Symmetrical Supervision with Transformer for Few-shot Medical Image Segmentation

1st Yao Niu
Department of Artificial Intelligence
Xiamen University
Fujian, China

2nd Zhiming Luo ✉
Department of Artificial Intelligence
Xiamen University
Fujian, China

3rd Sheng Lian
College of Computer and Data Science
Fuzhou University
Fujian, China

4th Lei Li
Department of Software Engineering
Xiamen University
Fujian, China

5th Shaozi Li
Department of Artificial Intelligence
Xiamen University
Fujian, China

6th Haixin Song
Department of Artificial Intelligence
Xiamen University
Fujian, China

***Abstract*—Few-shot learning can potentially learn the target knowledge in extremely few data regimes. Existing few-shot medical image segmentation methods fail to consider the global anatomy correlation between the support and query sets. They generally adopt a weak one-way information transmission that can not fully explore the knowledge to segment query data. To address this problem, we propose a novel Symmetrical Supervision network based on traditional two-branch methods. We raise two main contributions: (1) The Symmetrical Supervision Mechanism is leveraged to strengthen the supervision of network training; (2) A transformer-based Global Feature Alignment module is introduced to increase the global consistency between the two branches. Experimental results on two challenging datasets (abdominal segmentation dataset CHAOS and cardiac segmentation dataset MS-CMRSeg) show a remarkable performance compared to other comparing methods.**

***Index Terms*—few-shot segmentation, Symmetrical Supervision, Global Feature Alignment.**

I. Introduction

Accurate medical image segmentation plays an essential role in many clinical applications. Recently, deep learning-based segmentation methods [1], [2] have achieved superior performance in medical image segmentation tasks. However, deep neural networks need a large amount of pixel-wise labeled data for training, which is hard to obtain. Moreover, the trained networks generally have a weak generalization performance. Few-shot learning has been proposed to address these challenges in medical image segmentation, which learns a generalization model that can use just a few labeled examples to segment the target organs.

Existing few-shot segmentation methods [3]–[10] generally adopt a two-branch structure: the conditioner branch for support data and the segmenter branch for query data, as shown in Fig. 1**(a)**. In general, these methods usually perform a one-way transmission that the conditioner branch extracts the features of support data to guide the segmentation of query data in the segmenter branch. However, such a one-way transmission can not fully carry out the information exchange between the two branches. Furthermore, most methods [8], [9] in few-shot medical image segmentation fail to consider this global consistency information lies in support and query data.

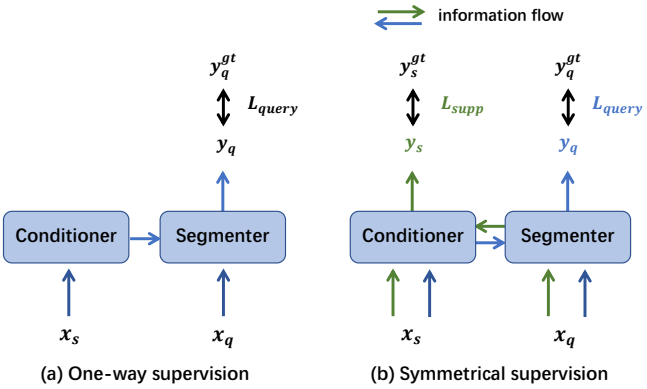

Fig. 1. Information flow in the few-shot segmentation. (a) shows the data flow in the previous method, which is one-way supervision from conditioner to segmenter. Our proposed Symmetrical Supervision Mechanism as shown in (b). The information can also flow from segmenter to conditioner.

To deal with the above limitations, we propose a few-shot segmentation model with a Symmetrical Supervision Mechanism (SSM) instead of one-way supervision as shown in Fig. 1**(b)**. In our model, we also compute the feature of the query image from the segmenter branch to guide the segmentation of the conditioner for support images. The experiments show that SSM effectively compensates for the shortcomings of the insufficient semantic information transmission between support and query images in the two-branch structures and the weak supervision.

On the other aspect, the same organ in the support and query images usually follow a consistent global anatomical structure. However, the newly proposed Transformer network [11] has the advantage of modeling long-range global information in the images. Therefore, we further leverage a light-weighted Transformer by introducing a novel design Global Feature Alignment(GFA) to increase the global consistency between

✉ Corresponding author. Email: zhiming.luo@xmu.edu.cn

support-query data pairs.

In summary, the main contributions of this work are as follows.

