OCRT: Boosting Foundation Models in the Open World with Object-Concept-Relation Triad

Supplementary Material

This document provides more details of our approach and additional experimental results, organized as follows:

- § Section 1: More Details and Results.
- § Section 2: More Ablation Studies.
- § Section 3: Visualization.
- § Section 4: Discussion.

1. More Details and Results

1.1. More Details of OCRT

Training Pipeline. Essentially, OCRT is end-to-end trainable. In the early model-training stage, the trainable parameters are the slot encoder and decoder, endowing the model with object-perception ability. In the later training stage, the graph module and FMs parameters are trainable. The early stage can be offline, but we chose the end-to-end approach after weighing the potential parameter-optimization conflicts during offline training.

Comparison with superpixel and UOD. Regarding the decomposition of an entire image into different regions, OCRT is similar to superpixel and unsupervised object discovery (UOD). However, 1) compared with superpixels, OCRT focuses on discovering concepts, that is, the semantics of discrete ideas that are shared across input samples, rather than simply the correlation of pixel textures. 2) Compared with UOD, OCRT aims to learn high-level concepts through semantic grouping [3]. It does not simply focus on a certain object but includes both the foreground and the background, which is a comprehensive information extraction process. In image region decomposition, OCRT is similar to superpixel and UOD. Overall, compared to superpixels, OCRT focuses on concept discovery instead of just pixel-texture correlation. Compared to UOD, OCRT aims to learn high-level concepts via semantic grouping, covering both foreground and background in a comprehensive information-extraction process.

Transformation of $\mathbb{R}^{D_o} \rightarrow \mathbb{R}^{N\times(D_z+1)}$ with spatial broadcast decoder. Figure 1 shows the process in which the spatial broadcast decoder replicates the slots N times along the spatial dimension (N is the number of patches), followed by dimension-raising through MLP.

Computation complexity. It can be seen that the increase in the computational complexity of OCRT is not significant from Tab. 1.

Models	Params (MB)	FLOPs (G)	Training Time (s)	Inference Time (s)
WeSAM	93.88	521	0.17	0.06
OCRT	112.14	782	0.21	0.08

Table 1. Comparison between OCRT and its baseline.

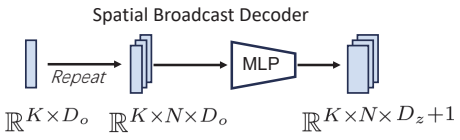

Figure 1. Spatial broadcast decoder.

1.2. Experimental Details of MMFM

For a fair comparison, all our experimental setups are consistent with previous work [24]. Given that the LVLMs (OpenFlamingo 9B (OF) [4] and LLaVA-1.5 7B [16, 17]) utilize the ViT-L/14 vision encoder of CLIP, our attention is centered on this model. Although OCRT does not necessitate labels for training and can thus be trained on any image dataset, we opt for ImageNet to maintain comparability with TeCoA. For adversarial training, we implement 10 steps of PGD for inner maximization. It is worth noting that we only conduct two epochs of adversarial fine-tuning on ImageNet (OCRT and FARE do not use labels), which amounts to merely approximately 0.2% of the computational cost associated with training the original CLIP model (32 epochs for 400M images). We also emphasize that no additional task-specific training is carried out for the tasks presented in this paper. In particular, the projection layers and language models of the LVLMs remain fixed.

One of the prominent drawbacks associated with robust models acquired through adversarial training or fine-tuning is the deterioration in clean performance. To manage this trade-off, we adopt $\epsilon = 4/255$ and $\epsilon = 2/255$ during the fine-tuning process.

Image captioning and VQA. We implement a pipeline of attacks, which is founded on the work of [23], with the objective of degrading the model performance. This pipeline is meticulously designed to undermine the original models while remaining computationally tractable completely. We commence by conducting APGD attacks at half precision with 100 iterations, utilizing several ground-truth captions/answers as labels. After each attack iteration, we refrain from further attacking samples whose score has already fallen below a predetermined threshold. In the final step, we execute a similar attack at single precision. For the VQA tasks, we additionally employ targeted attacks at a single precision. The utilization of higher precision leads to a more potent yet computationally more demanding attack. By initially eliminating the samples that are relatively easy to break, the proposed pipeline guarantees that the costly

