Pay Attention to the Foreground in Object-Centric Learning

Supplementary Material

A. Implementation details

A.1. Baselines

Slot Attention[4] makes its core contribution by introducing the Slot Attention module. This module, based on an attention mechanism, employs an iterative process to align a set of shared slots with input features, generating object-level representations. This approach enables effective learning of object segmentation and representation without the need for supervision, addressing the problem of unsupervised object discovery and learning. It has become the mainstream paradigm in object-centric learning.

SLASH[3] aims to address the stability issue in object-centric learning for single-view images, particularly the "bleeding issue" where attention leaks into the background. To solve this problem, SLASH introduces two key modules: Attention Refining Kernel (ARK) and Intermediate Point Predictor and Encoder (IPPE). ARK, a learnable low-pass filter, optimizes attention maps by reducing noise and enhancing object-like patterns, while IPPE provides positional guidance to slots via weak semi-supervision, injecting location information into the slots. Together, these modules enable SLASH to achieve stable and robust scene decomposition.

SLATE[6] aims to address the limitations of existing object-centric representation learning models in compositional systematic generalization for image generation. By combining the strengths of DALL·E and object-centric representation learning, SLATE proposes a slot-based autoencoder architecture that uses a slot-conditioned Image GPT decoder to handle complex interactions among image components, overcoming the slot-decoding dilemma and the pixel independence issue found in traditional decoders. SLATE learns composable representations directly from images, enabling stronger systematic generalization capabilities.

SlotDiffusion[7] proposes an object-centric learning method based on a Latent Diffusion Model, aiming to address the insufficient generative capabilities of existing object-centric models in image and video generation, while maintaining their performance in object segmentation, significantly improving image generation and temporal reasoning.

LSD (Latent Slot Diffusion)[1] replaces traditional slot decoders with a conditional latent diffusion model, enabling unsupervised compositional generation based on visual concepts extracted from images. This model addresses the challenge of applying diffusion models to object-centric learning and demonstrates superior performance compared to transformer-based autoregressive models, particularly in tasks such as object segmentation, property prediction, and image editing for complex natural scenes.

DINOSAUR[5] introduces an architecture that combines DINO's self-supervised feature reconstruction loss with the Slot Attention module, incorporating an inductive bias based on the homogeneity of features within objects. This approach addresses the limitations of existing image-based object-centric learning methods in handling complex real-world scene data, successfully bridging the gap between object-centric representation learning on synthetic and real-world datasets. It has gradually become a new paradigm in the field of object-centric learning.

SPOT[2] introduces a dual-stage strategy to enhance unsupervised object-centric learning in slot-based autoencoders, addressing challenges in handling complex real-world images. It improves slot generation through a self-training scheme that distills superior slot-attention masks from the decoder to the encoder, enhancing object segmentation precision. Additionally, it strengthens autoregressive decoders by incorporating sequence permutations, which amplify the role of slot vectors in reconstruction and provide more robust supervisory signals.

A.2. More implementation details

Foreground and background indicator. In the indicator, we use ViT-B/16 as the encoder (initialized with DINO by default) and optimize only the final layer. We employ the SGD optimizer with a learning rate set to 0.001 and a batch size of 512. The learning rate follows a linear warm-up for $10,000$ steps and then an exponential decay schedule. Additionally, we clip the gradient norm to 1 to stabilize the training process. We train for 100 epochs across all datasets.

For data augmentation in the indicator, we apply color jitter with a probability of 80%, convert the image to grayscale with a probability of 20%, flip the image horizontally with a probability of 50%, and invert the image pixel values with a probability of 20%. In the two branches of contrastive learning, Gaussian blur is applied with probabilities of 100% and 10%, respectively.

For all datasets, we set the loss weights of $\mathcal{L}^{\text{pixel}}$, $\mathcal{L}^{\text{stuff}}$, and $\mathcal{L}^{\text{sep}}$ to 0.5, 0.5, and 0.5, respectively.