- We tackle few-shot segmentation from the perspective of strengthening information exchange of support and query data. The supervisions are conducted symmetrically through Symmetrical Supervision Mechanism (SSM).
- We propose a novel Global Feature Alignment (GFA), which leverages the Transformer blocks to fully utilize the global anatomical consistency between image slices.
- Experimental results on two challenging datasets (abdominal segmentation dataset CHAOS and cardiac segmentation dataset MS-CMRSeg) demonstrate that our proposed method can outperform the state-of-the-art methods.

II. Problem Setting

Suppose we have two different medical sets D_{train} and D_{test} with non-overlapped category labels C_{train} and C_{test} (e.g. $C_{train} = \{\text{left kidney, right kidney, liver}\}$ and $C_{test} = \{\text{spleen}\}$), the goal of few-shot image segmentation is to learn a segmentation Model M on the D_{train} and can directly evaluate on the D_{test} without re-training. **In the training phase**, we construct the D_{train} into several episodes, and each episode contains a set of support images with labels and a query image, denoted as $(\mathbb{S}, \mathbb{Q})$. Besides, we define a N-way K-shot segmentation task as agreed in other papers. Specifically, the support set $\mathbb{S}$ has K (x_s^i, y_s^i) pairs per category with a total of N different categories from C_{train}. The $\mathbb{Q}$ usually contains one (x_q, y_q) pair from the same N categories as the support set. **During the testing phase**, we construct the testing episodes from D_{test} in the same manner, and evaluate the generalization performance of the segmentation M trained on D_{train}.

III. Approach

The overall architecture of our proposed method is shown in Figure 2, which consists of the Segmenter, the Conditioner, and the Global Feature Alignment (GFA) module. The Segmenter and Conditioner are used to extract features from the query and support data. Between them, we leverage the Symmetrical Supervision Mechanism to transmit knowledge from one to the other. The GFA module aligns high-level features from the Segmenter and the Conditioner to improve global consistency between support and query data. Additionally, in the Segmenter branch, we further adopt a Refinement Loop (RL) to refine the segmentation results of x_q.

A. Conditioner & Segmenter

The traditional U-Net structure [2] is used as the backbone for both the conditioner and the segmenter, as shown in Fig. 3. In each episode, we concatenate the support image x_s with y_s^{gt} and then feed it into the conditioner module. The query image x_q is fed into the segmenter module.

To symmetrically transmit the information between the conditioner and the segmenter, we adopt the SSE module proposed in [12] as the interaction block. Especially, the SSE module leverage a channel-wise squeeze and a spatially excite to suppress the irrelevant background areas in the support features from different levels, thus the segmenter can pay more attention to useful foreground information. For the information flow propagated from conditioner to segmenter (green flow in Fig. 3), we use the SSE module to re-calibrate the feature of the segmentation by:

$$\text{Sigmoid}\left(\text{Conv}_{1*1,1}(E_{i-1}^c)\right) \times E_{i-1}^s, \tag{1}$$

where E_{i-1}^c and E_{i-1}^s stand for intermediate features from the conditioner and the segmenter, respectively. $\text{Conv}_{1*1,1}$ is a convolution operation with 1×1 kernel and 1 output channel.

For the segmenter to conditioner flow (blue flow in Fig. 3), we compute the corresponding re-calibrated features of the conditioner by:

$$\text{Sigmoid}\left(\text{Conv}_{1*1,1}(E_{i-1}^s)\right) \times E_{i-1}^c. \tag{2}$$

Note that, we perform the feature re-calibration at each middle level in the encoder and decoder of the U-Net backbones.

B. Symmetrical Supervision Mechanism

The Symmetrical Supervision Mechanism is to use a dual-directional symmetrical information flow between the conditioner and segmenter to enhance supervision. Existing methods merely allow the query to learn the information of the labeled support data from the conditioner branch, thus wasting the ground truth supervision of the support data, and cannot fully dig out the relationship between support and query data. These factors make the supervision of few-shot learning weaker. Therefore, we introduce the Symmetrical Supervision Mechanism to enhance supervision. During the training process, the intermediate features of each stage in the segmenter branch are also transferred to the corresponding position of the conditioner branch through the same interaction block, guiding the segmentation task of support data. Fig. 1**(b)** shows the sketch map of our proposed Symmetrical Supervision Mechanism. Experiment results show that such mechanism can enhance the interaction between the upper and lower branch.

