Recognize Anything: A Strong Image Tagging Model

Youcai Zhang[*1], Xinyu Huang[*1], Jinyu Ma[*1], Zhaoyang Li[*1], Zhaochuan Luo[1], Yanchun Xie[1],
Yuzhuo Qin[1], Tong Luo[1], Yaqian Li[1], Shilong Liu[2], Yandong Guo[3], Lei Zhang[2]
[1]OPPO Research Institute, [2]International Digital Economy Academy (IDEA), [3]AI[2] Robotics
[*]Equal Contribution

(zhangyoucai,huangxinyu2,majinyu,lichaoyang1)@oppo.com

Abstract

We present the Recognize Anything Model (RAM): a strong foundation model for image tagging. RAM makes a substantial step for large models in computer vision, demonstrating the zero-shot ability to recognize any common category with high accuracy. RAM introduces a new paradigm for image tagging, leveraging large-scale image-text pairs for training instead of manual annotations.

The development of RAM comprises four key steps. Firstly, annotation-free image tags are obtained at scale through automatic text semantic parsing. Subsequently, a preliminary model is trained for automatic annotation by unifying the caption and tagging tasks, supervised by the original texts and parsed tags, respectively. Thirdly, a data engine is employed to generate additional annotations and clean incorrect ones. Lastly, the model is retrained with the processed data and fine-tuned using a smaller but higher-quality dataset.

We evaluate the tagging capabilities of RAM on numerous benchmarks and observe impressive zero-shot performance, significantly outperforming CLIP and BLIP. Remarkably, RAM even surpasses the fully supervised manners and exhibits competitive performance with the Google tagging API. We are releasing the RAM at https://recognize-anything.github.io/ to foster the advancements of large models in computer vision.

1. Introduction

Large language models (LLM) trained on large-scale web datasets have sparked a revolution in nature language processing (NLP). These models[20, 5] exhibit impressive zero-shot generalization, enabling them to generalize to tasks and data distributions beyond their training domain. When it comes to computer vision (CV), Segment Anything Model (SAM) [12] has also demonstrated remarkable zero-shot localization abilities through data scaling-up.

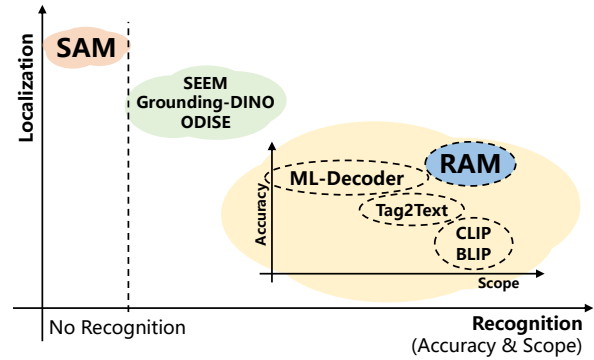However, SAM lacks the capability to output semantic

Figure 1. SAM excels in providing strong localization capabilities, while it falls short when it comes to recognition tasks. In contrast, RAM exhibits exceptional recognition abilities, surpassing existing models in terms of both accuracy and scope.

labels, which is another foundational task on par with localization. Multi-label image recognition, also known as image tagging, aims to provide semantic labels by recognizing multiple labels of a given image. Image tagging is a significant and practical computer vision task, as images inherently contain multiple labels encompassing objects, scenes, attributes, and actions. Regrettably, existing models in multi-label classification, detection, segmentation, and vision-language approaches have exhibited deficiency in tagging, characterized by limited scopes or poor accuracy, as illustrated in Figure 1.

Two core components impede the progress of image tagging. *1)* The difficulty lies in collecting large-scale high-quality data. Specifically, there is a lack of a universal and unified label system and an efficient data annotation engine, capable of semi-automatic or even automatic annotation of large-scale images with a vast number of categories. *2)* There is a lack of efficient and flexible model design that can leverage large-scale weakly-supervised data to construct an open-vocabulary and powerful model.

