Efficient Masked Attention Transformer for Few-Shot Classification and Segmentation

Dustin Carrión-Ojeda [1,2]   Stefan Roth [1,2]   Simone Schaub-Meyer [1,2]

[1]TU Darmstadt   [2]hessian.AI

https://visinf.github.io/emat

Abstract

*Few-shot classification and segmentation (FS-CS) focuses on jointly performing multi-label classification and multi-class segmentation using few annotated examples. Although the current state of the art (SOTA) achieves high accuracy in both tasks, it struggles with small objects. To overcome this, we propose the **E**fficient **M**asked **A**ttention **T**ransformer (EMAT), which improves classification and segmentation accuracy, especially for small objects. EMAT introduces three modifications: a novel memory-efficient masked attention mechanism, a learnable downscaling strategy, and parameter-efficiency enhancements. EMAT outperforms all FS-CS methods on the PASCAL-5^i and COCO-20^i datasets, using at least four times fewer trainable parameters.*

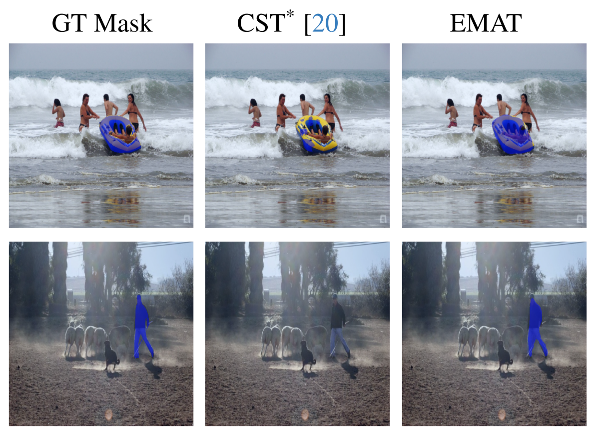

Figure 1. **Qualitative comparison of small objects** between the current SOTA FS-CS method (CST) [20] and our EMAT. CST* uses the same backbone as EMAT (*i.e.*, DINOv2 [31]). By processing high-resolution correlation tokens, EMAT preserves finer details, yielding more accurate segmentation masks.

1. Introduction

Recently, data-intensive methods have been introduced for various deep learning applications [5, 8, 22, 24, 31, 33, 40]. These methods rely on large training datasets, making them impractical in fields where collecting extensive datasets is challenging or costly [12, 13, 63]. Consequently, few-shot learning (FSL) methods have gained significant attention for their ability to learn from just a few examples and quickly adapt to new classes [1, 43, 50, 54]. In computer vision, FSL has been mostly applied to image classification (FS-C) [3, 17, 39, 42] and segmentation (FS-S) [10, 29, 53, 60, 61].

FS-C and FS-S often co-occur in real-world applications, *e.g.*, in agriculture, where crops must be segmented and classified by type or health status. Hence, recent works [18, 20] integrate multi-label classification and multi-class segmentation into a single few-shot classification and segmentation (FS-CS) task. While FS-CS addresses some limitations of FS-C (*e.g.*, assuming the query image contains only one class) and FS-S (*e.g.*, assuming the target class is always present in the query image), it also increases the task difficulty by simultaneously tackling classification and segmentation. Moreover, some applications, *e.g.*, medical imaging, rely on precise small-object analysis [12, 15, 63]. Thus, achieving high accuracy on small objects is a desired property for FS-CS methods. Yet, as shown in Fig. 1, the current state-of-the-art (SOTA) FS-CS method [20] struggles with small objects, a limitation we address in this work.

Contributions. *(1)* Building on the current SOTA FS-CS method [20], we propose an efficient masked attention transformer (EMAT), which enhances classification and segmentation accuracy, particularly for small objects, while using approximately four times fewer trainable parameters. *(2)* Our EMAT outperforms all FS-CS methods on the PASCAL-5^i and COCO-20^i datasets, supports the N-way K-shot configuration, and can generate empty segmentation masks when no target objects are present.