C. Global Feature Alignment Module

In general, the same organs in the support images and query images usually follow a consistent anatomical structure. This information can be utilized to enhance the ability of query images to learn from the precious support example. So we leverage a Global Feature Alignment (GFA) module to increase the similarity of their global features. Transformer [13]–[15] gets a powerful ability to extract global information. Thus we use the Transformer to perform the GFA. Before feeding the features from E^S and E^C into the transformer blocks, we first use a patch embedding and a position embedding to encode the features as in ViT [11],

$$f_q = \text{PE}(f_q^{enc}) + E_{pos}, \tag{3}$$

$$f_s = \text{PE}(f_s^{enc}) + E_{pos}, \tag{4}$$

where f_q^{enc} and f_s^{enc} are the features from E^S and E^C, PE increases the feature dimension at each spatial location into 768, and E_{pos} is the position embedding.

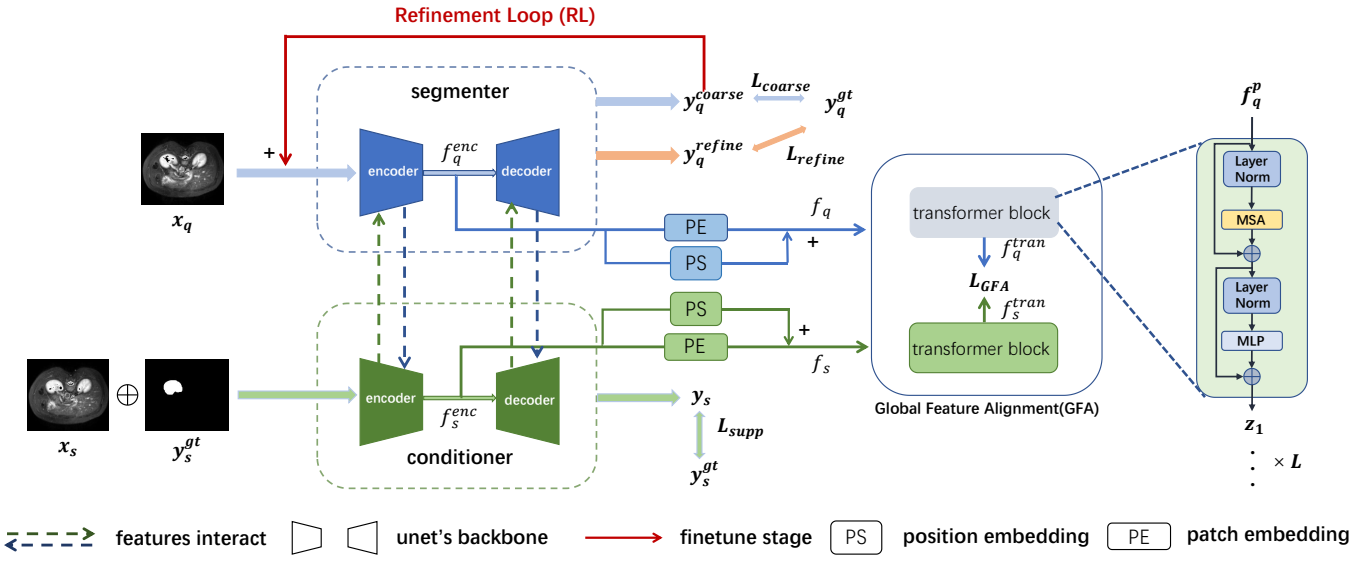

Fig. 2. The network architecture of our proposed method. It is mainly composed of three parts: The conditioner and the segmenter extract and analyze the features of support images and query images, respectively; and the Global Feature Alignment part shortens the distance between the global features from the two parts mentioned above to keep anatomical similarity between slices. The Symmetrical Supervision Mechanism shows by the blue and red lines.

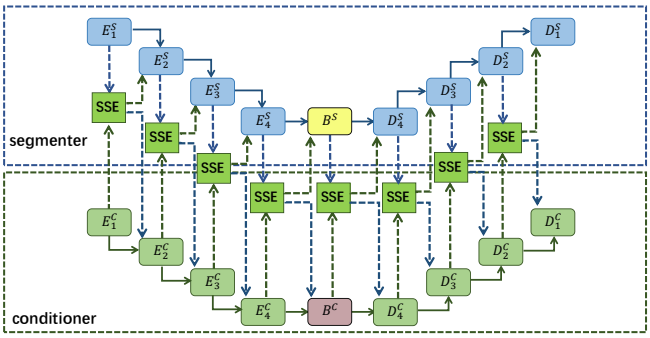

Fig. 3. Implementation details of conditioner and segmenter. Features from support data and query data at different levels are interacted by SSE modules and then passed through the U-Net architecture.