To address these key bottlenecks, this paper introduces

Figure 2. The comparison of recognition capability among tagging models. RAM recognize more valuable tags than other models without missing important part. ML-Decoder and Google tagging API tend to output redundant tags (*e.g.*, "human head") or less relevant tags (*e.g.*, "property") tags. BLIP's tag recall is limited as it relies on caption generation.

the Recognize Anything Model (RAM), a strong foundation model for image tagging. RAM overcomes the challenges related to data, including label system, dataset and data engine, as well as the limitations in model design.

Label System: We begin by establishing a universal and unified label system. We incorporate categories from popular academic datasets (classification, detection, and segmentation) as well as commercial tagging products (Google, Microsoft, Apple). Our label system is obtained by merging all the public tags with the common tags from texts, thus covering most of common labels with a moderate amount of 6,449. The remaining open-vocabulary labels can be identified through open-set recognition.

Dataset: How to automatically annotate large-scale images with the label system is another challenge [30]. Drawing inspiration from CLIP [22] and ALIGN [11], which leverage publicly available image-text pairs at scale to train powerful visual models, we adopt similar datasets for image tagging. To utilize these large-scale image-text data for tagging, following [9, 10], we parse the texts and obtain the image tags through automatic text semantic parsing. This process allows us to obtain a diverse collection of

annotation-free image tags in accordance with image-text pairs.

Data Engine: However, the image-text pairs from the web are inherently noisy, often containing missing or incorrect labels. To enhance the quality of annotations, we design a tagging data engine. In addressing missing label, we leverage existing models to generate additional tags. With regards to incorrect labels, we first localize specific regions corresponding to different tags within the image. Subsequently, we employ region clustering techniques to identify and eliminate outliers within the same class. Furthermore, we filter out tags that exhibit contrary predictions between whole images and their corresponding regions, ensuring a cleaner and more accurate annotation.

Model: Tag2Text [10] has demonstrated superior image tagging capabilities by the integration of image tagging and caption, employing a lightweight recognition decoder [18] in conjunction with the original image encoder. However, the effectiveness of Tag2Text is limited to recognizing fixed and predefined categories. In contrast, RAM enable generalization to previously unseen categories by incorporating semantic information into label queries. This model design
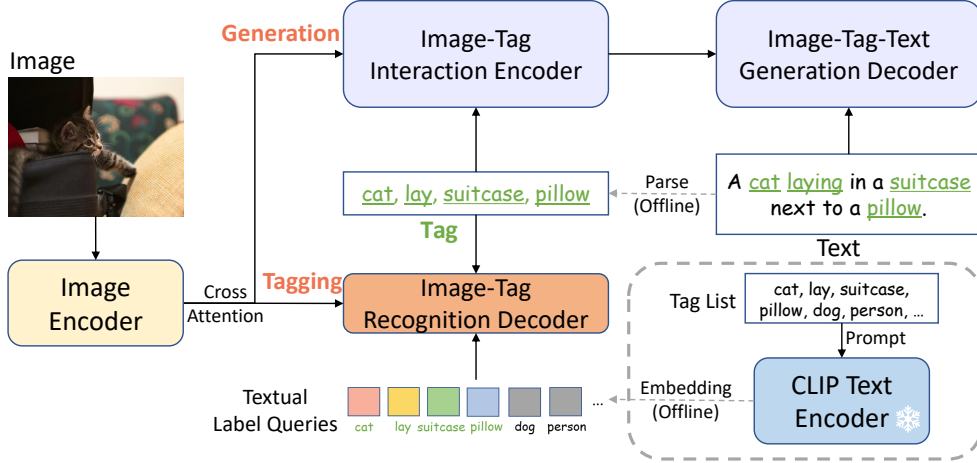
Figure 3. Illustration of RAM's model architecture. Large-scale image tags are obtained from image-text pairs through automatic text semantic parsing. With image-tag-text triplets, RAM unifies the captioning and tagging tasks. Furthermore, RAM introduces an off-the-shelf text encoder to encoder tags into textual label queries with semantically-rich context, empowering the generalization to unseen categories in training stage.

allows RAM to empower the recognition capabilities of any visual dataset, underlining its potential for diverse applications.

