully alleviate annotation costs. In addition, since videos are often not recorded and stored in medical imaging applications, publicly available medical video datasets remain relatively scarce and limited in scale, posing challenges for medical video object segmentation tasks.

In this paper, we introduce a novel task of few-shot medical video object segmentation with image datasets. The goal is to segment a medical video given sparse annotations—only a few frames (e.g. the first) have ground truth masks, along with abundant labeled images. This represents an extremely low-data regime for video segmentation that minimizes costly pixel-level video annotations. To address this, we propose a two-phase framework. First, we learn a few-shot segmentation model using labeled images. Specifically, a pre-trained backbone extracts features from support and query images. These features and support masks are integrated to generate coarse query masks that approximately localize target objects. The coarse masks are then fused with support and query features to output fine segmentation masks. However, without temporal modeling, the few-shot image segmentation model may perform sub-optimally on videos. Thus, we propose a spatiotemporal consistency relearning approach for medical videos that is performed in the second phase. We leverage the temporal continuity prior, which assumes consistency between consecutive frames in a video, to obtain effective spatiotemporal features. Furthermore, we introduce spatial consistency constraints between the model trained in the first phase and the relearning model at both feature and prediction levels. These effectively minimize discrepancies between the two models to ensure model performance on videos. Our key contributions are three-fold:

- We explore few-shot medical video object segmentation using image datasets, without reliance on full video annotations.

- We introduce a novel self-supervised framework that exploits spatiotemporal consistency as additional supervision to achieve improved segmentations in this challenging setting.
- Extensive experiments demonstrate the superiority of our approach over existing state-of-the-art few-shot segmentation methods.

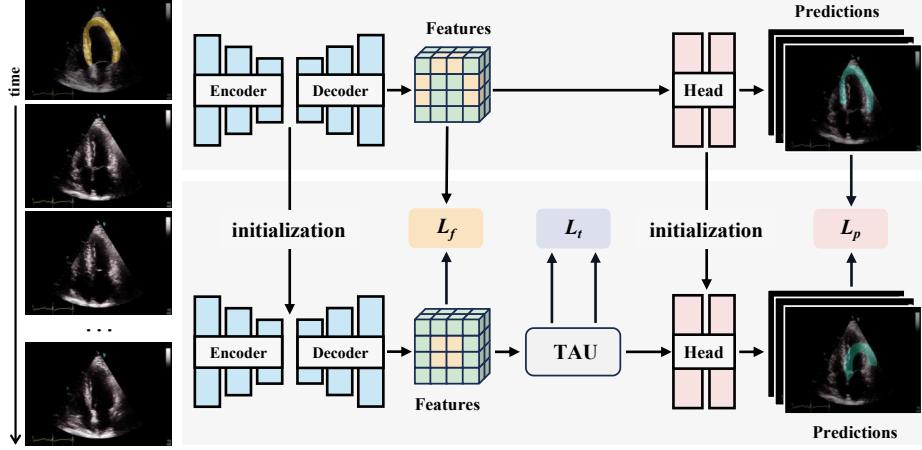


Fig. 1. Pipeline of the proposed few-shot segmentation method for medical video object segmentation.

2 Method

2.1 Task Definition

We address the task of few-shot video object segmentation, where given annotated frames from the beginning of a medical video sequence, the goal is to segment subsequent frames. Our approach leverages labeled images to train a model, which is then applied to test videos containing unseen object classes. Both training and testing are performed in an episodic manner, where each episode contains data sampled from the same class. Specifically, an episode comprises a support set $\mathcal{S} = \{(s_i, m_i)\}_{i=1}^N$ and a query set $\mathcal{Q} = \{(q_k, y_k)\}_{k=1}^K$, where s_i and m_i represent the i -th support image and its ground truth mask, respectively, while q_k and y_k denote the k -th query image and the corresponding annotation mask. The model segments q_k guided by $\mathcal{S}$. For video data, we use the first N frames as the support set and the remaining frames as query data.

