Unlocking the Potential of Pre-trained Vision Transformers for Few-Shot Semantic Segmentation through Relationship Descriptors

Supplementary Material

In this supplementary material, we provide more details to complement the manuscript, including the implementation details in Sec. 7 and additional experimental results in Sec. 8.

7. Implementation Details

7.1. The detailed architecture of the decoder

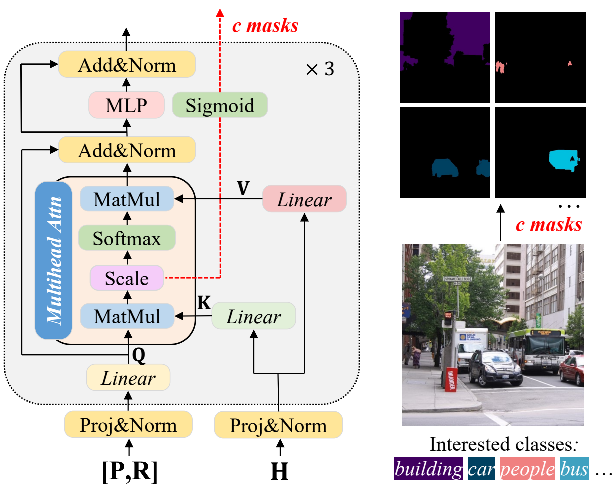

Figure 7. Details of decoder architecture used in our framework.

In the main paper, we propose a unified framework to generate semantic predictions by matching the class embeddings and visual embeddings in a vanilla transformer-based decoder as shown in Fig. 2 of the main paper. In this section, we provide more details of the decoder architecture as presented in Fig. 7.

Specifically, there are two inputs for the transformer-based decoder: the one input is $[\mathbf{P}, \mathbf{R}] \in \mathbb{R}^{C\times(2*d)}$, where $\mathbf{P} \in \mathbb{R}^{C\times d}$ and $\mathbf{R} \in \mathbb{R}^{C\times d}$ are class-wise prototype embedding and our proposed relationship descriptor (RD) embeddings respectively, and d is the feature dimension. The other input is $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_N] \in \mathbb{R}^{N\times d}$, where N is the number of patch tokens of an image and $\mathbf{h}_j$ denoting the jth patch. We can apply linear layers $\{\psi_q, \psi_k, \psi_v\}$ to generate $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ for query, key and value embeddings, respectively

$$\mathbf{Q} = \psi_q(\phi([\mathbf{P}, \mathbf{R}])) \in \mathbb{R}^{C\times d}, \tag{7}$$

$$\mathbf{K} = \psi_k(\varphi(\mathbf{H})) \in \mathbb{R}^{N\times d}, \tag{8}$$

$$\mathbf{V} = \psi_v(\varphi(\mathbf{H})) \in \mathbb{R}^{N\times d} \tag{9}$$

where $\{\phi, \varphi\}$ are the projection layers described in Sec. 4.1 and Eq. 3 of the main paper. The semantic masks could be obtained by calculating the scaled dot-product attention which is the intermediate product of the multi-head attention model (MHA):

$$\texttt{Masks} = \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} \in \mathbb{R}^{C\times N}. \tag{10}$$

where d_k is the dimension of the keys as a scaling factor. The final semantic segmentation results are obtained by applying `Argmax` operation on the class dimension of `Masks` logits.

7.2. Details of applying our trained GFSS model to (binary) FSS setting

In the main paper, we have successfully demonstrated the application of our optimized model, originally developed for the Generalized Few-Shot Segmentation (GFSS) setting, to the binary Few-Shot Segmentation (FSS) context. This was detailed in Section 5.5 and illustrated in Table 3 of the main paper. In this section, we offer further insights into the methods and processes we employed to accomplish this. Specifically, the binary FSS setting requires the model to segment out the target class objects and treat all other pixels as non-target (background) for a given testing image.

Therefore, we first adapt our optimized model to the FSS task by creating the novel class prototype from the target class objects in the support set and accumulating non-target-region features of the support set into the existing background class prototype. Moreover, since our optimized model has accommodated the knowledge for base classes, we propose to reinterpret predictions for these base classes as background class decisions in the binary FSS setting. This straightforward yet effective modification enabled our model to deliver remarkable results in the binary FSS setting. The complete results on both PASCAL-5^i and COCO-20^i datasets are provided in Tab. 7.