Fusion stage. In the fusion stage, we use ViT-B/16 as the encoder (initialized with DINO by default) and freeze it during training. We employ the Adam optimizer with a learning rate of 4e-4 and a batch size of 64. Following [5], we use 7, 6, and 11 slots for the COCO, PASCAL, and MOVi-C datasets, respectively.

Region Combination. Region Combination is an optional

module in this paper, used to refine the results from the fusion stage using spectral clustering. We primarily use an eigenvalue gap heuristic to determine the optimal number of clusters, N, and set a lower bound for N. In MOVi, PASCAL, and COCO, we set N to 10, 5, and 6, respectively.

B. Additional experimental results

B.1. Relationship between the indicator and the final results on COCO dataset

As shown in Table 1, the conclusions on COCO are consistent with those discussed for PASCAL in the main text,*i.e.*, the performance of the indicator is positively correlated with the final model results. However, there is one difference: our indicator shows a smaller improvement in instance-level scene segmentation. We believe this is because our foreground-background indicator only performs segmentation at the semantic level, and in the more complex scene of the COCO dataset, it is insufficient to provide instance-level information for slot attention-based methods. Achieving both semantic-level and instance-level scene decomposition is one of the challenges for current slot attention-based methods.

SA Method	Indicator	IoU	mBO^c	mBO^i
DINOSAUR	-	-	38.9	31.1
	Random	-	24.5	19.0
	ViT	31.3	39.4	31.1
	Ours	**48.4**	**40.1**	**31.2**

Table 1. *Relationship between the indicator and the final results.* This experiment is conducted on COCO dataset.

B.2. Impact of different losses on COCO dataset

As shown in Table 2, the results on COCO are consistent with those on PASCAL, *i.e.*, using only $\mathcal{L}^{\text{pixel}}$ and $\mathcal{L}^{\text{stuff}}$ leads to training collapse. It is necessary for $\mathcal{L}^{\text{sep}}$, $\mathcal{L}^{\text{pixel}}$, and $\mathcal{L}^{\text{stuff}}$ to work together in synergy to achieve effective foreground-background separation.

B.3. Performance across different foreground sizes

Figures 1 and 2 show the results of our SPOT-based method on the PASCAL and COCO datasets for different foreground sizes. We calculate the proportion of foreground pixels to total pixels and categorize the results accordingly. The results show that our method outperforms the baseline methods across different foreground sizes. On the PASCAL dataset, when the foreground size exceeds 40%, our method exhibits strong robustness, while its performance declines when the foreground size is below 40%. We believe this is due to the low resolution of the ViT encoder. On the COCO dataset, our method improves as the foreground size increases, indicating that our approach is more suitable for complex scenes with more foregrounds. However, it is still limited by the low resolution of the ViT encoder in scenes with smaller foreground sizes.

$\mathcal{L}^{\text{pixel}}$	$\mathcal{L}^{\text{stuff}}$	$\mathcal{L}^{\text{sep}}$	IoU	mBO^c	mBO^i
✓			33.2	35.4	28.3
✓	✓		32.9	34.9	26.9
✓		✓	46.3	39.3	30.8
✓	✓	✓	**48.4**	**39.9**	**31.5**

Table 2. *Analysis of different loss functions on COCO dataset.* IoU represents the results of the indicator on segmenting foreground and background regions. mBO^c and mBO^i represent the final results of DINOSAUR after fusing knowledge from different indicators, which are optimized using different types of loss functions.

