

REPRESENTATION MUTUAL LEARNING FOR END-TO-END WEAKLY-SUPERVISED SEMANTIC SEGMENTATION

Anonymous authors

Paper under double-blind review

1 APPENDIX

1.1 MORE EXPERIMENTAL RESULTS FOR CONSUMPTION TIME

To verify the efficiency of our RML framework, we conduct time-consuming comparison experiments with state-of-the-art methods in Tab. A1. It can be observed that our RML consumes the least time and has the highest segmentation accuracy. AdvCAM Lee et al. (2021) iteratively finds newly activated pixels and uses gradients to perturb the image, which is time-consuming. RIB Lee et al. (2021) requires 10 feedforward and feedback iterations for each test image, and its inference cost alone takes 8 hours, while the total cost of our RML is only 3 hours.

Table A1: Comparison of our RML with state-of-the-art methods in mIoU (%) and consumption time (*hours*) on PASCAL VOC 2012 val set. The time counted is the total time from training the model to generating segmentation masks.

Methods	PASCAL VOC	
	mIoU	Time
IRN Ahn et al. (2019)	63.5	8.1
AdvCAM Lee et al. (2021)	58.3	195.5
AdvCAM Lee et al. (2021) + IRN Ahn et al. (2019)	68.1	201.6
RIB Lee et al. (2021)	68.3	19.5
1Stage Araslanov & Roth (2020)	62.7	3.4
AFA Ru et al. (2022)	63.8	4.1
RML (Ours)	64.9	3.0

1.2 MORE EXPERIMENTAL RESULTS FOR HYPER-PARAMETERS

In this section, we conduct parameters studies on the PASCAL VOC 2012 dataset.

Impact of α in the CAM-driven Instance-leave Mutual Learning loss. As shown in Tab. A2, the model has the highest mIoU score when $\alpha = 0.01$. Therefore, our choice is favorable, that is, the weight of α should not be too large or too small.

Table A2: Influence of α in Eq. 4 on PASCAL VOC 2012 test set.

α	0.001	0.005	0.01	0.015	0.02
mIoU (%)	64.6	64.8	65.4	64.4	64.2

Impact of β_1 in the feature-leave mutual learning loss. The effect of β_1 defined in Eq. 8 on segmentation performance is shown in Tab. A3. If β_1 is too small, the contribution of minimizing mutual information to preserve specific information is modest, and the probability of pixels being

misclassified is high. And if β_1 is too large, it will lead to a decrease in segmentation accuracy. Therefore, our choice is favorable.

Table A3: Influence of β_1 in Eq. 8 on PASCAL VOC 2012 test set.

β_1	10	50	100	150	200
mIoU (%)	63.5	63.9	65.4	64.1	63.7

Impact of β_2 in the Affinity-aware Pixel-level Mutual Learning loss. We show the effect of β_2 on segmentation performance in Tab. A4. Similar to β_1 , β_2 cannot be too small or too large. If β_2 is too small, the contribution of pixel-level mutual information loss is weak, resulting in a decrease in segmentation accuracy. Therefore, we choose $\beta_2 = 100$ in the paper.

Table A4: Influence of β_2 in Eq. 14 on PASCAL VOC 2012 test set.

β_2	10	50	100	150	200
mIoU (%)	63.3	63.7	65.4	64.2	64.0

Impact of the weight factors in the total loss. In addition, we also conduct sensitivity studies on the weight factors (λ_1 , λ_2 , λ_3) in the total loss. As shown in Tab. A5, when $\lambda_1 = 0.01$, $\lambda_2 = 0.01$, $\lambda_3 = 0.01$, the mIoU of the model is the highest. From the experimental results, all three weight factors have a great impact on the performance of the model, proving the effectiveness of our proposed CIML, MFML and APML.

Table A5: Influence of the weight factors in the total loss term on PASCAL VOC val set.

	λ_1	λ_2	λ_3	<i>val</i>
	0.05			62.3
	0.2			60.1
	0.4			59.5
		0.05		62.7
		0.2		61.8
		0.4		60.2
			0.05	61.9
			0.2	61.8
			0.4	58.8
Default	0.1	0.1	0.1	64.9

1.3 MORE EXPERIMENTAL DETAILS

Our network architecture uses the MiT-B1 proposed in SegFormer Xie et al. (2021) as the backbone, and adopts the same subtle adjustments as AFA Ru et al. (2022) to the network structure for fair comparison. During training, we set the batch size to 8. The affine transformation in our method can be any spatial transformation. Without additional instructions, we employ a spatial transformation with a downsampling rate of 0.3.