2.2 Learning Few-Shot Segmentation with Images

We first train a few-shot medical image segmentation network that consists of three key components: a pseudo mask generation module, a cross-resolution fea-

ture fusion module, and a segmentation head. We extract visual features from support and query images using a pre-trained backbone. The pseudo mask generation module is designed to roughly localize query objects with minor cost. The cross-resolution feature fusion module then facilitates interaction between query features and support features along with pseudo masks at multiple scales, thereby preserving fine image details and precise localization cues. The output features from this module are fed into the segmentation head, comprising a 3×3 convolutional layer followed by a 1×1 convolution with softmax, to predict segmentation masks.

Pseudo Mask Generation Pseudo masks have been widely adopted in few-shot segmentation models [11,12] to provide coarse object localization. Specifically, they measure the cosine similarity between query and support features. We utilize high-level features to acquire pseudo masks, as they encode richer semantic information. Let $\mathcal{F}$ denote the backbone and $\mathbf{p}_k$ represent the generated pseudo mask for $\mathbf{q}_k$. We have

$$\mathbf{p}_k = \cos(\mathcal{F}(\mathbf{q}_k), \mathcal{F}(\mathbf{s}_i \odot \mathbf{m}_i)), \quad (1)$$

where $\odot$ is the Hadamard product. We then normalize $\mathbf{p}_k$ to the range of $[0, 1]$ using min-max normalization.

Cross-Resolution Feature Fusion First, we combine the resulting pseudo mask $\mathbf{p}_k$, containing localization cues, with support and query features as follows:

$$\mathbf{f} = \text{Conv}_{1 \times 1} \circ \text{Concat}(\text{up}(\mathbf{p}_k), \mathcal{G}(\mathbf{q}_k), \mathcal{G}(\mathbf{s}_i \odot \mathbf{m}_i)), \quad (2)$$

where $\circ$ is a composition function, $\text{up}(\cdot)$ upsamples the pseudo mask, $\mathcal{G}$ denotes the first half of the backbone network, and $\text{Concat}(\cdot)$ represents channel-wise concatenation. Note that we utilize intermediate features (via $\mathcal{G}$) instead of high-level ones as in Eq. (1), since they retain finer-grained details. Then, to derive multi-scale representations, we exploit the heavy neck module from GiraffeDet [13], and its output features are finally fed into the segmentation head to predict segmentation masks.

2.3 Inference on Individual Video Frames

A straightforward approach to applying the above-mentioned few-shot segmentation model trained on images to videos is to treat each video frame as an independent sample, using the first few annotated frames as support data, with the remaining frames as individual query inputs.

However, empirical evidence suggests that this direct application does not yield optimal segmentation results. The fundamental issue arises from treating query data, i.e., video frames, as isolated inputs, inherently disregarding the temporal continuity inherent in video content.

2.4 Spatiotemporal Consistency Relearning for Videos

Temporal Consistency As mentioned above, the few-shot segmentation model, which is merely trained on images, fails to consider inter-frame correlations. A common approach to mitigate this issue, while reducing computational complexity, involves integrating new temporal modules with a frozen base model, such as adapters [32]. However, such methods typically require extra dense video annotations for parameter learning, which undermines the intention of few-shot learning in medical videos.

To overcome these issues, we first introduce a temporal attention unit [14] placed between the cross-resolution feature fusion module and segmentation head of the trained image model, designed to capture connections between frames. Then, we relearn this new model in a self-supervised manner. Under the temporal continuity prior, we utilize cosine similarity to regularize the consistency of features of consecutive frames. This serves to capture temporal relationships without supervision while improving segmentation.

Formally, let $\mathbf{f}_t$ denote feature maps of the t -th frame from the inserted temporal attention module in a batch $\mathcal{B}$ which contains consecutive frames. We can have the following temporal consistency regularization:

$$\mathcal{L}_t = 1 - \frac{1}{|\mathcal{B}|} \sum_{t=1}^{|\mathcal{B}|-1} \cos(\mathbf{f}_t, \mathbf{f}_{t+1}). \quad (3)$$

Feature and Prediction Consistency Constraints In the relearning phase, altering model parameters could cause the model to diverge from its training, thus deteriorating its performance on video data. For stability, we introduce the following two consistency constraints.