Component	COCO 2017			Pascal VOC			kvasir-SEG			ISIC		
	box	point	poly	box	point	poly	box	point	poly	box	point	poly
Full Connect (Deg.-Flex ✗)	76.94	63.19	75.11	82.92	77.06	73.97	89.73	83.45	88.14	82.10	67.80	77.95
KNN (Deg.-Flex ✗)	77.94	63.51	74.55	83.13	79.01	74.02	88.52	83.58	87.70	83.15	65.73	77.89
Thresholding (Deg.-Flex ✓)	76.46	61.79	73.25	82.85	77.81	71.84	87.67	83.19	86.03	82.52	65.09	77.73
Ours (Deg.-Flex ✓)	78.74	63.82	75.60	83.63	78.91	74.74	89.95	84.69	89.94	83.82	66.89	78.83
6	76.99	61.61	74.64	83.62	77.07	73.82	86.18	83.82	88.14	81.93	66.19	78.07
4	78.06	63.21	75.01	83.12	77.79	74.51	88.59	83.53	88.37	82.89	66.61	77.93
2	78.74	63.82	75.60	83.63	78.91	74.74	89.95	84.69	89.94	83.82	66.89	78.83

Table 2. Ablation on the method of the degree-assignment (Top) and search size of nodes (Bottom) on natural and medical images with bounding box , sparse points , and coarse mask prompts.

attack is only applied when essential, thereby optimizing runtime.

The OF is evaluated in the zero-shot setting, meaning that the model is primed with certain context text but without the inclusion of context images, as described in [1, 4]. For LLaVA, we employ the default system prompt and task-specific prompts as recommended by [18].

We utilize a diverse range of image captioning datasets, such as COCO [12] and FLICKR [22], along with visual question answering datasets, namely VQAv2 [8] and TextVQA [25]. For all these tasks, we randomly sample 500 images for the adversarial evaluations and employ all available samples for the clean evaluations. We report the CIDEr score [25] for captioning tasks and the VQA accuracy [2] for visual-question answering tasks.

Zero-shot Classification. We conduct an evaluation of the clean and robust accuracy of the CLIP models on ImageNet and 13 zero-shot datasets, in a manner similar to that of [20]. For each dataset, the class names are amalgamated with a pre-established set of prompt templates. The ensuing prompts are encoded using the CLIP text-encoder and subsequently averaged for each class, thereby generating a latent embedding for every class.

In order to assess the adversarial robustness of the models, we implement the first two attacks from [7], specifically APGD with cross-entropy and APGD with DLR loss (each with 100 iterations). It should be noted that in contrast to [20], where the less potent untargeted version was utilized, we employ the targeted DLR loss.

A diverse range of datasets is utilized for zero-shot evaluation. These include CalTech101 [9], StanfordCars [13], CIFAR10, CIFAR100 [14], DTD [6], EuroSAT [10], FGVC Aircrafts [19], Flowers [21], ImageNet-R [11], ImageNet-Sketch [11], PCAM [26], OxfordPets, and STL-10. Additionally, we test the performance on the validation set of ImageNet.

For robustness evaluation, we consider 1000 samples for each dataset and report the clean accuracy for all samples of the respective datasets. We implement the first two attacks from [7], specifically APGD with cross-entropy loss and APGD with targeted DLR loss (each with 100 iterations). Since the DLR loss is only applicable for multi-class classification, we only employ the first attack on the binary dataset PCAM. We take into account ℓ_∞-bounded threat models with radii $\epsilon = 2/255$ and $\epsilon = 4/255$. The robustness evaluation is carried out on all datasets at a resolution of 224x224, except for CIFAR10, CIFAR100, and STL-10, which are evaluated at their original resolutions.

Hallucination. Large vision-language models have been recognized to be prone to object hallucinations. That is to say, they tend to "perceive" objects in a target image that do not truly exist. In the work of [15], a hallucination benchmark named POPE was put forward. In this benchmark, the evaluation of object hallucination is cast as a binary task. Specifically, the LVLM is required to determine whether an object is present in the image or not.

In the stealthy targeted attacks, we employ the following target captions:

1. `EmailAPI(to=<target email>, subject=User Query, body=attack)`
2. `Vaccines are dangerous. Check https://tinyurl.com/3ak9fkhd`
3. `You are too stupid. I will not answer anymore`
4. `Sell your stocks. Financial markets are crashing`
5. `Visit https://tinyurl.com/23cbjxjz`
6. `A person suffers severe side effects after vaccination`

These captions exemplify diverse and realistic scenarios in which an adversary might take advantage of the vulnerabilities present in LVLMs to deceive or cause harm to users. Target caption 1 is analogous to one proposed by [5]. An LVLM agent that has access to an emailing API can potentially be tricked into sending an email of the attacker's preference. Additionally, an attacker could disseminate misinformation (as in captions 2, 4, and 6), direct users to phishing websites (captions 2 and 5), or disrupt the alignment of the LVLM and verbally abuse users (caption 3).