1.4 MORE QUALITATIVE RESULTS

We provide more visualizations of semantic segmentation in Fig. A1 and Fig. A2. The results show that our method can produce fine and high-quality semantic segmentation results.

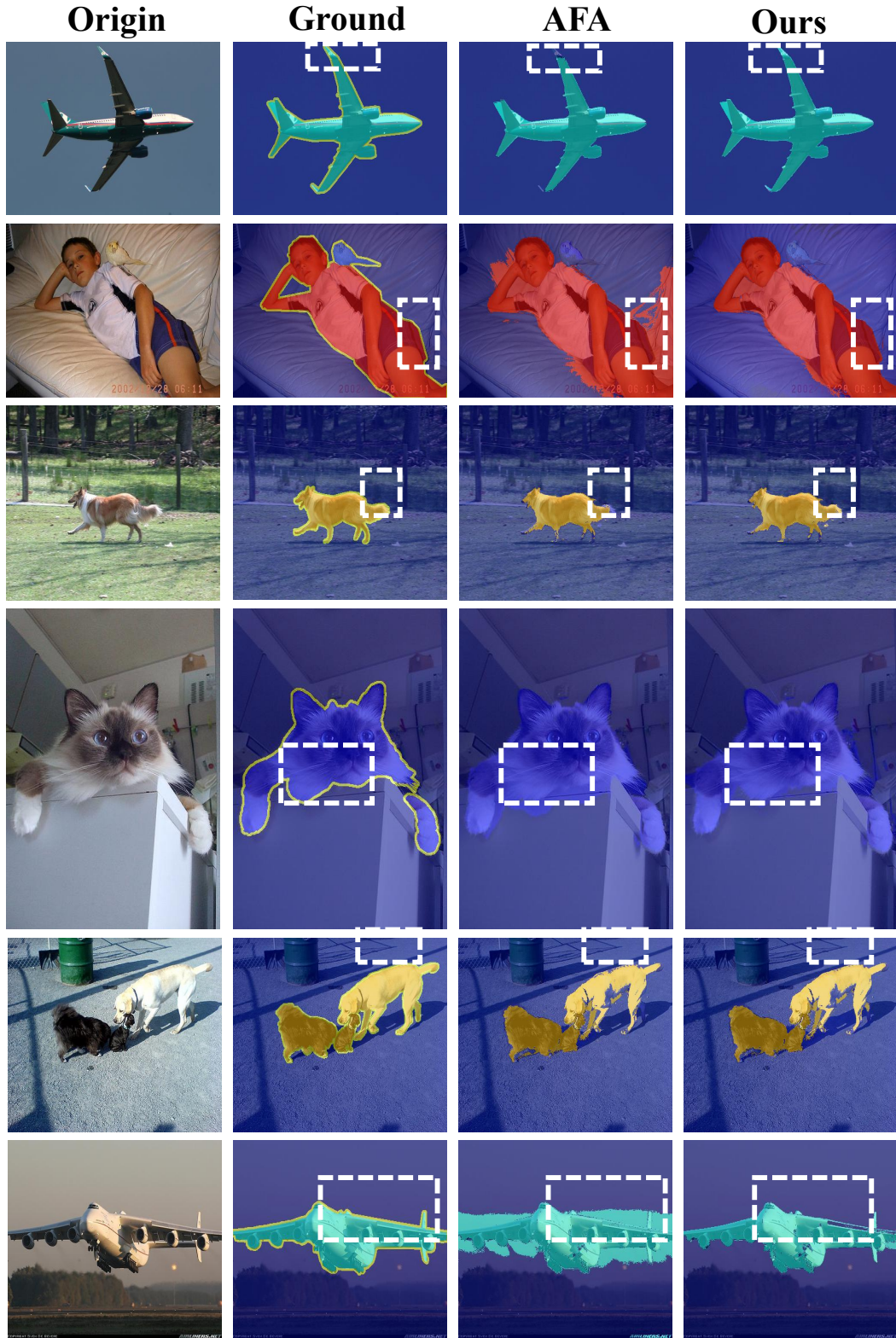


Figure A1: Semantic segmentation results of AFA Ru et al. (2022) and our RML on PASCAL VOC benchmark. From left to right: Original images; Ground truth; Segmentation results of AFA; Segmentation results of our RML. The white box shows the difference.