2. Related Work

Few-shot classification (FS-C) methods can be categorized into three groups based on what the model learns. *Representation-based* approaches learn class-agnostic, dis-

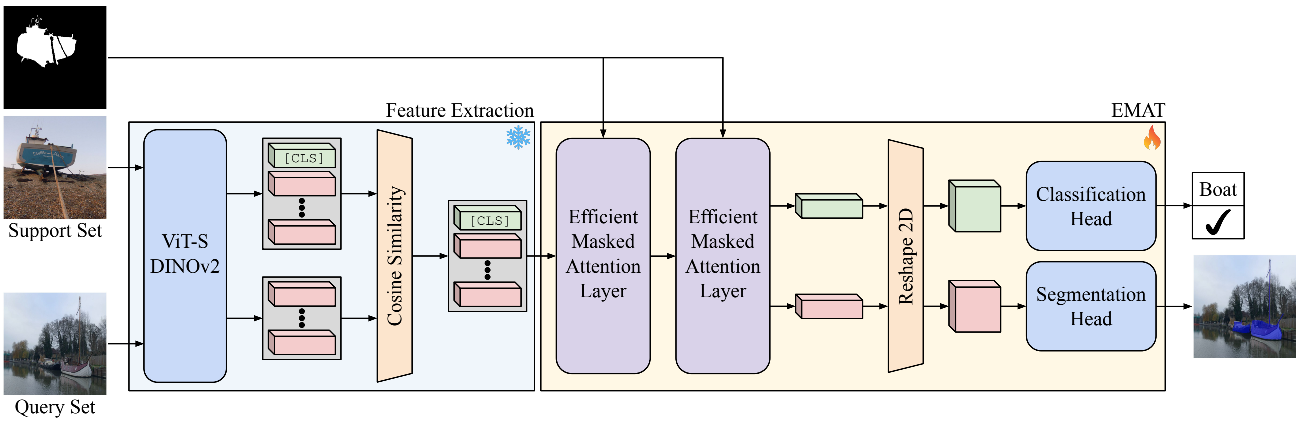

Figure 2. **FS-CS pipeline used by our EMAT.** A frozen, pre-trained ViT [11] extracts image and class tokens from support and query images, which are correlated via cosine similarity. The resulting correlation tokens are processed by a two-layer transformer equipped with our masked attention mechanism, learnable downscaling, and parameter-efficient design (see the supplementary material for details). Task-specific heads then predict the multi-label classification vector and multi-class segmentation mask.

criminative embeddings [3, 16, 19, 38, 45, 46, 58]. *Optimization-based* approaches learn the optimal set of weights that allow the model to adapt to new classes in just a few optimization steps [4, 14, 34, 39]. *Transfer-based* approaches adapt large pre-trained [6, 9, 23, 27, 42] or foundation models [17, 36, 62]. A major limitation of most FS-C methods is the assumption of a single label per image [2, 37], limiting them in multi-label settings.

Few-shot segmentation (FS-S) methods can also be categorized into three groups: *prototype matching*, which aligns support embeddings with query features [10, 25, 44, 48, 49, 56]; *dense correlation*, which constructs support–query correlation tensors [7, 28, 29, 32, 51, 52]; and *model-adaptation*, which fine-tunes large pre-trained models [26, 47, 55, 59, 60]. Despite the advancements in FS-S, most methods have two main limitations: *(1)* they target only the 1-way K-shot configuration and *(2)* they assume the query image contains the target class, preventing the models from predicting empty segmentation masks. Only a few recent works [41, 57] address the more general N-way K-shot configuration.

Few-shot classification and segmentation (FS-CS) focuses on jointly predicting the multi-label classification vector and multi-class mask without assuming support classes are present in the query image [18]. The current SOTA FS-CS method, CST [20], uses a memory-intensive masked-attention mechanism that requires significant downsampling of the correlation features, reducing its accuracy on small objects. In this work, we enhance CST by proposing an efficient masked-attention formulation and adding further refinements, resulting in a more memory- and parameter-efficient method with improved accuracy, especially for small objects.

