SA-Matching DETR: A Lightweight Transformer Detector with Enhanced Scale Adaptive Matching

Supplementary Material

Table 1. Top-1 accuracy comparison of ImageNet-1K [2] pre-trained models on multiple downstream classification benchmarks. We highlight the **best** and the second-best results.

Model	CIFAR-10	CIFAR-100	Food	Pet
ViT-B-16 [3]	98.1	87.1	–	93.8
DeiT-B [8]	99.1	90.8	–	–
DC-ViT-B [10]	98.0	87.5	–	92.2
EfficientNet-B7 [7]	98.9	**91.7**	–	–
ResMLP-S24 [9]	98.7	89.5	–	–
ECTFormer-×1.0 [6]	–	–	89.5	92.5
ECTFormer-×1.25 [6]	–	–	89.5	93.2
Partial ViT-S	98.7	90.0	90.9	94.5
Partial ViT-B	**99.2**	**91.7**	**93.7**	**95.8**

1. Supplementary Experiments

In this section, we first assess the generalization of our pre-trained representations through fine-tuning on a variety of downstream classification benchmarks. We then analyze the key hyperparameters of the SA-Matching DETR architecture, including the Partial ViT design and SA-Matching settings. Unless otherwise noted, all experiments use the **Partial ViT-S** backbone with a 12-epoch training schedule.

1.1. Downstream Classification Transferability

To assess the generalization capabilities of our pre-trained representations, we evaluated our models through fine-tuning on a diverse set of downstream classification benchmarks: CIFAR-10, CIFAR-100 [4], Food-101 [1], and Oxford-IIIT Pet [5]. As summarized in Tab. 1, our models demonstrate excellent transfer learning performance. The Partial ViT-B model achieves 99.2% and 91.7% accuracy on CIFAR-10 and CIFAR-100 respectively, surpassing prominent architectures like ViT [3] and DeiT [8]. The efficiency of our approach is highlighted by the lightweight Partial ViT-S, which, with only 17.1M parameters, achieves a highly competitive 90.9% on Food-101 and 94.5% on the Pet dataset. These results validate the strong generalization ability of Partial ViT, confirming that our channel decomposition strategy produces robust and transferable features.

1.2. Channel Decomposition Ratio

We analyze the impact of the channel decomposition ratio, r, on the performance and efficiency of our Partial ViT-S model. As shown in Tab. 2, this strategy effectively reduces redundant computations while retaining critical features. The results indicate that lower channel retention ratios (30%, 40%) offer high parameter efficiency but compromise overall accuracy, whereas higher ratios (60%, 70%) yield only marginal accuracy gains over the 50% configuration at the expense of efficiency. Consequently, a 50% channel retention strikes the optimal balance between accuracy and computational cost, and is thus adopted as our default setting.

Table 2. Top-1 accuracy and parameter count of Partial ViT-S on ImageNet-1K under different channel decomposition ratios.

r (%)	Acc	Param (M)	Acc/Param
30	79.5	16.0	5.0
40	80.6	16.4	4.9
50	81.3	17.1	4.8
60	81.6	17.9	4.6
70	81.6	18.9	4.3

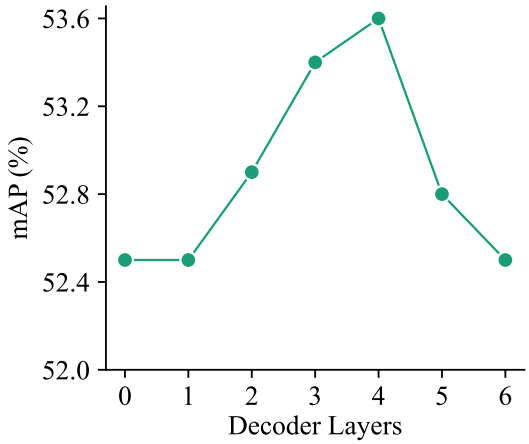

Figure 1. Impact of the number of decoder layers on detection performance.

Table 3. The impact of η_s and η_l on detection performance.