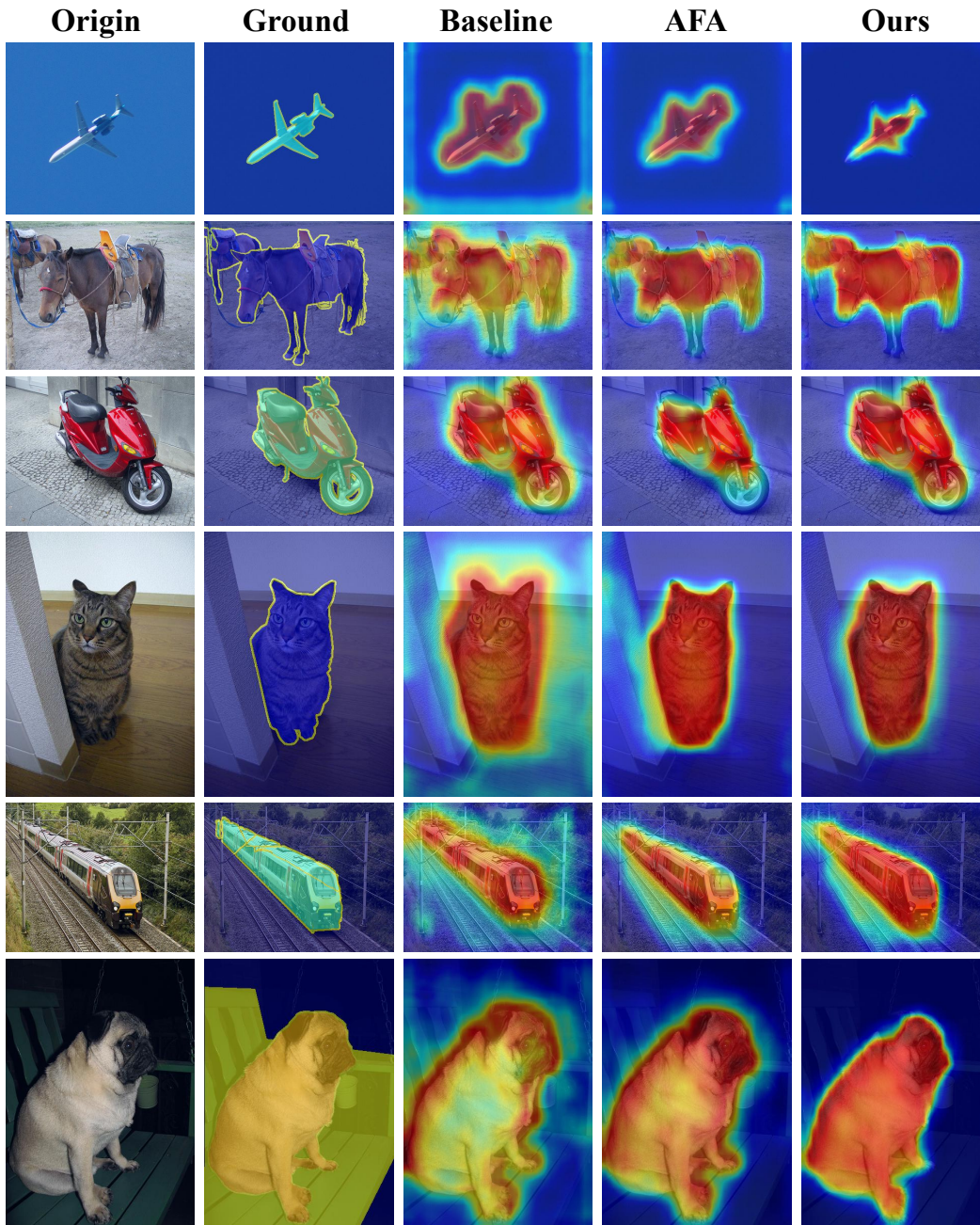


Figure A2: Qualitative results of CAMs. From left to right: Original images; Ground truth; CAMs generated by our baseline; CAMs generated by AFA Ru et al. (2022); CAMs generated by our RML. It can be observed that our method produces finer and higher quality activation maps than AFA.

REFERENCES

- Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2209–2218, 2019.
- Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4253–

4262, 2020.

Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4071–4080, 2021.

Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021.