3. Problem Definition

This work focuses on the FS-CS task [18], formulated as an N-way K-shot learning problem [46]. We assume two disjoint class sets: $\mathcal{C}_{\text{train}}$ for training and $\mathcal{C}_{\text{test}}$ for testing. Accordingly, training tasks are sampled from $\mathcal{C}_{\text{train}}$, and testing tasks from $\mathcal{C}_{\text{test}}$. Each task consists of a support set $\mathcal{S}$ and a query image $\mathbf{I}_q$, where $\mathcal{S}$ contains N classes $\mathcal{C}_{\text{s}}$ ($\mathcal{C}_{\text{s}} \subseteq \mathcal{C}_{\text{train}}$ or $\mathcal{C}_{\text{s}} \subseteq \mathcal{C}_{\text{test}}$), each represented by K examples:

$$\mathcal{S} = \left\{ \left\{ (\mathbf{I}_j^i, \mathbf{M}_j^i, i) \mid i \in \mathcal{C}_{\text{s}} \right\}_j^K \right\}_i^N , \tag{1}$$

where $\mathbf{I}_j^i$, $\mathbf{M}_j^i$, and i denote the support image, segmentation mask, and class label for the j^{th} example of the i^{th} class.

The goal of FS-CS is to learn from $\mathcal{S}$ so that, given $\mathbf{I}_q$, the model can *(i)* identify which support classes appear (multi-label classification), and *(ii)* segment them (multi-class segmentation). Moreover, FS-CS allows $\mathbf{I}_q$ to contain a subset of the support classes. Thus, when $N > 1$, $\mathbf{I}_q$ can contain: *(1)* none of the support classes, *(2)* a subset of them, or *(3)* all support classes. Note that case *(1)* requires models to predict empty segmentation masks when necessary.

4. Efficient Masked Attention Transformer

Fig. 2 illustrates the pipeline used by our proposed EMAT, which builds upon CST [20]. Both methods share the same feature extraction process: support and query images $\mathbf{I}_j^i, \mathbf{I}_q \in \mathbb{R}^{H \times W \times 3}$ are processed by a frozen, pre-trained ViT [11] with patch size p, producing support and query image tokens $\mathbf{T}_{s_{\text{i}}}, \mathbf{T}_{q_{\text{i}}} \in \mathbb{R}^{h \times w \times d}$, and a support class token $\mathbf{T}_{s_{\text{c}}} \in \mathbb{R}^{1 \times d}$, where $h = H/p$, $w = W/p$, and d is the token dimension of a single ViT head. The support tokens $\mathbf{T}_{s_{\text{i}}}$ are downsampled via bilinear interpolation and reshaped to $\mathbf{T}_{s_{\text{i}}}^f \in \mathbb{R}^{(h' \cdot w') \times d}$. Similarly, query image tokens $\mathbf{T}_{q_{\text{i}}}$ are reshaped to $\mathbf{T}_{q_{\text{i}}}^f \in \mathbb{R}^{(h \cdot w) \times d}$. Next, $\mathbf{T}_{s_{\text{i}}}^f$ and

$\mathbf{T}_{s_c}$ are concatenated to form $\mathbf{T}_s^c$. Finally, cosine similarity between $\mathbf{T}_s^c$ and $\mathbf{T}_{q_i}^f$ is computed across all ViT layers l and attention heads g, resulting in the correlation tokens $\mathbf{C} \in \mathbb{R}^{t_s \times t_q \times (l \cdot g)}$, where $t_s = h' \cdot w' + 1$ and $t_q = h \cdot w$.

EMAT differs from CST in its two-layer transformer that processes correlation tokens and feeds task-specific heads for multi-label classification and multi-class segmentation. We enhance this transformer with three key improvements. *(1)* A memory-efficient masked attention formulation:

$$\mathbf{O}_{ijk} = \sum_{p\varnothing} \left[\sigma\left(\mathbf{Q}^d_{ijk} \cdot (\mathbf{K}_{:jk} \oslash \mathbf{M}^f_{:})\right)\right]_{p\varnothing} \odot \left(\mathbf{V}_{:jk} \oslash \mathbf{M}^f_{:}\right)_{p\varnothing}, \quad (2)$$

where $\mathbf{Q}^d, \mathbf{K}, \mathbf{V}$ are the downscaled query, key, and value matrices, and $\mathbf{M}^f$ is the resized, flattened segmentation mask; $i \in \{1, \ldots, h''{\cdot}w''{+}1\}$, $j \in \{1, \ldots, t_q\}$, $k \in \{1, \ldots, e\}$ with e denoting the embedding size. The operators σ, $\odot$, and $\oslash$ denote softmax, element-wise multiplication, and our element-wise masking operator:

$$(\mathbf{Z}_{pjk} \oslash \mathbf{M}^f_p) = \begin{cases} \mathbf{Z}_{pjk} & \text{if } \mathbf{M}^f_p = 1, \\ \varnothing & \text{otherwise,} \end{cases} \quad \forall p \in \{1, \ldots, t_s\}, \quad (3)$$

with $p \in \{1, \ldots, t_s\}$ and $\varnothing$ indicating exclusion of the entry. By excluding masked-out tokens, EMAT supports much higher-resolution inputs than CST. *(2)* A learnable downscaling strategy that combines small convolutions with average pooling, avoiding large pooling kernels. *(3)* A reduction in the number of channels across attention layers and task-specific heads to improve parameter efficiency and mitigate overfitting. Further details of these three improvements are provided in the supplementary material.

Following CST, EMAT is trained using the 1-way 1-shot configuration. Since EMAT uses task-specific heads, it is trained with two losses:

$$\mathcal{L}_{\text{clf}} = -y \log \widehat{y}, \quad (4)$$

$$\mathcal{L}_{\text{seg}} = -\frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{M}_{ij} \log \widehat{\mathbf{M}}_{ij}, \quad (5)$$

where $y \in \{0, 1\}$ and $\mathbf{M}_{ij} \in \{0, 1\}$ are the ground-truth classification and segmentation labels, and $\widehat{y}$, $\widehat{\mathbf{M}}_{ij}$ are the corresponding predictions. The final loss function jointly optimizes both losses using a balancing hyperparameter λ:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{clf}} + \mathcal{L}_{\text{seg}}. \quad (6)$$

Inference on N-way K-shot tasks is performed as in CST [20], by treating each class as an independent 1-way K-shot task: class-wise logits and masks are averaged over the K examples, producing N predictions. Logits above a threshold δ=0.5 form the multi-label vector, and $\widehat{\mathbf{M}}_{ij}$ is assigned to the class with the highest score, or to background if all scores fall below δ, thereby allowing empty masks.

Dataset	Method	Train. Params.	1-way 1-shot Acc.	1-way 1-shot mIoU	2-way 1-shot Acc.	2-way 1-shot mIoU
PASCAL-5^i	PANet [48]	23.51	68.70	36.14	56.53	37.20
	PFENet [44]	31.96	74.38	43.08	39.35	35.57
	HSNet [28]	2.57	83.60	49.62	67.27	44.85
	ASNet [18]	1.32	84.85	52.32	68.30	47.87
	CST [20]	<u>0.37</u>	85.72	55.52	70.37	53.78
	CST*	<u>0.37</u>	<u>90.62</u>	<u>64.40</u>	<u>80.58</u>	<u>63.28</u>
	EMAT	**0.09**	**91.25**	**64.64**	**82.70**	**63.38**
COCO-20^i	PANet [48]	23.51	66.62	25.16	51.30	23.64
	PFENet [44]	31.96	71.40	31.86	36.45	23.37
	HSNet [28]	2.57	76.95	34.33	62.43	30.58
	ASNet [18]	1.32	78.60	35.82	63.05	31.62
	CST [20]	<u>0.37</u>	80.53	38.28	64.02	36.23
	CST*	<u>0.37</u>	<u>88.50</u>	<u>53.48</u>	<u>78.70</u>	<u>51.47</u>
	EMAT	**0.09**	**88.70**	**54.76**	**80.07**	**52.81**

Table 1. **Comparison of FS-CS methods** on PASCAL-5^i and COCO-20^i across different task configurations. CST* and EMAT were trained and evaluated, while other methods were only evaluated using the checkpoints from [18]. CST* uses the same backbone as EMAT (*i.e.*, DINOv2 [31]). All values, except the number of trainable parameters (in millions), are percentages (higher is better). Highlight indicates our proposed method. **Bold** and <u>underlined</u> values indicate the best and second best results.