η_s	η_l	mAP	η_s	η_l	mAP
24	96	49.3	32	88	49.0
28	96	49.2	32	92	49.2
32	96	49.0	32	96	49.0
36	96	48.8	32	100	49.3
40	96	48.8	32	104	49.0

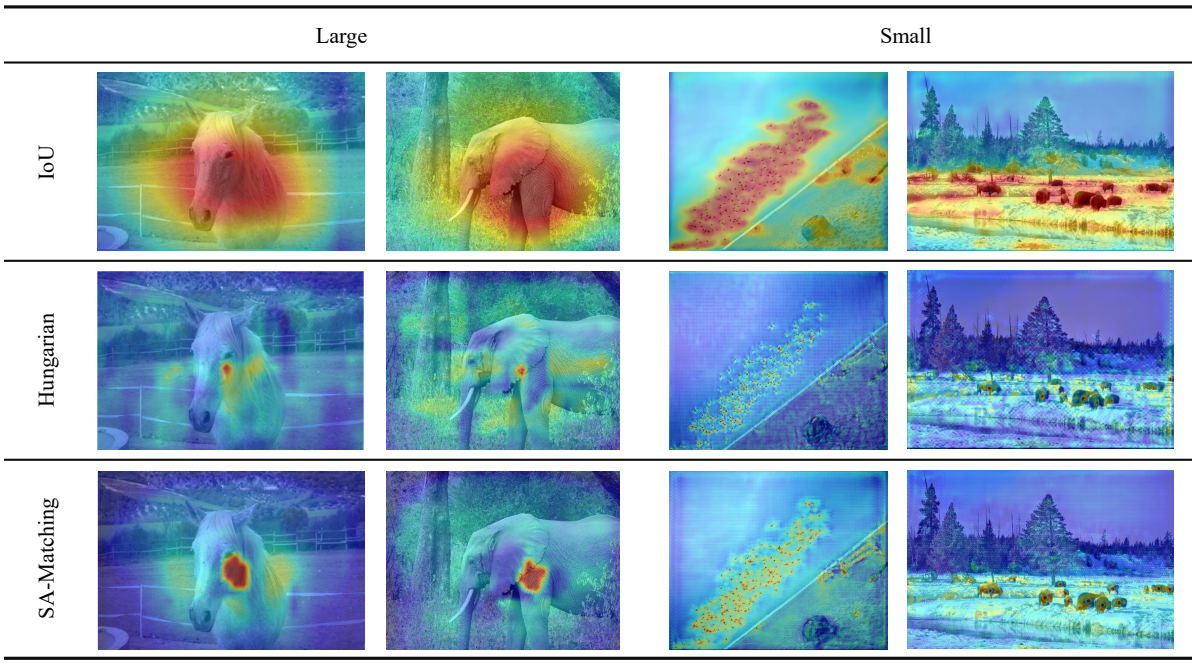

Figure 2. Class activation maps (CAMs) illustrating the object assignment of detectors under different matching algorithms, highlighting the optimization effect of SA-Matching on object assignment.