Benefitting from the large-scale, high-quality image-tag-text data and the synergistic integration of tagging with caption, we develop a strong recognize anything model (RAM). RAM represents a new paradigm for image tagging, demonstrating that a general model trained on noisy, annotation-free data can outperform fully supervised models. The advantages of RAM are summarized as follows:

- **Strong and general.** RAM exhibits exceptional image tagging capabilities with powerful zero-shot generalization as illustrated in Figure 2;

- **Reproducible and affordable.** RAM requires Low reproduction cost with open-source and annotation-free dataset. Moreover, the strongest version of RAM only requires 3-days 8 A100 GPUs training;

- **Flexible and versatile.** RAM offers remarkable flexibility, catering to various application scenarios. By selecting specific classes, RAM can be directly deployed to address specific tagging needs. Furthermore, when combined with localization model (Grounding DINO and SAM), RAM forms a strong and general pipeline for visual semantic analysis.

2. Recognize Anything Model

2.1. Model Architecture

As illustrated in Figure 3, we extract image tags through text semantic parsing, providing a large-scale of tags without expensive manual annotations. The overall architecture

of RAM is similar to that of Tag2Text[10], which consists of three key modules: an image encoder for feature extraction, following a image-tag recognition decoder [18] for tagging, and a text generation encoder-decoder for captioning. The image features interact with tags by the cross-attention layers in the image-tag interaction encoder and recognition decoder. In the training stage, the recognition head learns to predict the tags parsed from text, while in the inference stage, it serves as a image-to-tags bridge by predicting tags which provide a more explicit semantic guidance to image captioning.

Compared with Tag2Text [10], RAM's core advancement in model design is the introduction of open-vocabulary recognition. Tag2Text can only recognize the categories that it has seen during training, while RAM can recognize any category.

2.2. Open-Vocabulary Recognition

Textual Label Queries. Inspired by [23, 28], the pivotal enhancement lies in the incorporation of semantic information into the label queries of the recognition decoder, which facilitates generalization to previously unseen categories in training stage. To achieve this, we utilize an off-the-shelf text encoder to encode the individual tags from the tag list, consequently providing textual label queries with semantically-rich context. In contrast, the label queries employed in the original recognition decode [10, 18] are randomly learnable embeddings, lacking the semantic relationship with unseen categories, thus are confined to predefined seen categories.

Implementation Details. We adopt Swin-transformer [19] as the image encoder, as it demonstrated better performance
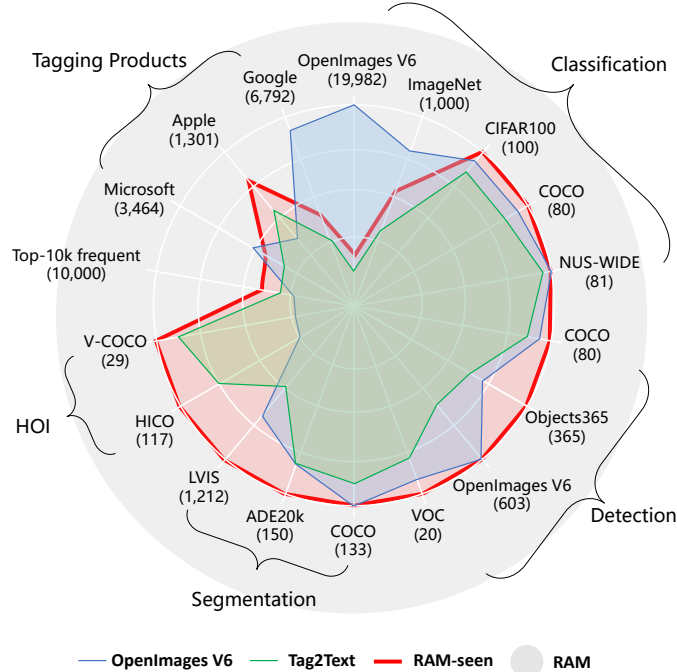
Figure 4. Recognition Scopes of different tagging models. Tag2Text recognizes 3,400+ fixed tags. RAM upgrades the number to 6,400+, covering more valuable categories than OpenImages V6. With open-set capability, RAM is feasible to recognize any common category.