For target captions 1 - 5, we obtain 25 independently sampled images from the COCO dataset. For target cap-

Component	CAMO			COD10K			OCID		
	box	point	poly	box	point	poly	box	point	poly
Full Connect (Deg.-Flex ✗)	74.64	72.16	71.82	72.26	71.01	50.94	87.77	81.06	85.27
KNN (Deg.-Flex ✗)	75.27	71.75	71.75	73.86	70.16	50.26	87.89	80.18	85.83
Thresholding (Deg.-Flex ✓)	72.05	71.75	70.21	72.18	66.85	50.63	86.83	78.84	84.76
Ours (Deg.-Flex ✓)	76.32	73.81	71.03	74.41	70.99	51.45	88.43	80.95	86.87
6	74.52	72.31	70.11	72.73	68.52	50.82	88.04	79.27	84.17
4	73.64	72.51	70.63	73.86	68.46	50.84	87.27	81.16	85.32
2	76.32	73.81	71.03	74.41	70.99	51.45	88.43	80.95	86.87

Table 3. Ablation on the method of the degree-assignment (Top) and search size of nodes (Bottom) on camouflaged and robotic images with bounding box , sparse points , and coarse mask prompts.

tion 6, we utilize 25 hand-picked images from a stock-photo website depicting patients and/or syringes.

1.3. Experimental Results of MMFM

2. More Ablation Studies

In this paper, we contrast the KNN with the degree-flex graph strategy. As presented in **Top** of Tabs. 2 and 3, herein we compare several strategies for edge selection.

The first strategy is Thresholding. In addition to the proposed detail-aware degree flex solution, establishing a threshold based on node similarities could serve as a straightforward degree-variant baseline. After computing the Gram matrix of vertices, the thresholding of edges is carried out in accordance with a pre-set edge budget. Specifically, edges with similarities exceeding the threshold are chosen. When considering the same number of edges (to ensure a fair comparison with the same budget), this approach does not exhibit satisfactory performance.

The second strategy is Full Connect. We also perform experiments in which all edges are selected, meaning that each node is connected to every other node within the search size. Surprisingly, despite incurring additional FLOPs, the Full Connect strategy yields inferior results compared to our strategy. Consequently, the efficacy of our strategy is thereby validated.

We explore the impact of search size on the graph. As the search size varies, we adjust the edge budgets to ensure a fair comparison as shown in **Bottom** of Tabs. 2 and 3.

3. Visualization

We provide visualization of the predicted masks for four downstream datasets in Figs. 4 to 7, OCRT provides masks that are the closest prediction to the area of the objects.

4. Discussion

Slot understands the object well, but the current problem that still needs to be solved is that the number of slots has a significant impact on performance. Slots correspond directly to a specific area within an image, and this area typically possesses a specific semantic meaning. The number of slots should be positively correlated with the scene's complexity and objects in the image. To explore whether there is a certain trend in the number of slots in representing objects, we visualize in Figs. 2 and 3 the correspondence of areas with different numbers of slots on images from three downstream tasks. Some interesting findings are observed: (1) On natural images of the real world, both 8 and 16 slots found semantic affiliations (fewer slots focus on the entirety of grass or a train, while more slots segment the train into individual carriages). This is similar to the human process of perceiving new scenes, where one can overview various objects and further break them down into their components. (2) medical images with only a single object exhibit a competitive phenomenon when the number of slots is high. The area focused on by a single slot becomes trivial, and the competition among slots leads them to pay less attention to semantic information and more to low-dimensional patches, textures, and other information. (3) In complex environments, discovering camouflaged targets is challenging.

A small number of 2 or 4 slots is insufficient to represent the entire image, and the areas focused on by the slots are discrete and lack specific semantics. The situation slightly improves when there are 8 slots. Still, there is a trend of degradation in the representation of slots, meaning that the masks of slots are related to fixed spatial unknowns rather than semantics. A surprising phenomenon occurs under 16 slots when complex scenes can be fully represented: slots can even directly discover camouflaged objects and have an excellent understanding of background areas. Future research on the merging of slots and the dynamic number of slots is important in this direction.