t_s^l per Layer	Method	ME	LD	PE	Mem. Usage	Train. Params.	Acc.	mIoU
t_s^1=145 t_s^2=10	CST*	–	–	–	**8.68**	<u>366.00</u>	80.58	63.28
t_s^1=401 t_s^2=101	CST*	–	–	–	≈ 63	<u>366.00</u>	N/A	N/A
	EMAT	✓	–	–	36.92	<u>366.00</u>	81.95	62.97
	EMAT	✓	✓	–	<u>36.53</u>	404.48	<u>82.17</u>	<u>63.36</u>
	EMAT	✓	✓	✓	38.31	**86.02**	**82.70**	**63.38**

Table 2. **Ablation study** on PASCAL-5^i using 2-way 1-shot tasks. "t_s^l" indicates the value of t_s for each layer $l \in \{1, 2\}$. The memory efficiency (ME), learnable downscaling (LD), and parameter efficiency (PE) columns correspond to the modifications of EMAT described in Sec. 4. "Mem. Usage" reports the average per-GPU memory used during training. CST* uses the same backbone as EMAT (*i.e.*, DINOv2 [31]). **(Top)** CST* with its original support dimension per layer t_s^l. **(Bottom)** successive modifications introduced by EMAT. Highlight indicates our complete EMAT. **Bold** and <u>underlined</u> values indicate the best and second best results.

5. Experiments

Datasets. We evaluated our EMAT on the widely used PASCAL-5^i [35] and COCO-20^i [30] datasets. Although they were designed for few-shot segmentation, both can also be used for few-shot classification and segmentation [18]. PASCAL-5^i comprises 20 classes and COCO-20^i 80 classes, each partitioned into four non-overlapping folds.
Implementation details. EMAT uses a frozen ViT-S encoder [11] pre-trained with DINOv2 [31]. The two-layer transformer uses our memory-efficient masked attention with 8 heads. We train for 80 epochs with a batch size of 9 using Adam [21] with learning rate 10^{-3}. Following

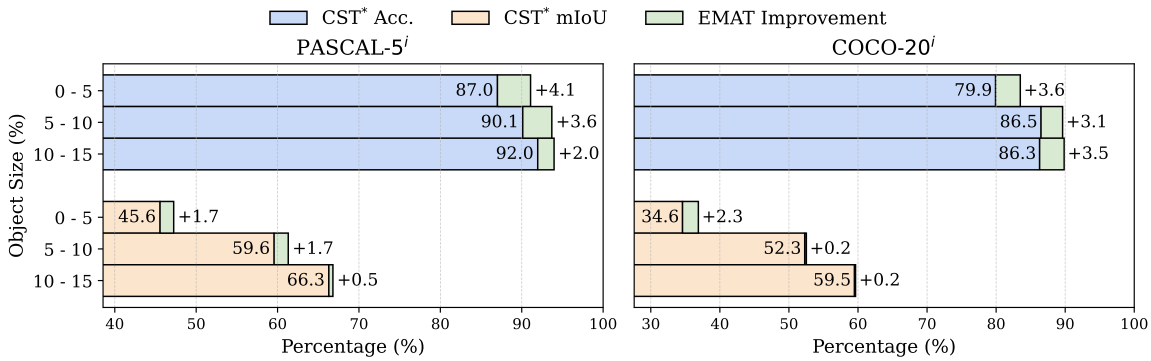

Figure 3. **Analysis of small objects** on PASCAL-5^i and COCO-20^i. Each bar represents the average across the four folds of each dataset, filtered by object size, using 1-way 1-shot tasks. To enable a more controlled analysis, we modified the setting described in Sec. 3 to ensure that the query image always contain the class of the support image. CST* uses the same backbone as EMAT (*i.e*., DINOv2 [31]).

[20], we use 1-way 1-shot tasks and set the loss weight λ in Eq. (6) to 0.1. Moreover, we re-train CST [20] with the same DINOv2 backbone used by EMAT and denote it as CST*. All training was conducted on three NVIDIA RTX A6000 GPUs, with evaluation performed on a single GPU.

5.1. Comparison to SOTA FS-CS

To evaluate the effectiveness of our EMAT, we compare it with CST [20] and other SOTA FS-CS methods. Tab. 1 shows mean classification accuracy (Acc.) and mean IoU (mIoU) over the four folds of PASCAL-5^i [35] and COCO-20^i [30]. Although DINOv2 pre-training [31] already significantly improves CST* over its original version, EMAT consistently outperforms all methods. These results validate the benefit of processing higher-resolution correlation tokens enabled by our memory-efficient masked attention (see Sec. 4). Moreover, EMAT requires at least four times fewer parameters than CST, making it the most parameter-efficient method among SOTA FS-CS models.