The feature consistency implies that given the same inputs, features generated by the relearning model should remain correlated to the feature distribution of the trained, frozen image model. Let $\mathbf{z}_{i,j}$ and $\hat{\mathbf{z}}_{i,j}$ denote features from cross-resolution feature fusion modules of the two models, respectively. The feature consistency constraint can be defined as:

$$\mathcal{L}_f = \frac{1}{|\mathcal{B}|D} \sum_{t=1}^{|\mathcal{B}|} \sum_{j=1}^N (\hat{\mathbf{z}}_{i,j} - \mathbf{z}_{i,j})^2, \quad (4)$$

where j is a spatial index, and N denotes the number of feature vectors.

The prediction consistency minimizes the discrepancy between segmentation masks predicted by the two models for the same frames. It can be formulated as:

$$\mathcal{L}_p = \frac{1}{|\mathcal{B}|} \sum_{t=1}^{|\mathcal{B}|} (\hat{\mathbf{y}}_t - \mathbf{y}_t)^2. \quad (5)$$

Overall Objective The weights of our relearning model, excluding the temporal attention unit, are initialized by the model pre-trained on images in the first phase. We fix the segmentation head and make the other parameters trainable. The overall objective for relearning is

$$\mathcal{L} = \lambda_1 \mathcal{L}_t + \lambda_2 \mathcal{L}_f + \lambda_3 \mathcal{L}_p, \quad (6)$$

where λ_1 , λ_2 , and λ_3 are trade-off hyperparameters.

Fig. 1 shows our method. The shift from medical images to videos leads to performance degradation when deploying an image model on videos. The proposed relearning method enables the model to adapt to test video characteristics, mitigating this domain shift. By fine-tuning a subset of parameters, coupled with newly introduced ones, on a few labeled frames from the test data, the model can better capture the underlying distribution of the test data, improving performance.

3 Experiment

3.1 Experimental Setups

Datasets For medical images, we consider the Breast Ultrasound Images (BUSI) dataset [20], the Multi-Modality Ovarian Tumor Ultrasound (MMOTU) dataset [23], the TN3K dataset [21], the Digital Database Thyroid Image (DDTI) dataset [22], and the Laryngeal Endoscopic dataset [24]. For medical videos, we use the HMC-QU dataset [25] and the ASU-Mayo Clinic Colonoscopy Video (ASU-Mayo) dataset [26]. To train a few-shot segmenter for ultrasound videos, we utilize BUSI, MMOTU, TN3K, and DDTI as training images, and perform relearning as well as testing on HMC-QU. Similarly, for experiments on endoscopic videos, we employ the Laryngeal Endoscopic dataset and ASU-Mayo. To enhance the performance of the image model learned in the first phase, we leverage a training strategy that combines natural image datasets, PASCAL-5ⁱ [18] and COCO-20ⁱ [19], with the above-mentioned medical image datasets. Note that models are trained on base image classes and evaluated on novel video classes with no class overlap, assessing generalization to unseen data.

Evaluation Metrics Following prior works [15,16], we adopt the Dice score and foreground-background IoU as evaluation metrics to assess the performance of few-shot segmentation models.

Implementation Details For a fair comparison, all experiments are conducted under the most challenging one-shot setting, where only one annotated image (first phase) or frame (second phase) is used as the support set. We employ ResNet50 [17] pre-trained on ImageNet as our backbone network. During training of the few-shot medical image segmentation model, the backbone’s weights are frozen except for block #4, which is fine-tuned to learn more robust feature

Table 1. Quantitative comparison with competing methods on the HMC-QU ultrasound video dataset for few-shot segmentation. Performance is reported in terms of Dice (%) and IoU (%).

Images	Thyroid		Breast		Ovary		All	
methods	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
PATNet [27]	32.91	21.82	32.57	20.37	35.65	22.44	36.61	22.76
HSNet [28]	67.97	52.24	70.60	55.28	68.63	53.01	69.71	54.15
DCAMA [29]	60.09	41.02	75.61	61.33	63.03	46.52	68.75	52.99
DCP [30]	51.34	35.26	47.47	31.63	45.43	30.10	43.23	27.74
AFA [31]	72.36	57.41	75.11	60.63	67.99	52.26	65.03	49.26
Ours	78.15	64.17	85.79	75.54	84.55	73.26	71.90	56.42

Table 2. Comparison with other methods on the ASU-Mayo endoscopic video dataset.