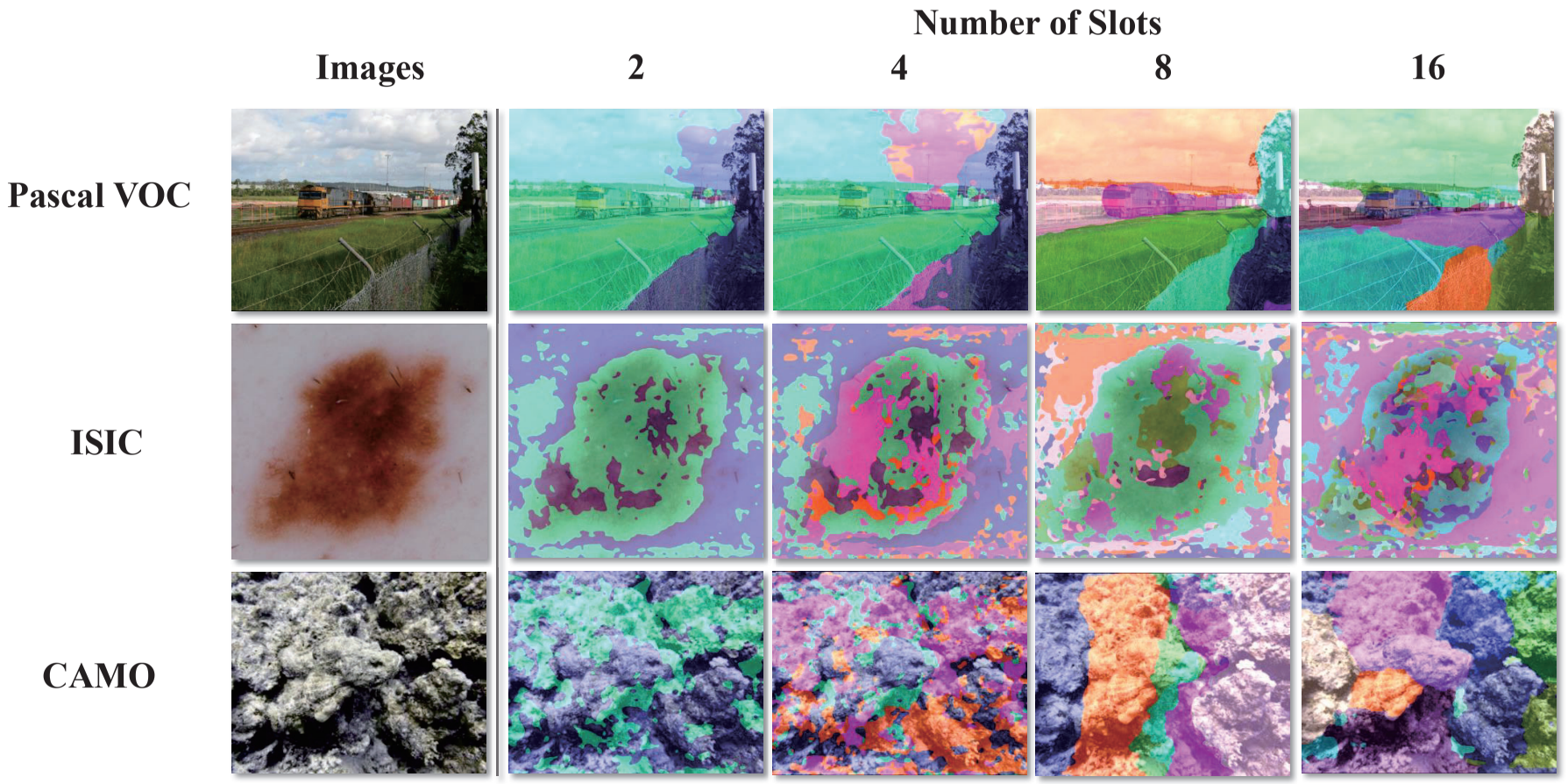

Figure 2. Semantic competition and semantic degradation exists among different numbers of slots.