5.2. Analysis of Small Objects

To analyze the impact of higher-resolution correlation tokens on small objects, we filter each fold of PASCAL-5^i and COCO-20^i based on object size, creating three splits: objects occupying 0–5 %, 5–10 %, and 10–15 % of the image. Fig. 3 shows the average accuracy and mIoU of CST* and the corresponding improvement achieved by EMAT across the three splits for both datasets. The results indicate that accuracy and mIoU increase with the object size, and EMAT provides the largest improvement over CST* for the smallest objects, gradually decreasing as object size increases. The enhanced classification and segmentation accuracy of EMAT is likely due to improved localization enabled by the increased resolution of the correlation tokens.

5.3. Ablation Study

Tab. 2 reports the results of CST* using its original support dimension per layer t_s^l. For fair comparison, we increased the t_s^l of CST* to use the same as EMAT, but it required about 63 GB of GPU memory, which exceeded the 48 GB capacity of our GPUs. For EMAT we progressively integrated the improvements described in Sec. 4: *(1)* memory-efficient masked attention, *(2)* learnable downscaling of the query matrix, and *(3)* parameter-efficiency modifications.

Adding our memory-efficient masked attention alone reduces memory usage by 26 GB ($\approx$ 41 %) and yields an absolute accuracy gain of +1.37 %, but it slightly lowers mIoU, likely because the model relies on large pooling windows for processing the higher-resolution correlation tokens. Incorporating our learnable downscaling removes these large windows and yields absolute gains of +1.59 % in accuracy and +0.08 % in mIoU over CST*. Since learnable downscaling increases the number of trainable parameters, we next apply our parameter-efficiency modifications that remove 318 K parameters ($\approx$ 79 %), while still saving about 39 % of the memory CST* would require for the same t_s^l as EMAT. These modifications also improve accuracy by +2.12 % and mIoU by +0.1 % compared to CST*.

6. Conclusion

In this work, we propose EMAT, an enhancement over CST, the state-of-the-art method for few-shot classification and segmentation (FS-CS). EMAT incorporates our novel memory-efficient masked attention mechanism that allows our model to process high-resolution correlation tokens while maintaining memory efficiency. Our learnable downscaling strategy and additional parameter-efficiency refinements enhance the classification and segmentation accuracy of EMAT while improving its parameter efficiency. Our results demonstrate that EMAT consistently outperforms all FS-CS methods across different task configurations while requiring at least four times fewer trainable parameters. Moreover, our qualitative results highlight that EMAT captures finer details more accurately, improving accuracy when dealing with small objects.

Acknowledgments. This work was funded by the Hessian Ministry of Science and Research, Arts and Culture (HMWK) through the project "The Third Wave of Artificial Intelligence – 3AI". The work was further supported by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) under Germany's Excellence Strategy (EXC 3057/1 "Reasonable Artificial Intelligence", Project No. 533677015). Stefan Roth acknowledges support by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 866008).

References

[1] Pranjal Aggarwal, Ameet Deshpande, and Karthik R. Narasimhan. SemSup-XC: Semantic supervision for zero and few-shot extreme classification. In *ICML*, pages 228–247, 2023. 1

[2] Amit Alfassy, Leonid Karlinsky, Amit Aides, Joseph Shtok, Sivan Harary, Rogério Schmidt Feris, Raja Giryes, and Alexander M. Bronstein. LaSO: Label-set operations networks for multi-label few-shot learning. In *CVPR*, pages 6548–6557, 2019. 2

[3] Kelsey R. Allen, Evan Shelhamer, Hanul Shin, and Joshua B. Tenenbaum. Infinite mixture prototypes for few-shot learning. In *ICML*, pages 232–241, 2019. 1, 2

[4] Antreas Antoniou, Harrison Edwards, and Amos J. Storkey. How to train your MAML. In *ICLR*, 2019. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 1