a) Transformer Block.: The Transformer unit is composed of (1) Multihead Self-Attention (MSA), (2) Multi-Layer Perception (MLP), and (3) two skip connections, as shown in Fig. 2. Both transformers for the query and support data contain L Transformer units sequentially. The final global features extracted from f_s and f_q by Transformer blocks are denoted as f_s^{tran} and f_q^{tran}.

b) Feature Alignment.: We intend to close the semantic relationship between the two branches using the Feature Alignment block so that the communication between the segmenter and the conditioner is more effective and closer. To increase the similarity between f_s^{tran} and f_q^{tran}, we perform a feature alignment between them by increasing the cosine similarity. The optimization loss function is as follows:

$$L_{GFA} = 1 - \text{Cosine}(f_s^{tran}, f_q^{tran}), \tag{5}$$

where $\text{Cosine}(f_s^{tran}, f_q^{tran})$ compute their cosine similarity.

D. Loss Function

Loss for Query Data. We acquire L_{coarse} for the query data by computing the Dice loss between the y_q^{coarse} and y_q^{gt}, and the L_{fine} between the y_q^{fine} and y_q^{gt}.

Loss for Support Data. The second loss L_{supp} is obtained by comparing the prediction of support data y_s with its ground-truth y_s^{gt}. In this process, features from the conditioner are fused with features from query data in the segmenter to add supervision symmetrically.

Final Loss. By further combining the Global feature alignment module, the final loss function for the training model is as follows:

$$loss = \lambda_1 L_{coarse} + \lambda_2 L_{fine} + \lambda_3 L_{supp} + \lambda_4 L_{GFA}, \tag{6}$$

where λ_1, λ_2, λ_3, λ_4 are hyper-parameters and all set to 1 in this study.

IV. Experiment

A. Experimental Setup

a) Dataset: We conduct experiments on two publicly available MRI datasets containing organs with various shapes, locations, and textures. The first one is from the ISBI 2019 Combined Healthy Abdominal Organ Segmentation Challenge (CHAOS) [20] with 20 3D T2-SPIR MRI scans. The other one is from the MICCAI 2019 Multi-sequence Cardiac MRI Segmentation Challenge (MS-CMRSeg) [21], containing 35 3D cardiac MRI scans. On the CHAOS dataset, we use four categories for training and testing, *i.e.*, Left kidney, Right kidney, Spleen, and Liver. While for the MS-CMRSeg dataset, we choose three categories to evaluate, *i.e.*, Left Ventricle Blood Pool (LV-BP), Left Ventricle Myocardium (LV-MYO), and Right Ventricle (RV).

TABLE I
COMPARISONS AGAINST THE STATE OF THE ARTS

Method	MS-CMRSeg				CHAOS				
	LV-BP	LV-MYO	RV	Mean	L.kildney	R.kidney	Spleen	Liver	Mean
SE-Net [8]	69.92	44.71	65.43	59.69	62.11	61.32	51.80	27.43	50.66
CANet [16]	78.99	43.61	61.10	61.07	69.53	77.15	67.05	72.88	71.65
PPNet [17]	67.78	42.61	60.80	57.06	62.13	71.78	66.57	**73.12**	68.40
PANet [18]	80.20	45.67	66.95	64.27	53.45	38.64	50.90	42.26	46.33
SSL-ALPNet [19]	**87.54**	60.19	76.08	74.60	73.63	78.39	67.02	73.05	73.02
Ours	86.79	**62.31**	**78.23**	**75.78**	**78.46**	**81.45**	**73.75**	72.90	**76.64**

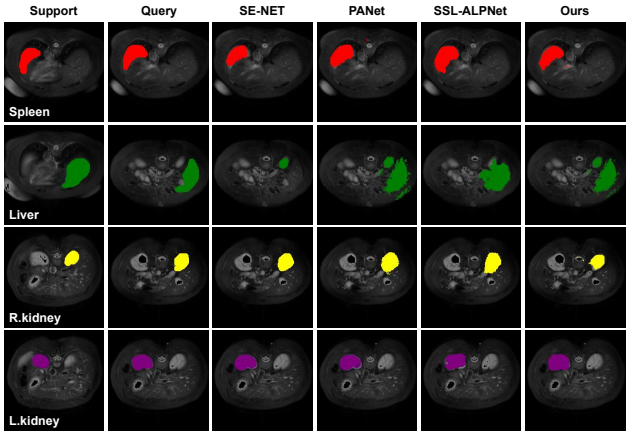

Fig. 4. Qualitative results of our method compared with other methods on abdominal MRI dataset CHAOS. We reproduced several representative methods for segmentation quality comparison. We see (left to right) the support image, the query image to be predicted with ground truth, and the segmentation results of the query slice of different models. Our proposed method achieves desirable results which are close to the ground truth.