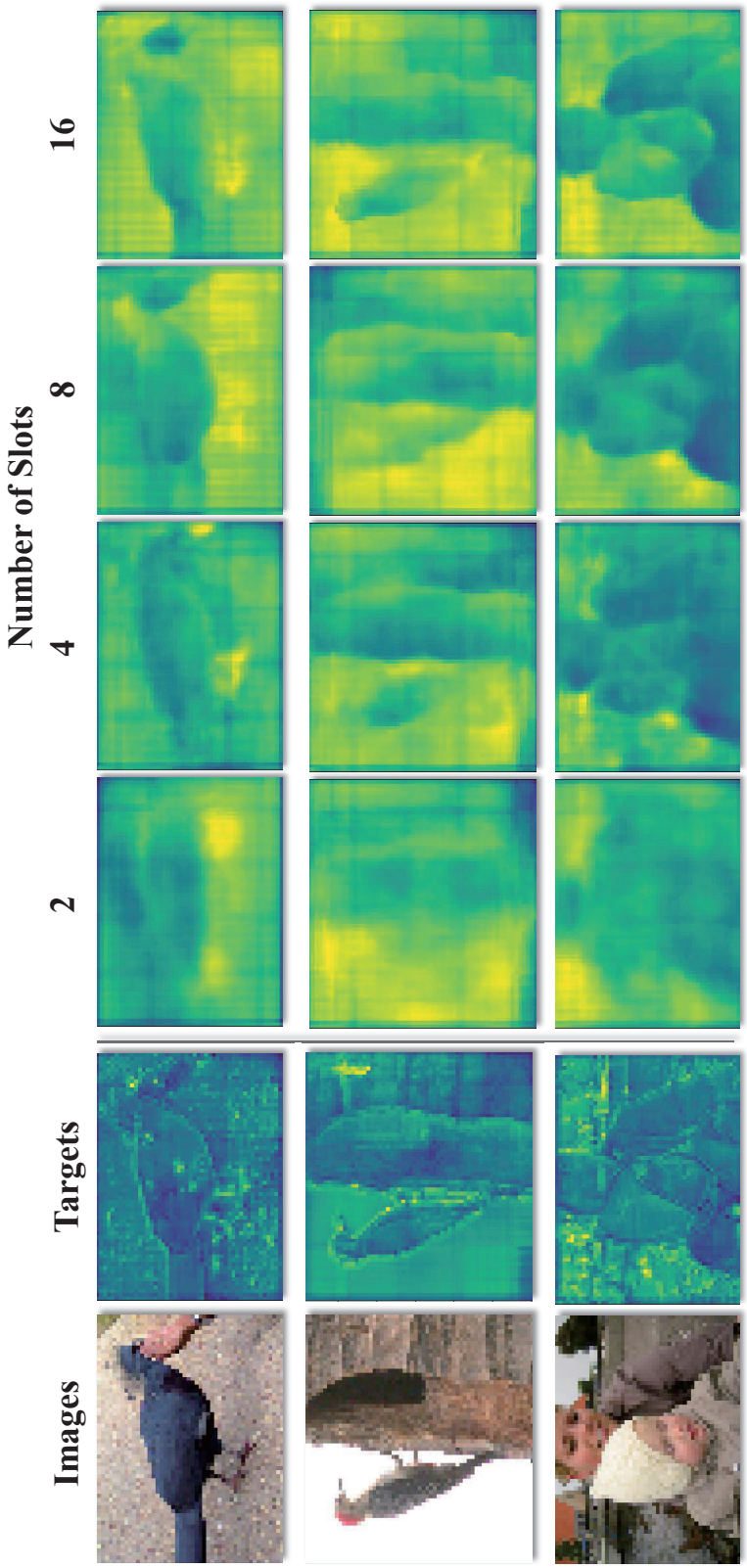

Figure 3. The preference for reconstructing details in object areas with different numbers of slots.

