Text Augmented Correlation Transformer For Few-shot Classification & Segmentation

Supplementary Material

7. More Implementation Details

We provide a comprehensive description of the implementation, including the hyperparameters utilized in training. Both the vision and text encoder branches are derived from CLIP-B/16, each comprising 12 transformer layers. The text encoder operates with six attention heads and a hidden dimension of 512, while the vision encoder uses eight attention heads with a hidden dimension of 768. The CLIP encoders remain frozen throughout the process, functioning solely for feature extraction. The total number of trainable parameters amounts to 3M, significantly exceeding CST's 0.4M but still minimal compared to the 86M parameters in the CLIP-B/16 vision encoder. By default, the classification loss scaling parameter λ is set to 0.1, with the temperature for the softmax cross-entropy loss fixed at $\tau = 0.1$. The learning rate is maintained at 1×10^{-3}.

The evaluation datasets include Pascal-5^i and COCO-20^i, each divided into four validation folds, where the classes in each fold are mutually exclusive. The sampling strategy adheres to the approach described in [12], with random generation for training and validation samples. A fixed random seed of 0 is used for validation to ensure consistency in the generated test episodes. For training, we adopt a 1-way 1-shot configuration, where each episode contains a single query image and one support image corresponding to one class. During validation, episodes are sampled in an N-way K-shot configuration, with the support set containing K images for each of the N classes. For images with multiple classes, only the N specified classes are treated as the foreground, with all others considered background. Metrics for evaluation follow the 0/1 exact match ratio for classification and mean Intersection over Union (mIoU) for segmentation, as outlined in [12].

For FS-CS training in the text-only setting on Pascal-5^i, the zero-shot distillation loss parameter κ is set to 0.1 for all folds except fold-2, where it is adjusted to 1. For COCO-20^i, κ is set to 0.01. In the text-augmented training configuration, the vision-text and vision-only correlation transformers are initialized with weights pre-trained on support text and images, respectively. These weights are specific to individual folds, such that initialization for fold-0 uses pre-training data from fold-0 alone. For the Pascal-5^i FS-CS text-augmented setting, the vision-only and vision-text correlation transformers are frozen, and only the scaling parameters α_1, α_2, β_1, and β_2 are optimized.

7.1. Results of N-way 1-shot FS-CS

We present extended results for the N-way 1-shot configuration, as depicted in Figure 4 and detailed in Table 14. As the value of N increases, a decline in performance is observed across all models. However, our model consistently demonstrates superior performance compared to others. This improvement stems from leveraging predictions derived from both text and image modalities. The integration of multimodal cues plays a pivotal role in enhancing classification and segmentation predictions, particularly across varying N values in the N-way 1-shot FS-CS setting.

7.2. Effect of different architectures on FS-CS

We evaluate CST [12] using various architectures, including ViT-B/16 (DINO), ViT-S/16 (DINO), ViT-S/8 (DINO), and CLIP-B/16, to examine how architectural choices impact FS-CS performance. Our analysis reveals that the pre-training methodology significantly influences performance in the FS-CS task. Additionally, smaller patch sizes are found to be more advantageous, as evidenced by DINO (ViT-S) with a patch size of 8 outperforming its counterpart with a patch size of 16. Notably, CLIP-B/32 experiences a substantial performance drop in fold 2, particularly in classification tasks, leading to the selection of CLIP-B/16 as the preferred backbone. Overall, CLIP, which leverages vision-language alignment during training, demonstrates superior performance compared to DINO for FS-CS tasks.

7.3. Does quality of text prompt effect FS-CS

Coop [43] emphasizes the critical role of prompt quality when evaluating CLIP in a zero-shot context. To determine whether the quality of handcrafted prompts impacts FS-CS performance, we conduct an analysis summarized in Table 10 across four folds. We test five different prompts, training each for 20 epochs on all Pascal-5^i folds. When compared to our default prompt, "This is a photo of {class-label}," the alternative prompts neither significantly enhance performance nor reduce variability. These findings suggest that handcrafted prompts may have limited utility in improving FS-CS results. Further exploration into creating more detailed or adaptive learnable prompts is left for future work

