AutoConcept: Unsupervised Extraction of Constituent Concepts from Single Image

Supplementary Material

Method	CLIP			DINO		
	SIM^I	$MSIM^I$	ACC^3	SIM^I	$MSIM^I$	ACC^3
(1) Using CLIP	0.60	0.69	0.24	0.15	0.25	0.25
(2) Using DINO	**0.71**	**0.78**	**0.79**	**0.39**	**0.50**	**0.88**

Table 1. Different feature extractor strategies on dataset D_2.

1. Additional Implementation Details

Obtaining Initial Clusters While obtaining initial clusters, as shown in Sec. 3.2 of paper, we start with the initial value of τ_{bz} equal to 0.75 for each given input image. We go on decreasing 0.025 from this value until we get at least one cluster for each master patch. This would ensure that clusters similar to each master patch are present in our candidate clusters so that clusters from all parts of the image are covered.

Computational Efficiency and Performance Analysis To analyze the computational efficiency and performance of our implementation, we evaluated both inference and training metrics. For inference, generating an image from a given prompt using the learned embeddings takes 47 seconds, with a peak GPU memory usage of 13.06 GiB. During training, the process takes 1046 seconds per image, with a peak GPU memory usage of 16.80 GiB.

2. Aditional Baseline Implementation Details

As shown in Table 6 of paper, we introduce two additional baselines: (1) We use CutLER [3] for mask extraction, while keeping all other components (Concept-wise Masked Denoising phase) identical to ConceptExpress [2]. CutLER extracts masks using an unsupervised edge-based approach that leverages contour cues and self-training. It refines object masks by progressively learning from high-confidence regions without requiring human annotations. (2) In ConceptExpress, we employ DINO [1] instead of CLIP as a text encoder to compute masked diffusion loss. We observe suboptimal performance in both baselines compared to ConceptExpress, while AutoConcept outperforms both.

3. Ablation Study on Feature Extractor

We used DINO and CLIP as the feature extractors in AutoConcept to obtain affinity matrix in the initial clustering phase. As Tab. 1 shows that our methodology aligns well with DINO, we use DINO as the feature extractor.

4. Additional Qualitative Results

Fig. 1 highlights the superior performance of our proposed method, AutoConcept, in generating more precise and accurate masks compared to ConceptExpress [2]. For instance, in the second example of Fig. 1, ConceptExpress fails to effectively separate the background, resulting in noisy masks. In the third example, it produces a single mask encompassing both birds, indicating a failure to differentiate between distinct instances of the same object. In the fourth example, ConceptExpress inaccurately generates two masks for the same object. In contrast, AutoConcept consistently produces tighter, more distinct, and semantically accurate masks, demonstrating its robustness across diverse scenarios.

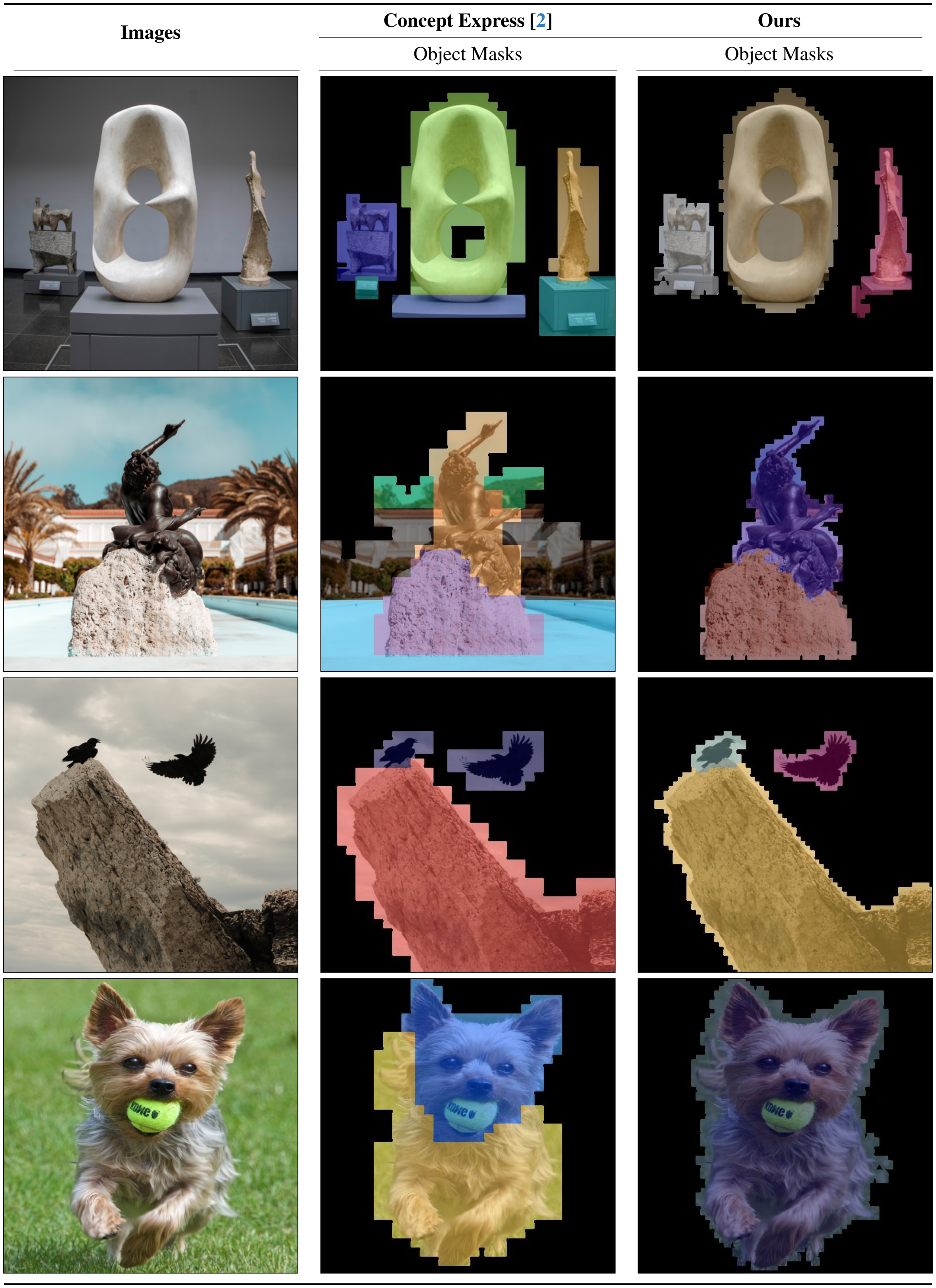

Figure 1. A qualitative comparison of the masks generated by Concept Express and our method. We show the generated mask images by ConceptExpress and AutoConcept for 4 examples. The object masks are represented by highlighting the concepts represented by the masks with different colors.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[2] Shaozhe Hao, Kai Han, Zhengyao Lv, Shihao Zhao, and Kwan-Yee K Wong. Conceptexpress: Harnessing diffusion models for single-image unsupervised concept extraction. *arXiv preprint arXiv:2407.07077*, 2024. 1, 2

[3] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3124–3134, 2023. 1