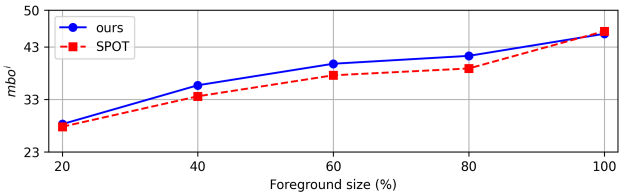

Figure 1. *Results across different foreground sizes on COCO dataset*

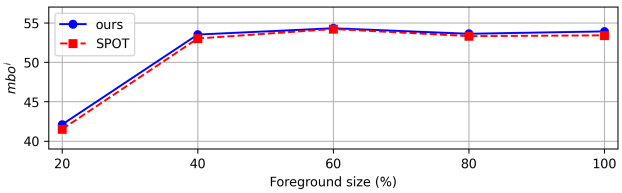

Figure 2. *Results across different foreground sizes on PASCAL dataset*

C. More visualization

To provide a more comprehensive understanding of our method, we present additional visualizations. Except for the indicator demonstration, all experiments are conducted on two real-world datasets (PASCAL and COCO) and two baseline methods (DINOSAUR and SPOT). Figure 3 shows the superiority of our proposed Foreground and Background indicator compared to the slot attention-based method using 2 slots. Figures 4 and 5 demonstrate the role of different components of our method. Figure 6 presents more qualitative results. From these results, it is evident that our method, compared to the baseline methods, can correctly fuse fore-

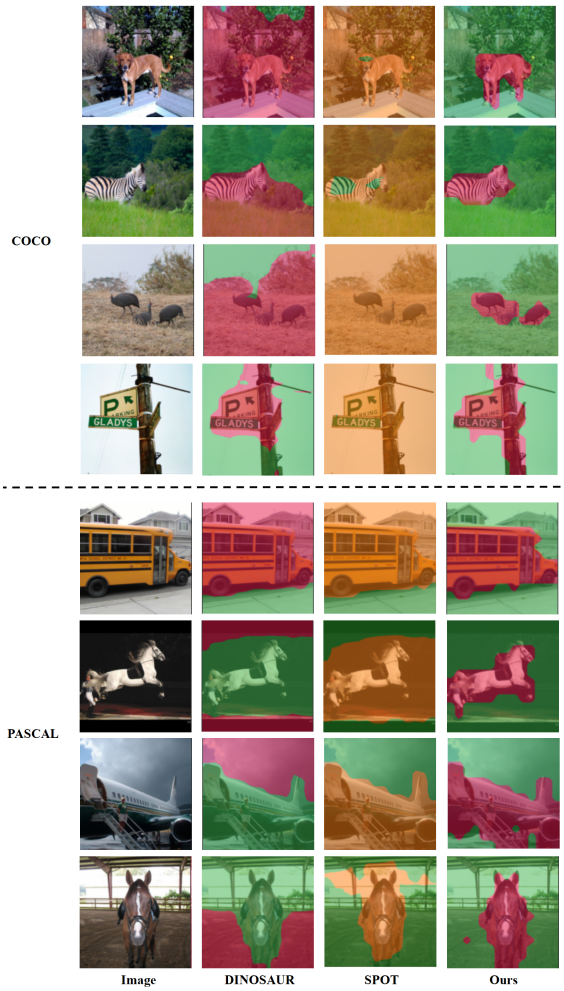

Figure 3. More example results of the proposed indicator and SOTA slot attention method with 2 slots on COCO and PASCAL dataset.

ground and background, leading to better scene decomposition.