7.4. Results of 1-way 5-shot, 2-way 5-shot FS-CS

We conduct experiments on Pascal-5^i using the 1-way 5-shot and 2-way 5-shot configurations. Our analysis suggests

Method	Pretraining	class. 0/1 exact ratio(%)					segmentation mIoU(%)				
		5^0	5^1	5^2	5^3	Avg	5^0	5^1	5^2	5^3	Avg
ViT-S/8	DINO	86.9	88.0	81.5	86.5	85.7	55.6	61.6	47.7	56.9	55.5
ViT-S/16	DINO	83.9	87.0	76.6	82.7	82.6	58.5	58.2	43.5	50.6	52.7
ViT-B/16	DINO	86.4	87.4	79.2	83.8	84.2	60.0	62.4	43.7	57.6	55.9
ViT-B/32	CLIP	90.1	91.3	76.0	87.5	86.2	60.6	63.8	44.3	57.2	56.5
ViT-B/16	CLIP	93.6	93.4	85.3	89.5	90.4	67.3	69.8	49.2	63.4	62.4

Table 9. Performance of our method across different architectures on Pascal-5^i.

Text prompts	1-way-1-shot		2-way-1-shot	
	cls. ER	seg. mIoU	cls. ER	seg. mIoU
This is a photo of a {CLASS}	85.4	56.9	74.4	56.8
a sculpture of a {CLASS}.	83.0	56.2	70.8	56.8
a drawing of a {CLASS}.	85.3	56.6	75.0	57.5
graffiti of a {CLASS}.	83.8	56.3	72.0	58.5
a tattoo of a {CLASS}	83.2	53.7	71.5	54.0

Table 10. Effect of various text prompts on classification and segmentation performance in Pascal-5^i.

that the inclusion of text does not significantly boost performance in the 5-shot setting, as only a single text instance is available compared to five image instances. The scaling factors α_1, α_2, β_1, and β_2, which are optimized for text predictions in the 1-shot setting, are not directly applicable to the 5-shot scenario. To address this, we manually adjust the scaling factors α_1 and β_2 for text classification and segmentation predictions to smaller values. Text-only evaluation is infeasible in this setting due to the presence of only one text instance. A comparative analysis, including CST [12] with CLIP-B/16 as the encoder, is presented in Table 11.

7.5. Dual Correlation Block vs Single Correlation Block

We train our model in a text-augmented setting, comparing two configurations: one employing a dual correlation transformer block and another using a single correlation transformer for both vision-only and vision-text correlations. Both models are trained for 50 epochs. Our findings indicate that the dual correlation transformer outperforms the single correlation transformer. This improvement can be attributed to the separation of vision-text and vision-only correlations, which allows each to specialize in handling distinct modality interactions. The detailed results are presented in Table 12.

7.6. Few-shot Segmentation Results

We assess the performance of our framework on Few-shot Segmentation (FSS) under the assumption that the query and support sets share at least one common class, using the COCO-20^i dataset. The results, shown in Table 13, are based on models trained for 50 epochs. Additionally, we include the performance of CST [12] with CLIP-B/16 as the backbone. When evaluated with both our text-based and text-augmented approaches, our framework significantly outperforms CST. These findings highlight the critical role of text in providing rich semantic cues for few-shot tasks like segmentation, demonstrating its potential to enhance performance effectively.

7.7. More segmentation Visualisations

Additional visualizations of segmentation predictions in the 2-way 1-shot setting are presented in Figure 7. Our analysis reveals that predictions based solely on text or images often fail to capture the entire foreground or result in under-segmentation. In contrast, our text-augmented model produces significantly improved predictions, demonstrating its ability to leverage multi-modal information effectively.