7.3. Pseudo code of our approach

For better comprehension, the complete pseudo code detailing the training and inference processes of our unified framework is presented in Algorithm 1.

Algorithm 1: Pseudo code of our framework

// Train on base & background classes C_B

Input: Dataset $\mathcal{D}_B$; Encoder with learnable prompts E, Decoder F, Relationship Descriptor Generator G;

Input: Initialized base class prototypes $\mathbf{P}_B$.

for sampled minibatch $\{\mathbf{I}, \mathbf{M}^{gt}\}_n^{bs}$ from $\mathcal{D}_B$ **do**

  $\mathbf{h}_{cls}, \mathbf{H} = E(\mathbf{I})$;

  // Generate Relationship Descriptors

  $\mathbf{R} = G(\mathbf{P}_B, \mathbf{H})$;

  $\mathbf{M} = sigmoid(F([\mathbf{P}_B, \mathbf{R}], \mathbf{H}))$;

  // Adamw update: E, F

  Updata parameters via $\mathcal{L}_{mask}(\mathbf{M}, \mathbf{M}^{gt})$;

  for class $c (c \in C_B)$ in $\{\mathbf{M}^{gt}\}_n^{bs}$ **do**

    // Momentum update: $\mathbf{P}_B$

    $\mathbf{P}^c = \frac{1}{\sum_{i,j}(\mathbf{M}^{gt})_{i,j}^c} \sum_{i,j}(\mathbf{M}^{gt})_{i,j}^c \mathbf{H}_{i,j}^c$

    $\mathbf{P}_B^c \leftarrow (1-\eta) * \mathbf{P}^c + \eta * \mathbf{P}_B^c$

  end

end

// Register novel classes C_N, $C_B \cap C_N = \emptyset$

Input: K-shot novel support set $\mathcal{S}_N = \{\mathbf{I}_\mathcal{S}, \mathbf{M}_\mathcal{S}^{gt}\}_n^{C_N * K}$

$\mathbf{h}_\mathcal{S}^{cls}, \mathbf{H}_\mathcal{S} = E(\mathbf{I}_\mathcal{S})$;

for class $c (c \in C_N)$ in $\{\mathbf{M}_\mathcal{S}^{gt}\}_n^{C_N * K}$ **do**

  $\mathbf{P}_N^c += \frac{1}{K} \frac{1}{\sum_{i,j}(\mathbf{M}_\mathcal{S}^{gt})_{i,j}^c} \sum_{i,j}(\mathbf{M}_\mathcal{S}^{gt})_{i,j}^c (\mathbf{H}_\mathcal{S})_{i,j}^c$

end

$\mathbf{P} = concat[\mathbf{P}_B, \mathbf{P}_N]$;

/* If test-time tuning: */

Adamw Updata $\{E, F\}$ via $\mathcal{L}_{mask}(\mathbf{M}_\mathcal{S}, \mathbf{M}_\mathcal{S}^{gt})$;

Momentum update: $\mathbf{P}$;

/* End if */

// Generalized segmentation on $C_B \cup C_N$

Input: Sampled testing image $\mathbf{I}$

$\mathbf{h}_{cls}, \mathbf{H} = E(\mathbf{I})$;

$\mathbf{M} = argmax(F(\phi([\mathbf{P}, G(\mathbf{P}, \mathbf{H})]), \varphi(\mathbf{H})))$;

Table 6. Effectiveness of our approach with an unsupervised pre-trained vision transformer model (i.e., ViT-B/16 pre-trained with DINO) on the PASCAL-5^i dataset. "Tuning" denotes the test-time tuning on the novel support set before inference under the generalized few-shot segmentation (GFSS) setting.