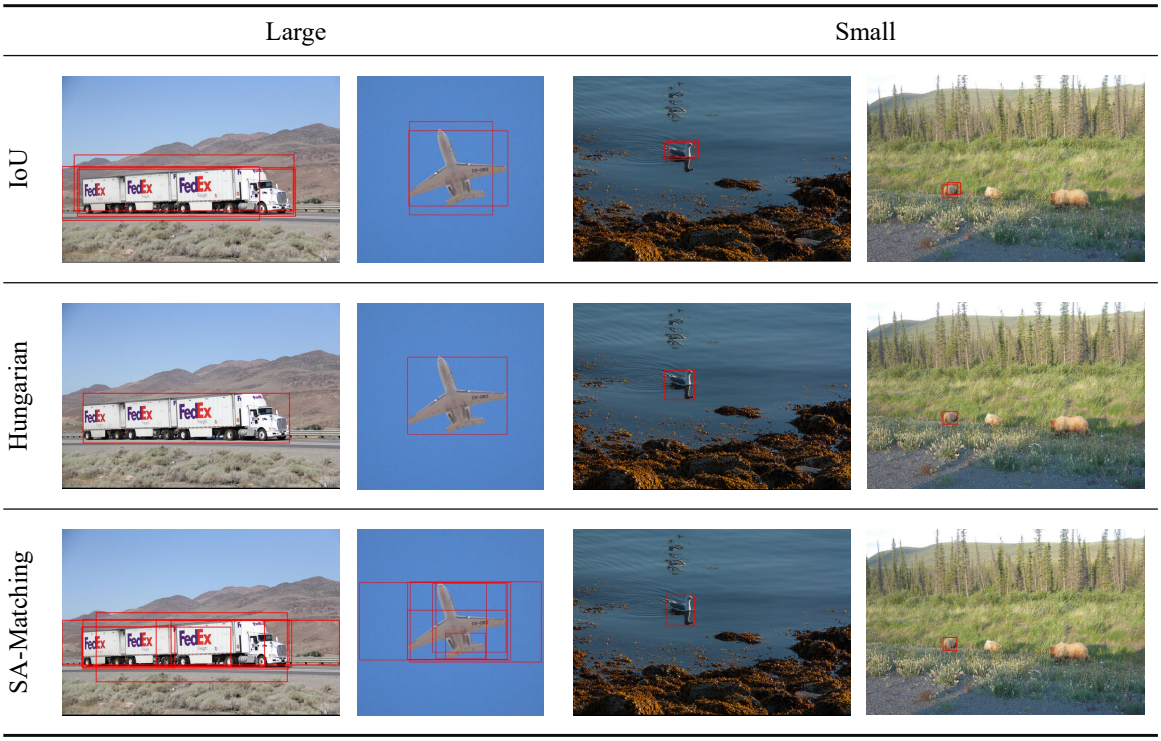

Figure 3. Visualization of predicted bounding boxes positively matched to ground-truth objects, illustrating the object assignment of detectors under different matching algorithms and highlighting the optimization effect of SA-Matching.

1.3. Decoder Depth

We analyze the impact of decoder depth using our SA-Matching DETR with a Partial ViT-B backbone under a 12-epoch training schedule. As shown in Fig. 1, the depth of the decoder critically influences multi-stage feature integration. Experimental results indicate that a 4-layer de-

coder attains 53.6% mAP, whereas a 3-layer configuration achieves nearly identical performance (53.4%) with lower model complexity. Balancing efficiency and accuracy, we adopt the 3-layer decoder as the default setup, ensuring high inference speed without compromising precision.

1.4. Scale Thresholds

The performance of SA-Matching algorithm is governed by the scale-dependent augmentation factors (s, m, l) and the object scale thresholds (η_s, η_l). For the scale thresholds, we adopt the standard definitions from the COCO dataset, which categorizes small objects as having an area less than 32^2 pixels and large objects as having an area greater than 96^2 pixels.

To validate this choice, we conducted an ablation study to assess the model's sensitivity to these thresholds, with the results presented in Tab. 3. Our experiments show that varying η_s (from 24 to 40) and η_l (from 88 to 104) around the COCO-defined values results in only minor performance fluctuations, with mAP remaining stable within a narrow 0.5-point range. Given this robustness and to maintain alignment with standard evaluation protocols, we set our final thresholds to $\eta_s = 32^2$ and $\eta_l = 96^2$.

1.5. Matching Strategy Analysis via Visualization

The SA-Matching algorithm achieves a remarkable mAP increase over the standard Hungarian algorithm and IoU-based matching. The underlying reasons for this performance gain are illustrated in our visualizations. The Class Activation Maps (CAMs) in Fig. 2 reveal the limitations of traditional methods: Hungarian matching produces overly constrained and under-activated regions for large objects, indicating sparse supervision, while IoU matching generates diffuse activations for small objects, a characteristic that often leads to FPs. In contrast, SA-Matching yields precise and comprehensive activation maps for objects of all sizes.

Furthermore, Fig. 3 visualizes the training-time assignment process. For medium and large objects, SA-Matching successfully assigns multiple predicted boxes as positive samples. While some of these assigned boxes may be less precise than the top prediction, their inclusion provides richer and more diverse gradient signals during training. This enhanced supervision allows the model to learn more robust object representations, which leads to higher detection accuracy at inference time. This approach effectively increases the density of positive samples while maintaining high precision, highlighting its advantages in enhancing overall performance.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[3] 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

[4] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. Technical Report. 1

[5] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 1

[6] Jaewon Sa, Junhwan Ryu, and Heegon Kim. Ectformer: An efficient conv-transformer model design for image recognition. *Pattern Recognition*, 159:111092, 2025. 1

[7] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 1

[8] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 1

[9] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):5314–5321, 2022. 1

[10] Hanxiao Zhang, Yifan Zhou, and Guo-Hua Wang. Dense vision transformer compression with few samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15825–15834, 2024. 1