[6] Dustin Carrión-Ojeda, Mahbubul Alam, Sergio Escalera, et al. NeurIPS'22 Cross-Domain MetaDL Challenge: Results and lessons learned. In *NeurIPS Competition Track*, pages 50–72, 2022. 2

[7] Hao Chen, Yonghan Dong, Zheming Lu, Yunlong Yu, and Jungong Han. Pixel matching network for cross-domain few-shot segmentation. In *WACV*, pages 978–987, 2024. 2

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 1

[9] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *ICLR*, 2020. 2

[10] Nanqing Dong and Eric P. Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, page 79, 2018. 1, 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 3

[12] Xin Fan, Xiaolin Wang, Jiaxin Gao, Jia Wang, Zhongxuan Luo, and Risheng Liu. Bi-level learning of task-specific decoders for joint registration and one-shot medical image segmentation. In *CVPR*, pages 11726–11735, 2024. 1

[13] Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, and Jimin Xiao. FastRecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *ICCV*, pages 17481–17490, 2023. 1

[14] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135, 2017. 2

[15] Xuan Gong, Xin Xia, Wentao Zhu, Baochang Zhang, David Doermann, and Li'an Zhuo. Deformable Gabor feature networks for biomedical image classification. In *WACV*, pages 4004–4012, 2021. 1

[16] Fusheng Hao, Fengxiang He, Liu Liu, Fuxiang Wu, Dacheng Tao, and Jun Cheng. Class-aware patch embedding adaptation for few-shot image classification. In *ICCV*, pages 18905–18915, 2023. 2

[17] Jonas Herzog. Adapt before comparison: A new perspective on cross-domain few-shot segmentation. In *CVPR*, pages 23605–23615, 2024. 1, 2

[18] Dahyun Kang and Minsu Cho. Integrative few-shot learning for classification and segmentation. In *CVPR*, pages 9979–9990, 2022. 1, 2, 3

[19] Dahyun Kang, Heeseung Kwon, Juhong Min, and Minsu Cho. Relational embedding for few-shot classification. In *ICCV*, pages 8822–8833, 2021. 2

[20] Dahyun Kang, Piotr Koniusz, Minsu Cho, and Naila Murray. Distilling self-supervised vision transformers for weakly-supervised few-shot classification & segmentation. In *CVPR*, pages 19627–19638, 2023. 1, 2, 3, 4

[21] Diederick P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 3

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, pages 4015–4026, 2023. 1

[23] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Cross-domain few-shot learning with task-specific adapters. In *CVPR*, pages 7161–7170, 2022. 2

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1

[25] 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 *CVPR*, pages 11553–11562, 2022. 2

[26] Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. In *ICLR*, 2024. 2

[27] Tianyi Ma, Yifan Sun, Zongxin Yang, and Yi Yang. ProD: Prompting-to-disentangle domain knowledge for cross-domain few-shot image classification. In *CVPR*, pages 19754–19763, 2023. 2

[28] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *ICCV*, pages 6941–6952, 2021. 2, 3

[29] Seonghyeon Moon, Samuel S. Sohn, Honglu Zhou, Sejong Yoon, Vladimir Pavlovic, Muhammad Haris Khan, and Mubbasir Kapadia. MSI: Maximize support-set information for few-shot segmentation. In *ICCV*, pages 19266–19276, 2023. 1, 2

[30] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *ICCV*, 2019. 3, 4

[31] Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024. 1, 3, 4

[32] Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical dense correlation distillation for few-shot segmentation. In *CVPR*, pages 23641–23651, 2023. 2

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1

[34] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? Towards understanding the effectiveness of MAML. In *ICLR*, 2020. 2

[35] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *BMVC*, 2017. 3, 4

[36] Julio Silva-Rodríguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *CVPR*, pages 23681–23690, 2024. 2

[37] Christian Simon, Piotr Koniusz, and Mehrtash Harandi. Meta-learning for multi-label few-shot classification. In *WACV*, pages 346–355, 2022. 2

[38] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *NIPS*, pages 4077–4087, 2017. 2

[39] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *CVPR*, 2019. 1, 2

[40] Google Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv:2312.11805 [cs.CL]*, 2023. 1