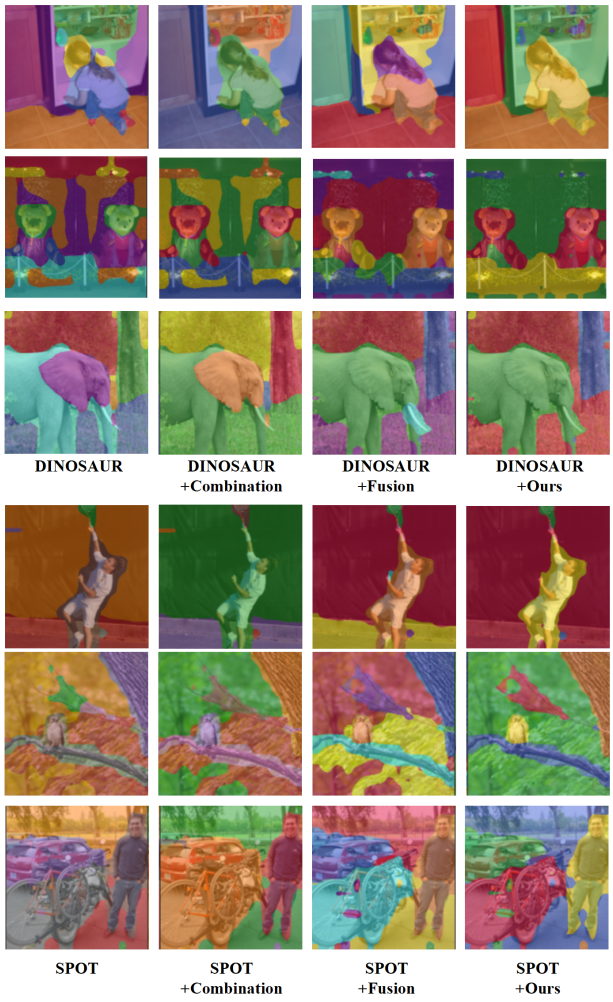

Figure 4. *The visualization results in our different components on COCO dataset*

References

[1] Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 8563–8601, 2023. 1

[2] Ioannis Kakogeorgiou, Spyros Gidaris, Konstantinos Karantzalos, and Nikos Komodakis. Spot: Self-training with patch-order permutation for object-centric learning with autoregressive transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22776–22786, 2024. 1

[3] Jinwoo Kim, Janghyuk Choi, Ho-Jin Choi, and Seon Joo Kim. Shepherding slots to objects: Towards stable and robust object-centric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19198–19207, 2023. 1

[4] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit,

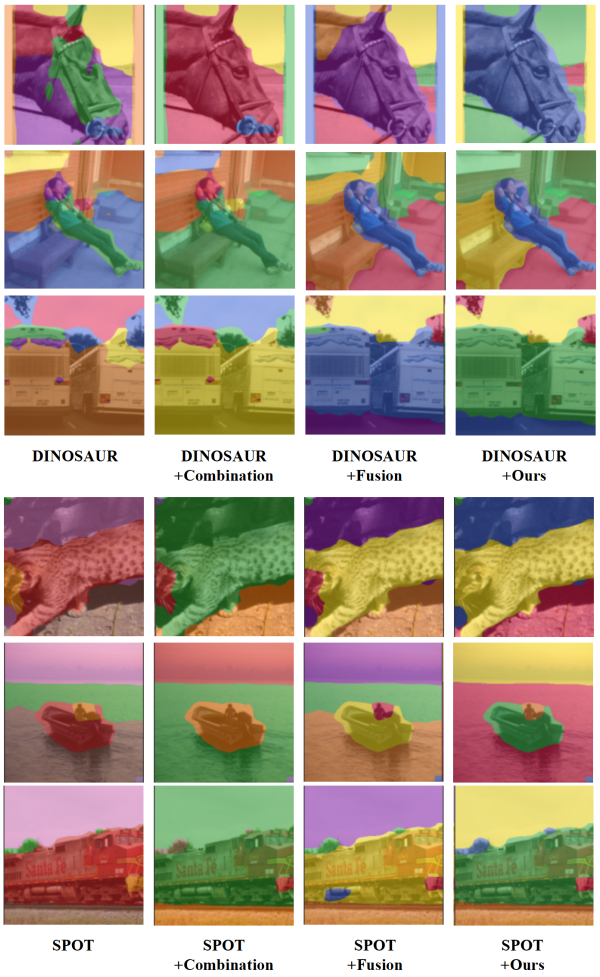

Figure 5. *The visualization results in our different components on PASCAL dataset*

Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020. 1

[5] Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. In *The 11th International Conference on Learning Representations*, 2023. 1

[6] Gautam Singh, Fei Deng, and Sungjin Ahn. Illiterate dall-e learns to compose. In *10th International Conference on Learning Representations, ICLR 2022*, 2022. 1

[7] Ziyi Wu, Jingyu Hu, Wuyue Lu, Igor Gilitschenski, and Animesh Garg. Slotdiffusion: Object-centric generative modeling with diffusion models. *Advances in Neural Information Processing Systems*, 36:50932–50958, 2023. 1