Images	Laryngeal	
methods	Dice	IoU
PATNet [27]	44.39	33.54
HSNet [28]	59.21	46.96
DCAMA [29]	56.67	45.60
DCP [30]	49.74	37.49
AFA [31]	60.38	47.42
Ours	63.97	48.40

Table 3. Ablation studies on the effectiveness of each proposed consistency regularization term.

	HMC-QU		ASU-Mayo	
$\mathcal{L}_t$ $\mathcal{L}_f$ $\mathcal{L}_p$	Dice	IoU	Dice	IoU
$\times$ $\times$ $\times$	83.55	71.89	62.69	48.05
$\times$ $\checkmark$ $\checkmark$	82.10	70.17	60.36	46.43
$\checkmark$ $\times$ $\checkmark$	83.37	71.26	61.85	47.65
$\checkmark$ $\checkmark$ $\times$	1.25	1.07	0.02	0.01
$\checkmark$ $\checkmark$ $\checkmark$	85.79	75.54	63.97	48.4

maps. The model is trained using the Adam optimizer with a learning rate of $1e-4$ for the first 5K iterations, followed by the SGD optimizer with a learning rate of $1e-5$ for subsequent iterations. The batch size is set to 8. In the second phase, we use the SGD optimizer with a learning rate of $1e-5$ and a batch size of 4. Our method is implemented in PyTorch and runs on NVIDIA RTX 4090 GPUs.

3.2 Results

We compare the proposed model with state-of-the-art few-shot segmentation approaches, including PATNet [27], HSNet [28], DCAMA [29], DCP [30], and AFA [31]. PATNet and AFA utilize the prototype learning paradigm, whereas HSNet, DCAMA, and DCP introduce novel mechanisms to fully leverage information from support image-mask pairs. In Table 1, “thyroid”, “breast”, and “ovary” denote the use of image datasets TN3K and DDTI, BUSI, and MMOTU, respectively, to train the few-shot medical image segmentation model. We also train the model with the combined ultrasound image datasets. In Table 2, “laryngeal” indicates the usage of the endoscopic dataset [24]. Experimental results

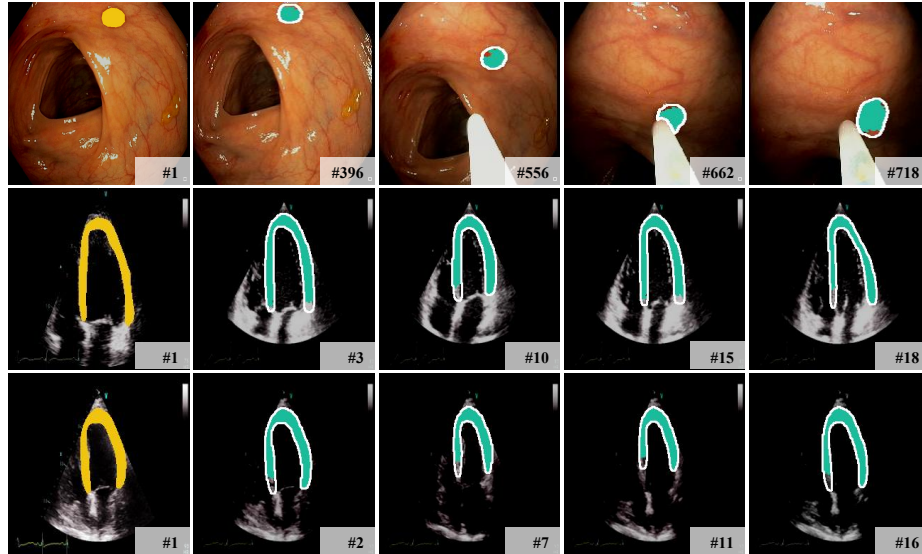


Fig. 2. Qualitative results of our few-shot video object segmentation model on the HMC-QU and ASU-Mayo datasets. The first column shows annotated support frames with ground truth masks (yellow). The remaining columns illustrate our model’s segmentation predictions (green masks) on sampled query frames from videos. Ground truth masks for the query frames are outlined in white for reference.