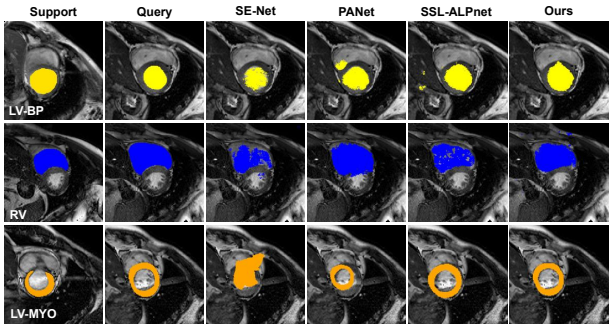

Fig. 5. Qualitative results of our method compared with other methods on Cardiac MRI dataset MS-CMRSeg. The proposed method achieves desirable results which are close to the ground truth.

We employ the mean dice score to compare the model predictions to the ground truth segmentation, which is commonly used in medical segmentation scenes.

b) Implementation Details.: Similar to [19], to get a fair result, we adopt a 5-fold cross-validation method and consider only 1-shot learning. We resized the MRI slices to a size of 256×256. The number of Transformer units is 6, and the number of headers in multi-head attention is 6.

B. Comparisons with the state of the arts

In this section, we compared our method with five other different methods, *i.e.*, SE-Net [8], CANet [16], vanilla PANet [18], PPNet [17] and SSL-ALPNet [19].

a) Quantitative Comparison: From Table I, we can observe that: **1)** Our proposed method consistently outperforms other methods, especially when compared with baseline SE-Net. our method improves the performance by 16.9% on MS-CMRSeg dataset and 25.98% on CHAOS dataset. **2)** In general, our method performs best for the left kidney and right kidney due to their relatively regular shape and the slight change along the slice direction, which are 78.46% and 81.45% respectively **3)** Our method is neither using prototypes nor pre-trained ResNet101 on the large-scale ImageNet and MS-COCO dataset. These results demonstrate that considering the anatomical global information and symmetrical supervision plays an important role for accurate few-show medical image segmentation.

b) Qualitative Comparison: In Figure 4 and Figure 5, we further show the qualitative results of different methods on two datasets. We can find that: **1)** Our proposed framework yields satisfying results on organs with various shapes, sizes, and intensities. **2)** The irregular edge that occurred in other methods is relieved by our proposed Global Feature Alignment module and Symmetrical Supervision Mechanism through enhancing information exchange and utilization of supervision information. **3)** In the traditional baseline network SE-Net (Row.2, Col.3 in Figure 4), the liver is wrongly segmented as a kidney, which is a typical phenomenon of overfitting that constantly occurs in few-shot learning caused by extremely low data regime.

C. Ablation Study

In this section, we conduct an ablation study on the abdominal MRI dataset CHAOS. To evaluate the effectiveness of the Refinement Loop (RL), Global Feature Alignment (GFA), and Symmetrical Supervision Mechanism (SSM). From Table II, we have following findings: **1)** Incorporating each component separately into the baseline model can improve the performance. Especially, the GFA and SSM can increase the Dice Score by 18.37% and 16.47%, respectively. **2)** When using two components, we can see that the GFA & SSM can significantly boost the Dice Score to 70.47, while the RL & GFA and RL & SSM only can slightly increase the performance. **3)** By jointly considering these three components, the mean Dice Score is

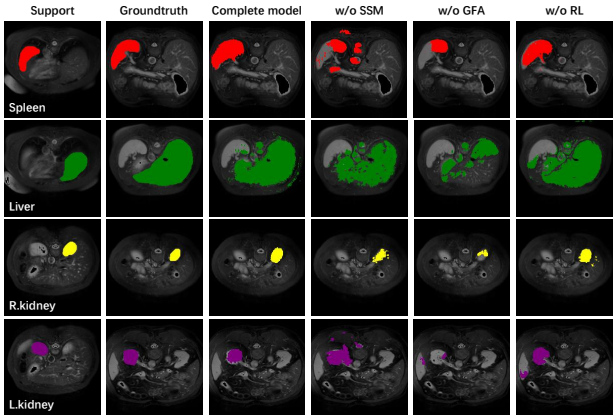

Fig. 6. Ablation study visualizations. (Col.1) Support images and the associated ground-truth masks. (Col.2) Query images and the associated ground-truth masks. (Col.3) Model results without Symmetrical Supervision Mechanism (Col.4) Model results without Refinement loop (Col.5) Model results without Global Feature Alignment.

further increased to 76.64%. The above findings validate the mutual benefits of the RL, GFA, and SSM.