RD	Tuning	1-shot			5-shot		
		mIoU(N)	mIoU(B)	hIoU	mIoU(N)	mIoU(B)	hIoU
N/A	✗	14.2	68.2	23.5	16.5	68.3	26.6
single	✗	24.3	60.4	34.7	24.6	61.0	35.1
	✓	41.5	65.7	50.9	42.6	66.2	51.8
multiple	✗	34.4	69.4	46.0	37.1	70.3	48.6
	✓	47.8	68.8	56.4	49.0	69.6	57.5

8. Additional Experimental Results

8.1. Effect of various pre-trained vision transformer model for FSS task

In the main paper, we demonstrate that our proposed relationship descriptor (RD) module can unlock the potential of the supervised pre-trained ViT-B/16 model and improve the generalization ability for FSS tasks. In this section, we further present the effectiveness of the proposed method on the unsupervised pre-trained transformer model DINO [3]. As shown in Tab. 6, our method can consistently achieve better performance on both 1-shot and 5-shot settings.

8.2. Complete results under FSS setting

Due to space constraints, several represented literature works are included in Tab. 3 of the main paper for the binary FSS task. In this section, we present the full FSS results in Tab. 7.

Table 7. Comparison of our proposed method with the state-of-the-art FSS methods. Note that our method has not been trained on binary segmentation as well as test-time tuning on novel classes.

Method	Backbone	PASCAL-5^i		COCO-20^i	
		1-shot	5-shot	1-shot	5-shot
PANet [56]	RN-50	48.1	55.7	20.9	29.7
PFENet [53]	RN-50	60.1	61.4	32.4	37.4
SCL [66]	RN-50	61.8	62.9	-	-
RePri [1]	RN-50	59.1	66.8	34.0	42.1
MMNet [58]	RN-50	61.8	63.4	37.5	38.2
CMN [60]	RN-50	62.8	63.7	39.3	43.1
DPCN [28]	RN-50	66.7	69.9	43.0	49.8
BAM [22]	RN-50	67.8	70.9	46.2	51.2
MSANet [20]	RN-50	69.1	74.0	51.1	56.8
SVF [50]	RN-50	69.0	72.3	48.5	53.9
SiGCN [29]	RN-50	65.3	68.5	41.4	48.0
FECANet [27]	RN-50	69.3	74.9	50.9	58.3
ASGNet [23]	RN-101	59.3	63.9	34.5	42.5
SAGNN [59]	RN-101	62.1	62.8	37.2	42.7
CWT [33]	RN-101	58.0	64.7	32.4	42.0
Mining [63]	RN-101	62.6	68.8	36.4	44.4
HSNet [35]	RN-101	66.2	70.4	41.2	49.5
CAPL [54]	RN-101	63.6	68.9	42.8	50.4
IPMT [32]	RN-101	66.1	69.2	42.6	47.9
HM [36]	RN-101	67.8	70.9	45.9	50.6
VAT [19]	RN101	67.9	72.0	41.3	47.9
DACM [61]	RN-101	69.1	73.3	43.0	49.2
CLIPSeg [34]	CLIP-ViT/B	52.3	-	33.2	-
CLIPSeg+ [34]	CLIP-ViT/B	59.3	-	33.2	-
PGMA-Net [47]	CLIP-ViT/B	74.1	74.6	-	-
PGMA-Net [47]	CLIP-RN50	74.1	75.2	54.3	57.1
PGMA-Net [47]	CLIP-RN101	77.6	<u>78.6</u>	<u>59.4</u>	**61.8**
FPTrans [71]	ViT-B/16	64.7	73.7	42.0	53.8
FPTrans [71]	DeiT-B/16	68.8	78.0	47.0	58.9
HSNet [46]	Swin-B	67.3	71.6	47.3	55.1
DCAMA [46]	Swin-B	69.3	74.9	50.9	58.3
IPRNet [41]	ResNet101	67.5	70.9	46.9	53.3
Ours-single	ViT-B/16	<u>77.7</u>	78.0	57.1	59.2
Ours-multiple	ViT-B/16	**78.9**	**80.3**	**60.1**	<u>61.2</u>

References

[1] Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13979–13988, 2021. 2

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901, 2020. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 1, 2

[4] Weiyu Chen, Yencheng Liu, Zsolt Kira, Yuchiang Frank Wang, and Jiabin Huang. A closer look at few-shot classification. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 2

[5] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 2

[6] Philip Chikontwe, Soopil Kim, and Sang Hyun Park. Cad: Co-adapting discriminative features for improved few-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14554–14563, 2022. 2