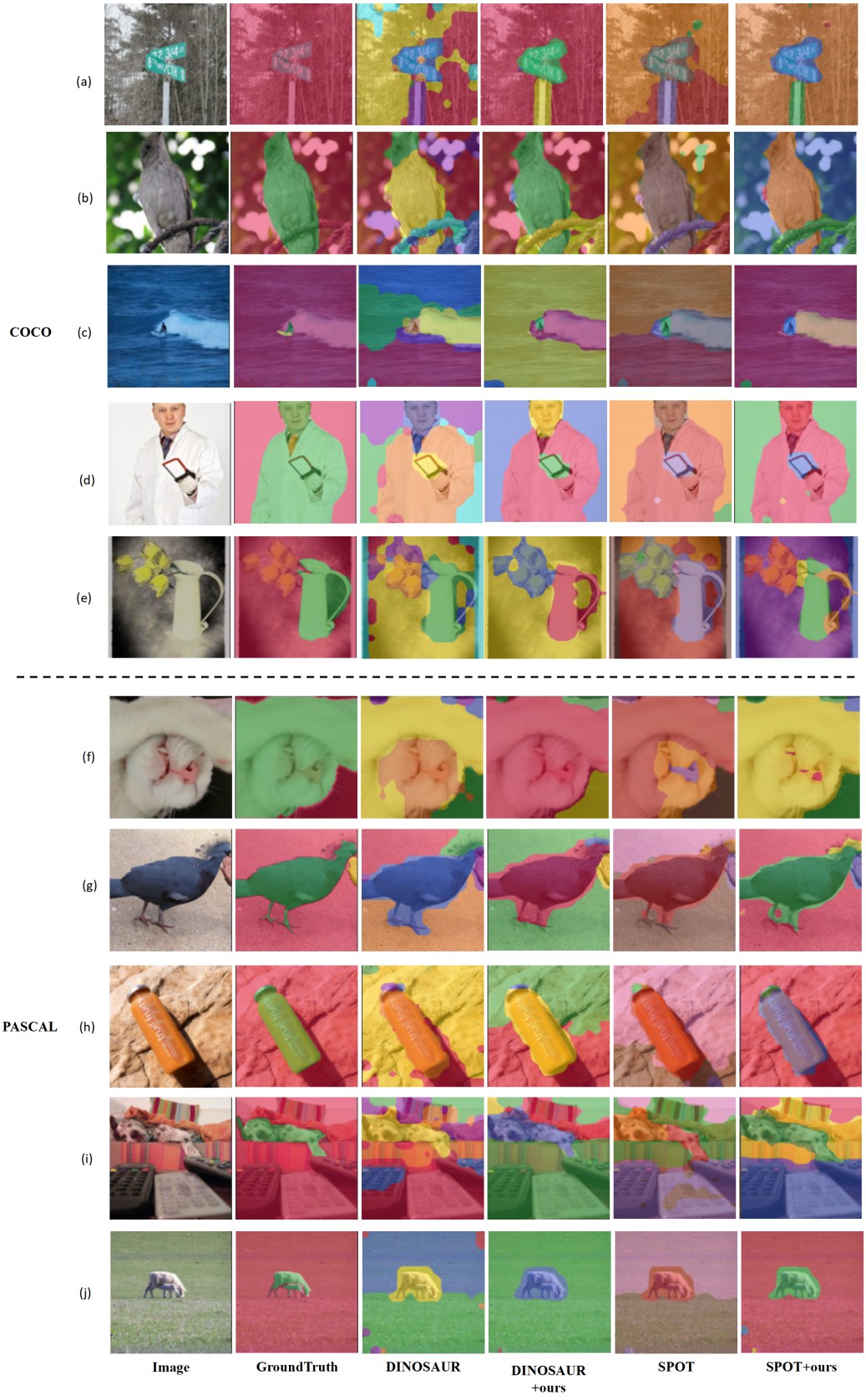

Figure 6. More qualitative results on COCO and PASCAL datasets. Ours: fusion stage + region combination.