TABLE II
ABLATION STUDY

RL	GFA	SSM	Liver	R.kidney	L.kidney	Spleen	Mean
			27.43	61.32	62.11	51.80	50.66
✓			38.83	50.69	59.09	59.49	52.03
	✓		58.72	59.33	57.24	58.65	58.49
		✓	61.32	59.58	56.94	54.74	58.15
	✓	✓	69.01	72.14	71.25	69.48	70.47
✓		✓	62.57	61.74	62.83	60.58	61.93
✓	✓		67.18	65.09	53.76	54.63	60.17
✓	✓	✓	**72.90**	**81.45**	**78.46**	**73.75**	**76.64**

Besides, we provide some qualitative results of removing one component in Figure 6. **1)** From Col.4, the absence of the SSM always leads to incomplete segmentation results, which suggests that one-direction supervision is insufficient in medical image analysis. **2)** From Col.5, the performance is the worst when lacking the GFA. These results indicate that Global Feature Alignment is essential for maintaining the anatomical structure of the organs. **3)** By comparing Col.3 and the last Col.6, we can find that the refinement loop helps remove some noisy predictions. These qualitative results further validate the effectiveness of these three components.

V. CONCLUSION

In this work, we propose a symmetrical supervision few-shot segmentation network with Transformers for medical images. Based on the traditional two-branch structure, we successfully introduce the Symmetrical Supervision Mechanism and the Global Feature Alignment module to strengthen the learning ability of the model and use a refinement loop to refine the segmentation results further. The proposed method can outperform many prototype-based methods when properly interacting the information between the support and query data.

ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (No. 61876159, No. 62076116, No. 62276221), and the Natural Science Foundation of Fujian Province of China (No. 2022J01002).

REFERENCES

[1] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert *et al.*, "nnu-net: Self-adapting framework for u-net-based medical image segmentation," *arXiv preprint arXiv:1809.10486*, 2018.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[3] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.

[4] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia, "Prior guided feature enrichment network for few-shot segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[5] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "Sg-one: Similarity guidance network for one-shot semantic segmentation," *IEEE Transactions on Cybernetics*, 2020.

[6] X. Li, T. Wei, Y. P. Chen, Y.-W. Tai, and C.-K. Tang, "FSS-1000: A 1000-class dataset for few-shot segmentation," in *CVPR*, 2020.

[7] K. Rakelly, E. Shelhamer, T. Darrell, A. A. Efros, and S. Levine, "Few-shot segmentation propagation with guided networks," *arXiv preprint arXiv:1806.07373*, 2018.

[8] A. G. Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger, "'squeeze & excite'guided few-shot segmentation of volumetric images," *Medical Image Analysis*, 2020.

[9] Q. Yu, K. Dang, N. Tajbakhsh, D. Terzopoulos, and X. Ding, "A location-sensitive local prototype network for few-shot medical image segmentation," in *ISBI*, 2021.

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

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks," *IEEE Transactions on Medical Imaging*, 2018.

[13] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[14] Y. Ji, R. Zhang, H. Wang, Z. Li, L. Wu, S. Zhang, and P. Luo, "Multi-compound transformer for accurate biomedical image segmentation," in *MICCAI*, 2021.

[15] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *MICCAI*, 2021.

[16] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, "Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *CVPR*, 2019.

[17] Y. Liu, X. Zhang, S. Zhang, and X. He, "Part-aware prototype network for few-shot semantic segmentation," in *ECCV*, 2020.

[18] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *CVPR*, 2019.

[19] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert, "Self-supervision with superpixels: Training few-shot medical image segmentation without annotation," in *ECCV*, 2020.

[20] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan *et al.*, "Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation," *Medical Image Analysis*, 2021.

[21] X. Zhuang, "Multivariate mixture model for myocardial segmentation combining multi-source images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.