[7] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 6

[8] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11583–11592, 2022. 3

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88:303–308, 2009. 5

[11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1126–1135. PMLR, 2017. 2

[12] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014. 2

[13] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 33: 21271–21284, 2020. 5

[14] Sina Hajimiri, Malik Boudiaf, Ismail Ben Ayed, and Jose Dolz. A strong baseline for generalized few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11269–11278, 2023. 2, 5, 6, 7

[15] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 991–998. IEEE, 2011. 5

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 5

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 1

[19] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *Proceedings of the IEEE conference on European Conference on Computer Vision (ECCV)*, pages 108–126. Springer, 2022. 2

[20] Ehtesham Iqbal, Sirojbek Safarov, and Seongdeok Bang. Msanet: Multi-similarity and attention guidance for boosting few-shot segmentation. *arXiv preprint arXiv:2206.09667*, 2022. 2, 7

[21] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *Proceedings of the International Conference on Machine Learning Deep Learning Workshop (ICML workshop)*. Lille, 2015. 2

[22] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8057–8067, 2022. 2, 7

[23] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8334–8343, 2021. 2, 7

[24] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021. 2

[25] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *Proceedings of the IEEE conference on European Conference on Computer Vision (ECCV)*, pages 280–296. Springer, 2022. 2

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the IEEE conference on European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 5

[27] Huafeng Liu, Pai Peng, Tao Chen, Qiong Wang, Yazhou Yao, and Xian-Sheng Hua. Fecanet: Boosting few-shot semantic segmentation with feature-enhanced context-aware network. *IEEE Transactions on Multimedia (TMM)*, 2023. 7, 2

[28] Jie Liu, Yanqi Bao, Guo-Sen Xie, Huan Xiong, Jan-Jakob Sonke, and Efstratios Gavves. Dynamic prototype convolution network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11553–11562, 2022. 2

[29] Jie Liu, Yanqi Bao, Wenzhe Ying, Haochen Wang, Yang Gao, Jan-Jakob Sonke, and Efstratios Gavves. Few-shot semantic segmentation with support-induced graph convolutional network. *arXiv preprint arXiv:2301.03194*, 2023. 2

[30] Sun-Ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. Learning orthogonal prototypes for generalized few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11319–11328, 2023. 2, 5, 6, 7

[31] Weide Liu, Zhonghua Wu, Yang Zhao, Yuming Fang, Chuan-Sheng Foo, Jun Cheng, and Guosheng Lin. Harmonizing base and novel classes: A class-contrastive approach for generalized few-shot segmentation. *arXiv preprint arXiv:2303.13724*, 2023. 2, 5, 6, 7

[32] Yuanwei Liu, Nian Liu, Xiwen Yao, and Junwei Han. Intermediate prototype mining transformer for few-shot semantic segmentation. *arXiv preprint arXiv:2210.06780*, 2022. 2

[33] Zhihe Lu, Sen He, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8741–8750, 2021. 2

[34] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7076–7086, 2022. 2, 7

[35] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6941–6952, 2021. 7, 2

[36] Seonghyeon Moon, Samuel S Sohn, Honglu Zhou, Sejong Yoon, Vladimir Pavlovic, Muhammad Haris Khan, and Mubbasir Kapadia. Hm: Hybrid masking for few-shot segmentation. In *Proceedings of the IEEE conference on European Conference on Computer Vision (ECCV)*, pages 506–523. Springer, 2022. 2

[37] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2554–2563. PMLR, 2017. 2

[38] Josh Myers-Dean, Yinan Zhao, Brian Price, Scott Cohen, and Danna Gurari. Generalized few-shot semantic segmentation: All you need is fine-tuning. *arXiv preprint arXiv:2112.10982*, 2021. 2, 5, 6, 7

[39] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 34:23296–23308, 2021. 2

[40] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 2

[41] Atsuro Okazawa. Interclass prototype relation for few-shot segmentation. In *Proceedings of the IEEE conference on European Conference on Computer Vision (ECCV)*, pages 362–378. Springer, 2022. 2