[41] Pinzhuo Tian, Zhangkai Wu, Lei Qi, Lei Wang, Yinghuan Shi, and Yang Gao. Differentiable meta-learning model for few-shot semantic segmentation. In *AAAI*, pages 12087–12094, 2020. 2

[42] 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 *ECCV*, pages 266–282. Springer, 2020. 1, 2

[43] Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. In *CVPR*, pages 11563–11572, 2022. 1

[44] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE T. Pattern Anal. Mach. Intell.*, 44(2):1050–1065, 2022. 2, 3

[45] Ihsan Ullah, Dustin Carrión-Ojeda, Sergio Escalera, Isabelle Guyon, Mike Huisman, Felix Mohr, Jan N. van Rijn, Haozhe Sun, Joaquin Vanschoren, and Phan Anh Vu. Meta-Album: Multi-domain meta-dataset for few-shot image classification. In *NeurIPS*, pages 3232–3247, 2022. 2

[46] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NIPS*, pages 3630–3638, 2016. 2

[47] Jin Wang, Bingfeng Zhang, Jian Pang, Honglong Chen, and Weifeng Liu. Rethinking prior information generation with CLIP for few-shot segmentation. In *CVPR*, pages 3941–3951, 2024. 2

[48] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. PANet: Few-shot image semantic segmentation with prototype alignment. In *ICCV*, 2019. 2, 3

[49] Yuan Wang, Naisong Luo, and Tianzhu Zhang. Focus on query: Adversarial mining transformer for few-shot segmentation. In *NeurIPS*, pages 31524–31542, 2023. 2

[50] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Universal-prototype enhancing for few-shot object detection. In *ICCV*, pages 9567–9576, 2021. 1

[51] Guo-Sen Xie, Huan Xiong, Jie Liu, Yazhou Yao, and Ling Shao. Few-shot semantic segmentation with cyclic memory network. In *ICCV*, pages 7293–7302, 2021. 2

[52] Qianxiong Xu, Wenting Zhao, Guosheng Lin, and Cheng Long. Self-calibrated cross attention network for few-shot segmentation. In *ICCV*, pages 655–665, 2023. 2

[53] Yong Yang, Qiong Chen, Yuan Feng, and Tianlin Huang. MIANet: Aggregating unbiased instance and general information for few-shot semantic segmentation. In *CVPR*, pages 7131–7140, 2023. 1

[54] Chuangguan Ye, Hongyuan Zhu, Yongbin Liao, Yanggang Zhang, Tao Chen, and Jiayuan Fan. What makes for effective few-shot point cloud classification? In *WACV*, pages 1829–1838, 2022. 1

[55] Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi Liu, and Yunchao Wei. Bridge the points: Graph-based few-shot segment anything semantically. In *NeurIPS*, 2024. 2

[56] Bingfeng Zhang, Jimin Xiao, and Terry Qin. Self-guided and cross-guided learning for few-shot segmentation. In *CVPR*, pages 8312–8321, 2021. 2

[57] Miao Zhang, Miaojing Shi, and Li Li. MFNet: Multiclass few-shot segmentation network with pixel-wise metric learning. *IEEE T. Circuits Syst. Video Tech.*, 32(12):8586–8598, 2022. 2

[58] Fei Zhou, Peng Wang, Lei Zhang, Wei Wei, and Yanning Zhang. Revisiting prototypical network for cross domain few-shot learning. In *CVPR*, pages 20061–20070, 2023. 2

[59] Ziqin Zhou, Hai-Ming Xu, Yangyang Shu, and Lingqiao Liu. Unlocking the potential of pre-trained vision transformers for few-shot semantic segmentation through relationship descriptors. In *CVPR*, pages 3817–3827, 2024. 2

[60] Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. LLaFS: When large language models meet few-shot segmentation. In *CVPR*, pages 3065–3075, 2024. 1, 2

[61] Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, and Jun Liu. Addressing background context bias in few-shot segmentation through iterative modulation. In *CVPR*, pages 3370–3379, 2024. 1

[62] Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In *ICCV*, pages 2605–2615, 2023. 2

[63] Yazhou Zhu, Shidong Wang, Tong Xin, and Haofeng Zhang. Few-shot medical image segmentation via a region-enhanced prototypical transformer. In *MICCAI*, pages 271–280. Springer, 2023. 1