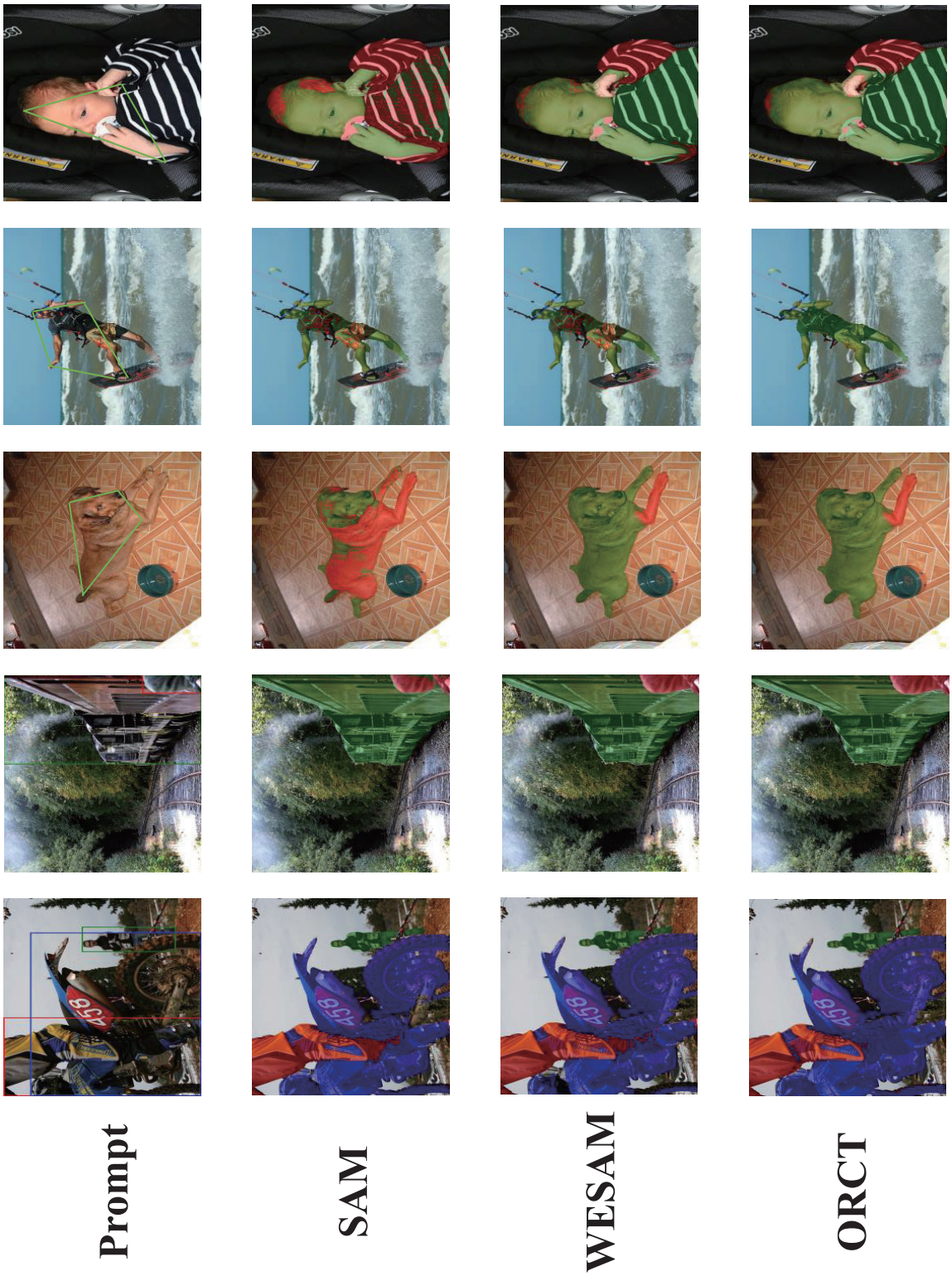

Figure 4. Comparison between OCRT and SOTAs of the fineness of the predicted masks on Pascal VOC.

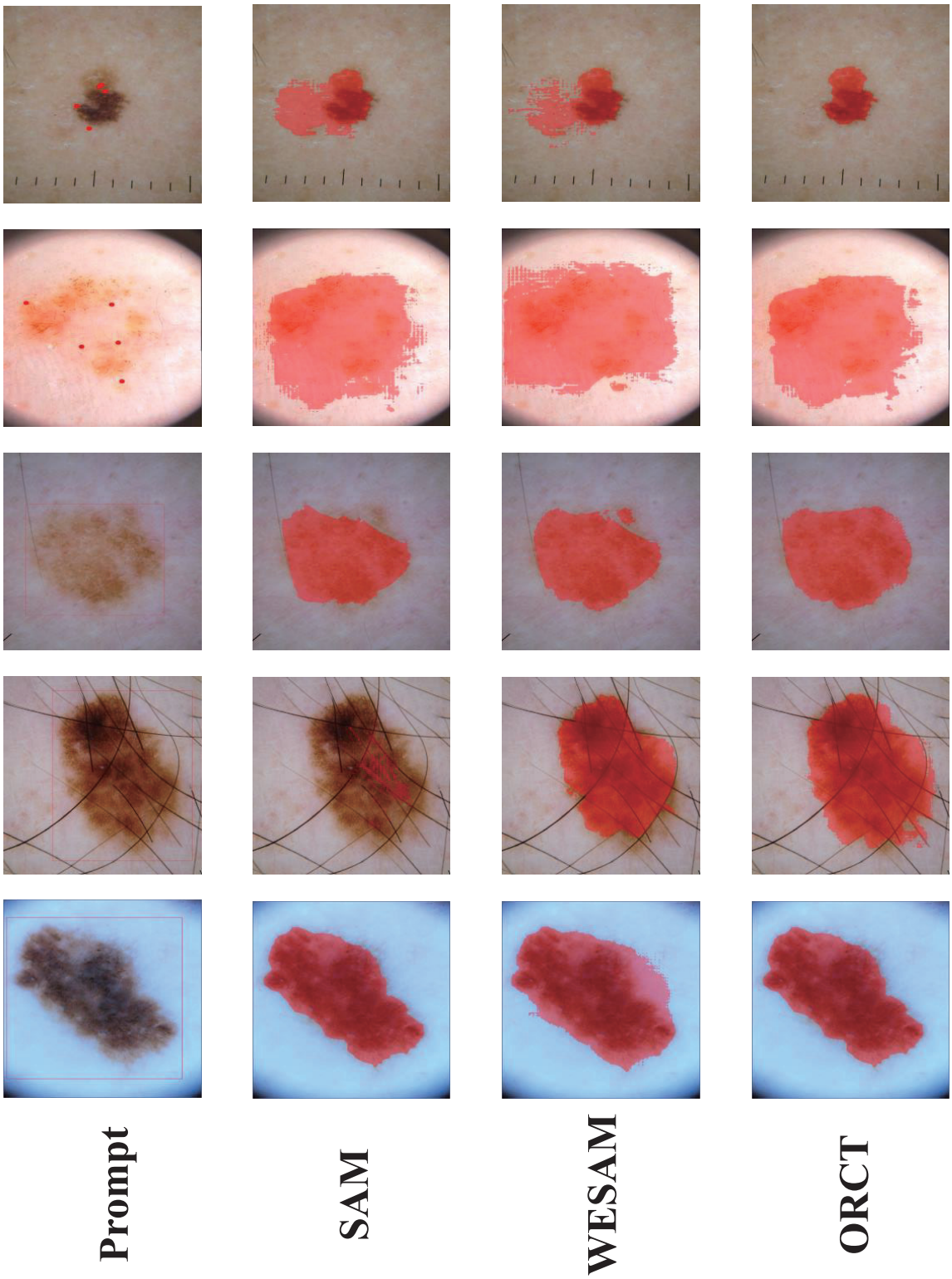

Figure 5. Comparison between OCRT and SOTAs of the fineness of the predicted masks on ISIC.