[42] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 3

[44] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 2

[45] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065*, 2016. 2

[46] Xinyu Shi, Dong Wei, Yu Zhang, Donghuan Lu, Munan Ning, Jiashun Chen, Kai Ma, and Yefeng Zheng. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In *Proceedings of the IEEE conference on European Conference on Computer Vision (ECCV)*, pages 151–168. Springer, 2022. 2, 7

[47] Chen Shuai, Meng Fanman, Zhang Runtong, Qiu Heqian, Li Hongliang, Wu Qingbo, and Xu Linfeng. Visual and textual prior guided mask assemble for few-shot segmentation and beyond. *arXiv preprint arXiv:2308.07539*, 2023. 7, 2

[48] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 4077–4087, 2017. 2

[49] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7262–7272, 2021. 2

[50] Yanpeng Sun, Qiang Chen, Xiangyu He, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Jian Cheng,

Zechao Li, and Jingdong Wang. Singular value fine-tuning: Few-shot segmentation requires few-parameters fine-tuning. *arXiv preprint arXiv:2206.06122*, 2022. 2

[51] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1199–1208, 2018. 2

[52] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Proceedings of the IEEE conference on European Conference on Computer Vision (ECCV)*, pages 266–282. Springer, 2020. 2

[53] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44 (2):1050–1065, 2020. 2, 6, 7

[54] Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11563–11572, 2022. 2, 3, 5, 6, 7

[55] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. In *Proceedings of the IEEE conference on European Conference on Computer Vision (ECCV)*, pages 730–746. Springer, 2020. 2

[56] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9197–9206, 2019. 2, 6, 7

[57] Davis Wertheimer, Luming Tang, and Bharath Hariharan. Few-shot classification with feature map reconstruction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8012–8021, 2021. 2

[58] Zhonghua Wu, Xiangxi Shi, Guosheng Lin, and Jianfei Cai. Learning meta-class memory for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 517–526, 2021. 2

[59] Guo-Sen Xie, Jie Liu, Huan Xiong, and Ling Shao. Scale-aware graph neural network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5475–5484, 2021. 7, 2

[60] Guo-Sen Xie, Huan Xiong, Jie Liu, Yazhou Yao, and Ling Shao. Few-shot semantic segmentation with cyclic memory network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7293–7302, 2021. 2

[61] Zhitong Xiong, Haopeng Li, and Xiao Xiang Zhu. Doubly deformable aggregation of covariance matrices for few-shot segmentation. In *Proceedings of the IEEE conference on European Conference on Computer Vision (ECCV)*, pages 133–150. Springer, 2022. 7, 2

[62] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757*, 2021. 3

[63] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. Mining latent classes for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8721–8730, 2021. 2

[64] Xianghui Yang, Bairun Wang, Kaige Chen, Xinchi Zhou, Shuai Yi, Wanli Ouyang, and Luping Zhou. Brinet: Towards bridging the intra-class and inter-class gaps in one-shot segmentation. *arXiv preprint arXiv:2008.06226*, 2020. 2

[65] Ze Yang, Ya-Li Wang, Xian-Yu Chen, Jian-Zhuang Liu, and Yu Qiao. Context-transformer: Tackling object confusion for few-shot detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 12653–12660, 2020. 2

[66] Bingfeng Zhang, Jimin Xiao, and Terry Qin. Self-guided and cross-guided learning for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8312–8321, 2021. 6, 2

[67] Bowen Zhang, Zhi Tian, Quan Tang, Xiangxiang Chu, Xiaolin Wei, Chunhua Shen, and Yifan Liu. Segvit: Semantic segmentation with plain vision transformers. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 5, 6

[68] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9587–9595, 2019. 2

[69] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5217–5226, 2019. 2, 6

[70] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, Shijian Lu, and Eric P Xing. Meta-detr: Image-level few-shot detection with inter-class correlation exploitation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 2

[71] Jian-Wei Zhang, Yifan Sun, Yi Yang, and Wei Chen. Feature-proxy transformer for few-shot segmentation. *arXiv preprint arXiv:2210.06908*, 2022. 2, 7

[72] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE transactions on cybernetics*, 50(9):3855–3865, 2020. 2

[73] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11175–11185, 2023. 2, 3, 4, 5, 6, 7, 8