	1-way-5-shot										2-way-5-shot									
	class. 0/1 exact ratio(%)					segmentation mIoU(%)					class. 0/1 exact ratio(%)					segmentation mIoU(%)				
Method	5^0	5^1	5^2	5^3	Avg	5^0	5^1	5^2	5^3	Avg	5^0	5^1	5^2	5^3	Avg	5^0	5^1	5^2	5^3	Avg
HSNet[18]	89.3	90.1	66.3	90.7	84.1	12.5	24.7	19.4	18.1	18.7	81.3	78.4	44.0	81.4	71.3	13.0	25.4	22.2	18.7	19.8
ASNet[11]	84.3	89.1	66.2	90.0	82.4	11.5	22.0	14.0	17.4	16.2	72.5	80.6	41.8	76.8	67.9	8.7	23.1	11.8	18.0	15.4
CST[12]	92.7	89.4	70.3	89.2	85.4	42.1	40.8	30.8	31.2	36.2	86.2	77.4	48.5	73.9	71.5	40.9	40.1	29.8	31.3	35.5
CST(CLIP-B/16)*	96.1	95.1	84.4	93.6	92.3	74.2	72.7	46.7	67.4	65.2	92.0	91.3	72.0	86.5	85.4	73.7	72.4	47.4	67.0	65.1
(Ours)Text Aug	**97.0**	**95.8**	**87.4**	**95.2**	**93.9**	**74.3**	**74.3**	**52.9**	**68.2**	**67.4**	**95.2**	**91.9**	**76.5**	**89.2**	**88.2**	**74.2**	**73.7**	**52.3**	**67.4**	**66.7**

Table 11. Comparison of N-way 5-shot performance between our method and prior approaches on Pascal-5^i. The "*" indicates our baseline, CST [12], with CLIP-B/16 as the backbone.

	class. 0/1 exact ratio(%)					segmentation mIoU(%)				
Method	5^0	5^1	5^2	5^3	Avg	5^0	5^1	5^2	5^3	Avg
Single Correlation Trans.	91.6	91.8	68.5	93.7	86.4	63.2	70.4	49.0	62.7	61.3
Dual Correlation Trans.	92.4	92.6	69.1	94.0	87.0	68.8	72.5	49.5	62.7	63.4

Table 12. Comparison of results between dual correlation transformer and single correlation transformer on Pascal-5^i.

Method	Arch	20^0	20^1	20^2	20^3	Mean
PFENet [28]	ResNet-50	34.3	33.0	32.3	30.1	32.4
HSNet [18]	ResNet-50	36.3	43.1	38.7	38.7	39.2
HSNet [18]	ResNet-101	41.5	44.1	42.8	40.6	42.2
ASNet [11]	ResNet-50	41.5	44.1	42.8	40.6	42.2
ASNet [11]	ResNet-101	41.8	45.4	43.2	41.9	43.1
MCL [32]	ResNet-50	46.8	35.3	26.2	27.1	33.9
MCL [32]	ResNet-101	50.2	37.8	27.1	30.4	36.4
FPTrans [39]	DeiT-B/16	44.4	48.9	50.6	44.0	47.0
CST [12]	ViT-S/8	39.6	45.8	45.0	45.5	44.0
CST*	CLIP-B/16	47.9	54.8	52.1	50.1	51.2
Text-Only	CLIP-B/16	49.1	54.6	50.0	48.3	50.5
Text-Aug	CLIP-B/16	**49.8**	**59.6**	**56.2**	**54.3**	**55.0**

Table 13. Fold-wise FS-S results for the 1-way 1-shot configuration on COCO-20^i. The "*" indicates our baseline CST [12] using CLIP-B/16 as the backbone.

	class. 0/1 exact ratio(%)					segmentation mIoU(%)				
Method	1	2	3	4	5	1	2	3	4	5
PANet [31]	69.0	50.9	39.3	29.1	22.2	36.2	37.2	37.1	36.6	35.3
PFENet [28]	74.6	41.0	24.9	14.5	7.9	43.0	35.3	30.8	27.6	24.9
HSNet [18]	82.7	67.3	52.5	45.2	36.8	49.7	43.5	39.8	38.1	36.2
CST [12]	85.7	70.4	57.3	47.3	36.9	55.5	53.7	52.6	52.0	50.3
CST(CLIP-B)*	89.0	79.0	71.7	64.6	57.5	62.4	61.4	60.7	60.5	59.4
Ours)Text-Only	90.2	82.2	75.4	70.0	65.3	59.3	58.6	58.3	57.9	57.3
Ours)Text-Aug	**91.2**	**83.3**	**77.8**	**73.0**	**66.2**	**66.2**	**65.4**	**65.3**	**65.0**	**64.4**

Table 14. Average FS-CS performance across four folds for N-way 1-shot on Pascal-5^i.