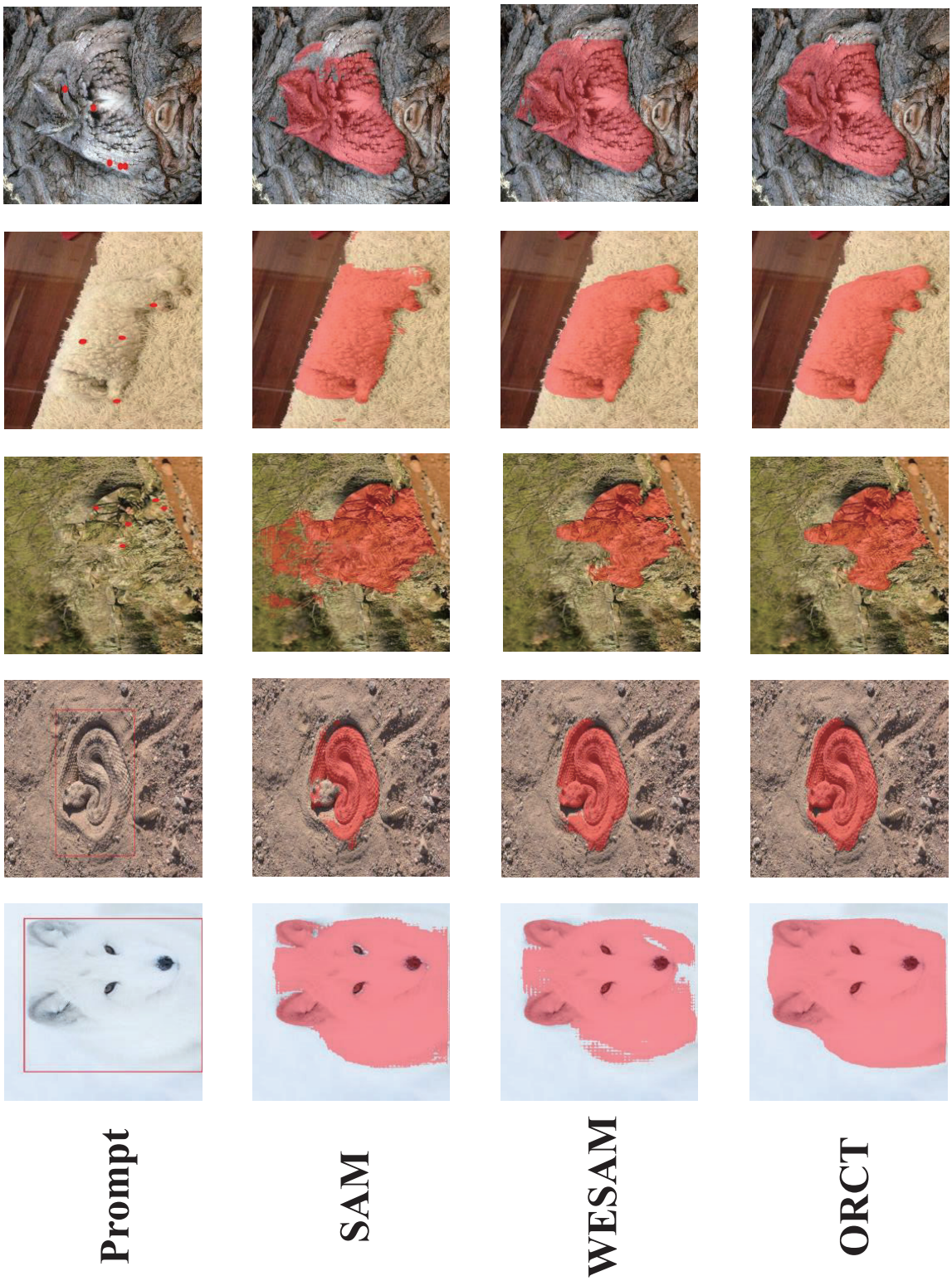

Figure 6. Comparison between OCRT and SOTAs of the fineness of the predicted masks on CAMO.

