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] TECHNISCHE UNIVERSITÄT DARMSTADT [2] hessian.AI

Project Page

TL;DR

EMAT, **E**fficient **M**asked **A**ttention **T**ransformer, processes high-resolution correlation tokens, boosting few-shot classification and segmentation, especially for small objects, while using at least four times fewer parameters than existing methods. It supports *N*-way *K*-shot tasks and outputs empty masks when no target is present.

Figure 1. **Qualitative comparison** of CST* (previous SOTA) and EMAT.

Introduction

Few-shot classification and segmentation (FS-CS) [2] focuses on jointly performing *multi-label classification* and *multi-class segmentation* using few annotated examples.

Motivation:

- FS-C and FS-S often co-occur in real-world applications.
- Applications such as medical imaging require precise small-object analysis, yet the current SOTA in FS-CS, CST [1], performs poorly on small objects.
- Most FS-S methods are limited to single-class (1-way) segmentation, and the standard multi-class (*N*-way) evaluation setting discards useful annotations.

Goal: Enhance efficiency and FS-CS accuracy, particularly for small objects, and better utilize annotations during evaluation.

Problem Definition

In *N*-way *K*-shot FS-CS, a **support set** provides *N* classes with *K* examples each, and the goal is to identify which support classes appear in the **query image** (multi-label classification) and segment them (multi-class segmentation). Here, unlike the standard definition, the query image may contain: *(1)* none, *(2)* a subset, or *(3)* all of the support classes.

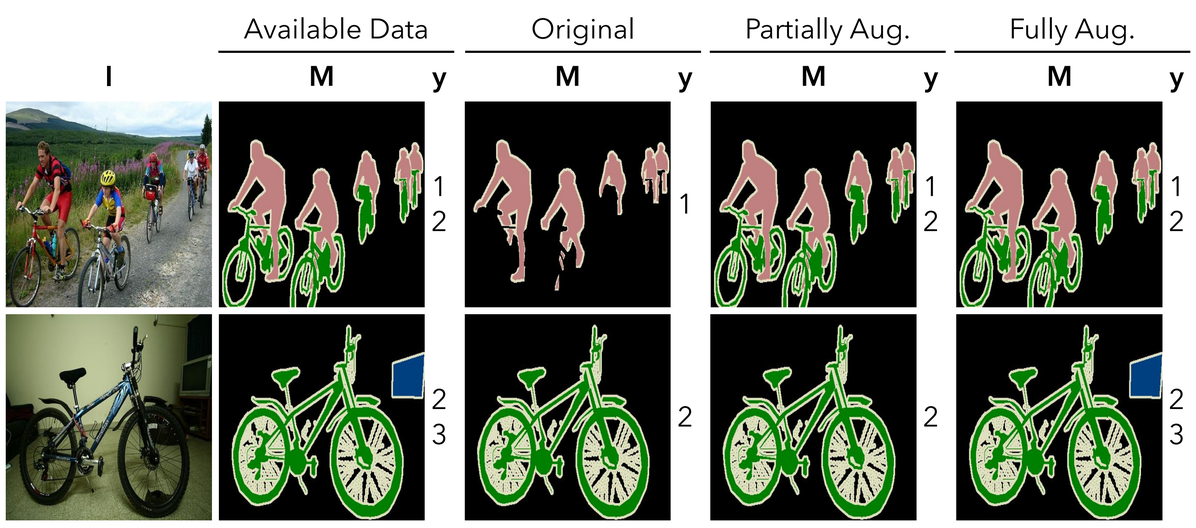

Figure 2. **Support set** of a 2-way 1-shot task across all evaluation settings.

Method

EMAT processes correlation features extracted with a frozen pre-trained ViT-S [3] using a two-layer transformer with:

(1) A memory-efficient masked attention formulation,

(2) A learnable downscaling strategy,

(3) Modifications for improved parameter efficiency.

Task-specific heads then predict the multi-label classification vector and the multi-class segmentation mask.