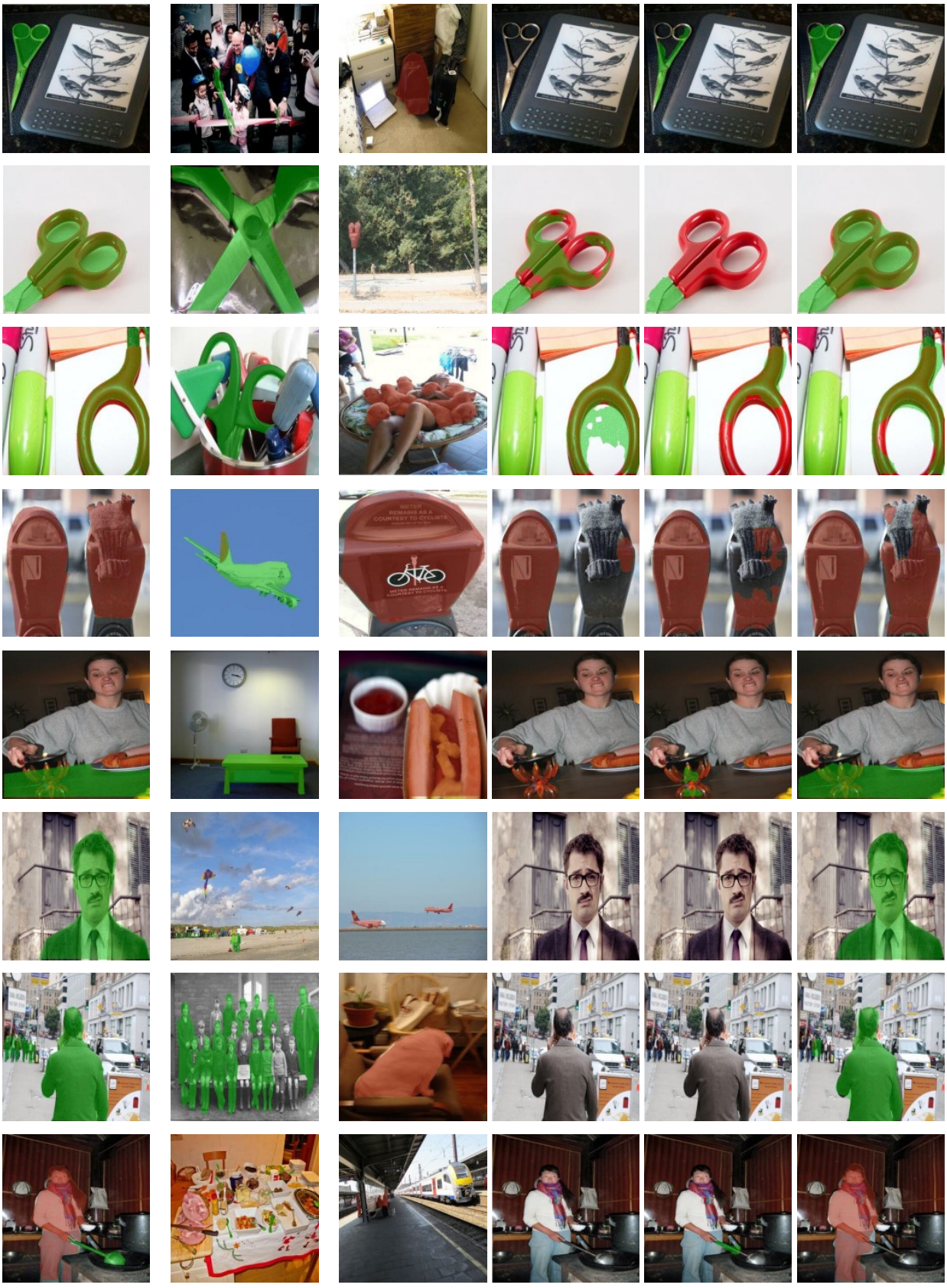

Figure 7. Segmentation predictions in the 2-way 1-shot setting with support sets comprising only images (image-only), only text (text-only), or both images and text (text-augmented). From left to right, the visualization includes the query image, support image 1, support image 2 with segmentation maps, image-only predictions, text-only predictions, and text-augmented predictions.