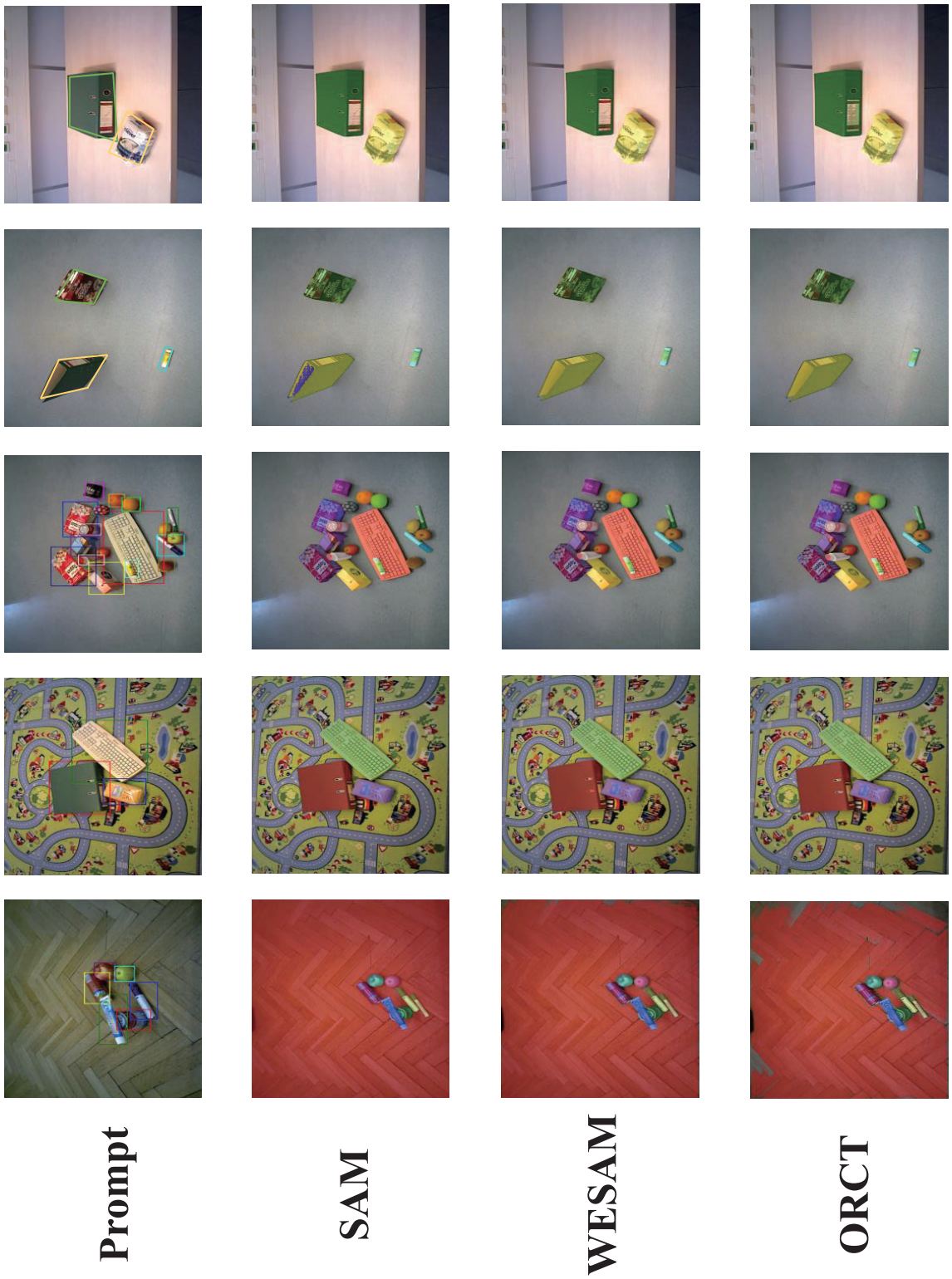

Figure 7. Comparison between OCRT and SOTAs of the fineness of the predicted masks on OCID.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 3

[3] Md Rifat Arefin, Yan Zhang, Aristide Baratin, Francesco Locatello, Irina Rish, Dianbo Liu, and Kenji Kawaguchi. Unsupervised concept discovery mitigates spurious correlations. *arXiv preprint arXiv:2402.13368*, 2024. 2

[4] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 2, 3

[5] Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacks: Adversarial images can control generative models at runtime. *arXiv preprint arXiv:2309.00236*, 2023. 3

[6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2013. 3

[7] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 3

[8] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[9] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 3

[10] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 3

[11] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8340–8349, 2021. 3

[12] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[13] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2013. 3

[14] Alex Krizhevsky. Convolutional deep belief networks on cifar-10. 2010. 3

[15] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 3

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2

[17] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[19] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *ArXiv*, abs/1306.5151, 2013. 3

[20] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*, 2022. 3

[21] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008. 3

[22] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 3

[23] Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3677–3685, 2023. 2

[24] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 43685–43704. PMLR, 2024. 2

[25] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 3

[26] Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018. 3