in Table 1 and Table 2 demonstrate that the proposed method outperforms competing approaches. The consistent superiority of our method across different datasets indicates its effectiveness and generalizability. Additionally, qualitative results (see Fig. 2) show that our method generates better segmentation results for medical videos. Regarding time complexity, our model achieves 78 fps, comparable to baselines.

3.3 Ablation Study

We conduct ablation studies to investigate the effectiveness of each component in the proposed spatiotemporal consistency relearning approach. Table 3 reports numerical results. We set the few-shot segmentation model pre-trained on images as the baseline. Then, we evaluate our relearning network under different loss combination settings. Compared to the baseline, lacking the temporal consistency regularization, i.e., $\mathcal{L}_t$, the Dice and IoU scores on both datasets decrease. Moreover, we observe that the constraints on feature consistency and prediction consistency, especially the latter, are essential for aligning distributions between the two phases, thereby guaranteeing model performance. By learning with all the introduced consistency losses, our model outperforms all competing approaches.

4 Conclusion

In conclusion, this work presents a novel approach to few-shot video object segmentation that bridges the gap between abundant annotated medical images and sparse video annotations. By first performing few-shot segmentation pre-training on labeled images and then introducing a spatiotemporal consistency relearning strategy on medical videos, our two-phase framework achieves strong performance in this extremely low-data regime without requiring costly dense video annotations. The relearning phase enforces consistency across consecutive frames while also maintaining constraints with the pre-trained image model. Experiments validate the superiority of our method over existing state-of-the-art few-shot segmentation techniques for video data. Overall, this approach provides an effective solution to leverage existing image labels for the video domain, minimizing costly new annotations. By enabling accurate video segmentation from just a few labeled frames, it opens up new possibilities for applications in the medical field where video data is prevalent but annotations are limited.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this paper.

References

1. Che, H., Chen, S., Chen, H.: Image quality-aware diagnosis via meta-knowledge co-embedding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19819–19829. (2023)
2. Che, H., Cheng, Y., Jin, H., Chen, H.: Towards generalizable diabetic retinopathy grading in unseen domains. In: International Conference on Medical Image Computing and Computer Proceted Intervention, pp. 430–440. (2023)
3. Sun, L., Li, C., Ding, X., Huang, Y., Chen, Z., Wang, G., Yu, Y., Paisley, J.: Few-shot medical image segmentation using a global correlation network with discriminative embedding. *Computers in Biology and Medicine* **140**, 105067 (2022)
4. Rakelly, K., Shelhamer, E., Darrell, T., Efros, A., Levine, S.: Conditional networks for few-shot semantic segmentation. *arXiv preprint arXiv:1806.07373* (2018)
5. Tritrong, N., Rewatbowornwong, P., Suwajanakorn, S.: Repurposing GANs for one-shot semantic part segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4475–4485. (2021)
6. Dong, N., Xing, E.P.: Few-shot semantic segmentation with prototype learning. In: British Machine Vision Conference, p. 79 (2018)
7. Roy, A.G., Siddiqui, S., Pölsterl, S., Navab, N., Wachinger, C.: ‘Squeeze & excite’ guided few-shot segmentation of volumetric images. *Medical Image Analysis* **59**, 101587 (2020)
8. Li, Y., Fu, Y., Gayo, I.J.M.B., Yang, Q., Min, Z., Saeed, S.U., Yan, W., Wang, Y., Noble, J.A., Emberton, M.: Prototypical few-shot segmentation for cross-institution male pelvic structures with spatial registration. *Medical Image Analysis* **90**, 102935 (2023)
9. Chen, H., Wu, H., Zhao, N., Ren, S., He, S.: Delving deep into many-to-many attention for few-shot video object segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14040–14049. (2021)