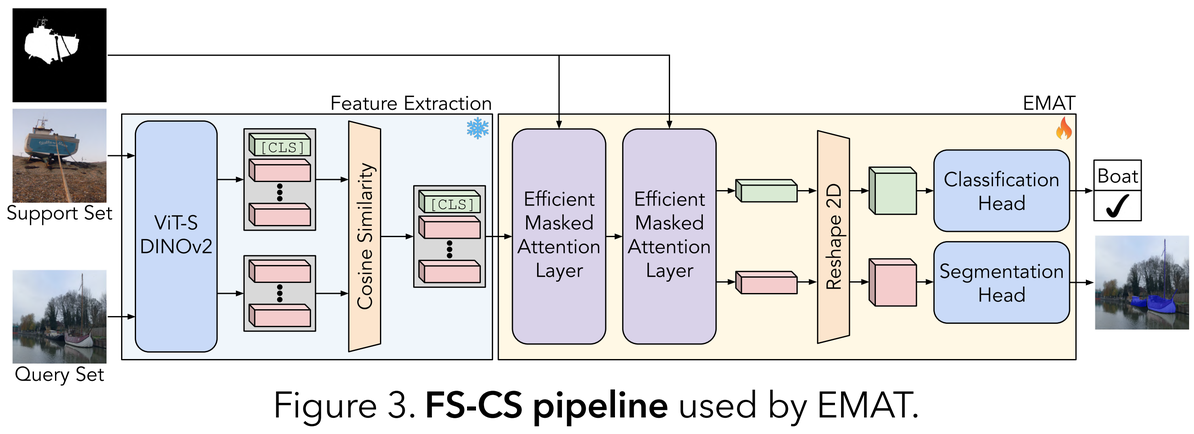

Figure 3. **FS-CS pipeline** used by EMAT.

Results

Table 1. **Comparison of FS-CS methods** on 2-way 1-shot tasks across all evaluation settings using COCO-20^i [4].

Method	Trainable Params.	Original		Partially Aug.		Fully Aug.	
		Acc.	mIoU	Acc.	mIoU	Acc.	mIoU
PANet [5]	23.51	51.30	23.64	51.32	23.78	45.07	23.17
HSNet [6]	2.57	62.43	30.58	62.40	30.66	55.15	29.44
ASNet [2]	1.32	63.05	31.62	63.03	31.64	55.47	30.47
CST* [1]	<u>0.37</u>	<u>78.70</u>	<u>51.47</u>	<u>78.87</u>	<u>51.53</u>	<u>71.18</u>	<u>50.76</u>
EMAT	**0.09**	**80.07**	**52.81**	**80.25**	**52.82**	**73.00**	**51.99**

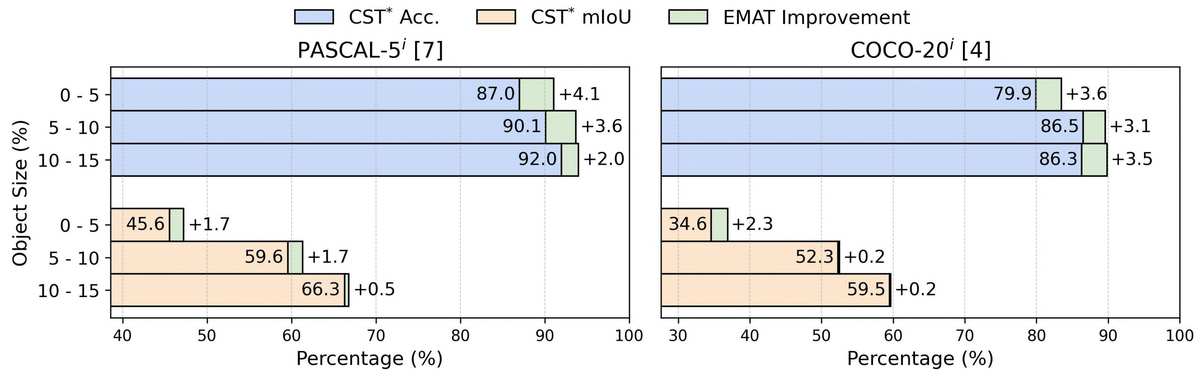

Figure 4. **Small-object analysis** with 1-way 1-shot tasks.

Conclusion

- EMAT achieves SOTA performance with ~4× fewer parameters.
- High-resolution tokens boost accuracy on small-objects.
- Our evaluation settings maximize annotation usage.

References & Acknowledgements

[1] Dahyun Kang et al. Distilling self-supervised vision transformers for weakly-supervised FS-CS. In *CVPR*, 2023.
[2] Dahyun Kang et al. Integrative few-shot learning for classification and segmentation. In *CVPR*, 2022.
[3] Maxime Oquab et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024.
[4] Khoi Nguyen et al. Feature weighting and boosting for few-shot segmentation. In *ICCV*, 2019.
[5] Kaixin Wang et al. PANet: Few-shot image semantic segmentation with prototype alignment. In *ICCV*, 2019.
[6] Juhong Min et al. Hypercorrelation squeeze for few-shot segmentation. In *ICCV*, 2021.
[7] Amirreza Shaban et al. One-shot learning for semantic segmentation. In *BMVC*, 2017.

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).

Code: github.com/visinf/emat