10. Yan, K., Li, X., Wei, F., Wang, J., Zhang, C., Wang, P., Lu, Y.: Two-shot video object segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2257–2267. (2023)
11. Liu, J., J., Bao, Y., Xie, G., Xiong, H., Sonke, J.-J. and Gavves, E.: Dynamic prototype convolution network for few-shot semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11543–11552. (2022)
12. Zhang, G., Kang, G., Yang, Y., Wei, Y.: Few-shot segmentation via cycle-consistent Transformer. In: Advances in Neural Information Processing Systems, pp. 21984–21996. (2021)
13. Jiang, Y., Tan, Z., Wang, J., Sun, X., Lin, M. and Li, H.: GiraffeDet: A heavy-neck paradigm for object detection. arXiv preprint arXiv:2202.04256 (2022)
14. Tan, C., Gao, Z., Li, S., Xu, Y., Li, S. Z.: Temporal attention unit: Towards efficient spatiotemporal predictive learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18770–18782. (2022)
15. Liu, Y., Zhang, X., Zhang, S., He, X.: Part-aware prototype network for few-shot semantic segmentation. In: European Conference on Computer Vision, pp. 142–158. (2020)
16. Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., Rueckert, D.: Self-supervised learning for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging* **41**(7), 1837–1848 (2022)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. (2016)
18. Shaban, A., Bansal, S., Liu, Z., Essa, I., Boots, B.: One-shot learning for semantic segmentation. arXiv preprint arXiv:1709.03410 (2017)
19. Nguyen, K., Todorovic, S.: Feature weighting and boosting for few-shot segmentation. In: IEEE/CVF International Conference on Computer Vision, pp. 622–631. (2019)
20. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. *Data in Brief* **28**, 104863 (2020)
21. Gong, H., Chen, J., Chen, G., Li, H., Chen, F., Li, G.: Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Computers in Biology and Medicine*, **155**, 106389 (2023)
22. Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., Romero, E.: An open access thyroid ultrasound image database. In: International Symposium on Medical Information Processing and Analysis, pp. 188–193. (2015)
23. Zhao, Q., Lyu, S., Bai, W., Cai, L., Liu, B., Wu, M., Sang, X., Yang, M., Chen, L.: A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. arXiv preprint arXiv:2207.06799 (2022)
24. Laves, M.H., Bicker, J., Kahrs, L.A., Ortmaier, T.: A dataset of laryngeal endoscopic images with comparative study on convolution neural network-based semantic segmentation. *International Journal of Computer Assisted Radiology and Surgery* **14**, 483–492 (2019)
25. Degerli, A., Zabihi, M., Kiranyaz, S., Hamid, T., Mazhar, R., Hamila, R., Gabbouj, M.: Early detection of myocardial infarction in low-quality echocardiography. *IEEE Access* **9**, 34442–34453 (2021)
26. Tajbakhsh, N., Gurudu, S. R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging* **35**(2), 630–644 (2015)

27. Wang, K., Liew, J. H., Zou, Y., Zhou, D., Feng, J.: PANet: Few-shot image semantic segmentation with prototype alignment. In: IEEE/CVF International Conference on Computer Vision, pp. 9197–9206. (2019)
28. Min, J., Kang, D., Cho, M.: Hypercorrelation squeeze for few-shot segmentation. In: IEEE/CVF International Conference on Computer Vision, pp. 6941–6952. (2021)
29. Shi, X., Wei, D., Zhang, Y., Lu, D., Ning, M., Chen, J., Ma, K., Zheng, Y.: Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In: European Conference on Computer Vision, pp. 151–168. (2022)
30. Lang, C., Cheng, G., Tu, B., Han, J.: Few-shot segmentation via divide-and-conquer proxies. *International Journal of Computer Vision* **132**(1), 261–283 (2024)
31. Karimijafarbigloo, S., Azad, R., Merhof, D.: Self-supervised few-shot learning for semantic segmentation: An annotation-free approach. In: International Conference on Medical Image Computing and Computer Assisted Intervention Workshops, pp. 159–171. (2023)
32. Park, J., Lee, J., Sohn, K.: Dual-path adaptation from image to video Transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2203–2213. (2023)