than naive ViT in both vision-language [10] and tagging domains [18]. The encoder-decoder used for text generation are 12-layer transformers, and the tag recognition decoder is a 2-layer transformer. We utilize the off-the-shelf text encoder from CLIP [22] and perform prompt ensembling [22] to obtain textual label queries. We also adopt the CLIP image encoder to distill image feature, which further improves the model's recognition ability for unseen categories via image-text feature alignment.

2.3. Model Efficiency

Training Phase. RAM is pretrained on large-scale datasets with a resolution of 224 and fine-tuning at a resolution of 384 using small and high-quality datasets. Empirical evidence suggests that RAM converges rapidly, often with convergence achieved after a minimal number of epochs (typically less than 5 epochs). This accelerated convergence enhances the reproducibility of RAM with limited computational resources. To illustrate, the version of RAM pretrained on 4 millions necessitate 1-day of computation, and the strongest version of RAM pretrained on 14 million images necessitate a mere 3-days of computation on 8 A100 GPUs.

Inference Phase. The lightweight image-tag recognition decoder effectively ensures the inference efficiency of RAM on image tagging. Furthermore, we eliminate the self-attention layers from the recognition decoder, which not

only further improves efficiency but also circumvents potential interference between label queries. Consequently, instead of fixed categories and quantities, RAM allows customization of label queries for any category and quantity which want to automatically recognize, enhancing its utility across various visual tasks and datasets.

3. Data

3.1. Label System

This work adopts three guiding principles for the formulation of the label system: *1)* Tags that frequently appear in image-text pairs are more valuable due to their representational significance in image description. *2)* A variety of domains and contexts should be represented in the tags. Our conception of a *tag* includes objects, scenes, attributes, and actions from a range of sources, which aids model generalization to complex, unseen scenarios. *3)* The quantity of tags needs to be moderate. Excessive tag numbers can incur heavy annotation costs.

Initially, we parse 14 million sentences from our pretraining datasets into tags by utilizing a SceneGraph-Parser [25] with minor modifications. We then hand-pick tags from the top-10k most frequently occurring tags. Our selection intentionally covers tags from numerous popular datasets for classification, detection, and segmentation, as illustrated in Figure 4. While most are fully covered,

4

exceptions include ImageNet and OpenImages V6, due to their unusual tag presence. Additionally, we partially cover tags from leading tagging products, which were obtained via public APIs [2, 3, 1] using open-source images. Consequently, RAM can recognize up to 6449 fixed tags, which is substantially more than Tag2Text [10], and includes a higher proportion of valuable tags. To reduce redundancy, we collected synonyms via various methodologies including manual checks, referring to WordNet [7], translating and merging tags, etc. Tags within the same synonym group are assigned the same tag ID, resulting in 4585 tag IDs in the label system.

3.2. Datasets

Similar to BLIP [15] and Tag2Text [10], we pre-train our model on widely-used open-source datasets. 4 million (4M) image and 14 million (14M) image settings are adopted. The 4M setting includes two human-annotated datasets, COCO [16] (113K images, 557K captions) and Visual Genome [13] (101K images, 822K captions), along with two large-scale web-based datasets, Conceptual Captions [6] (3M images, 3M captions) and SBU Captions [21] (849K images, 849K captions). The 14M setting builds upon the 4M, with the addition of Conceptual 12M [6] (10M images, 10M captions).

3.3. Data Engine

Given the predominantly open-source nature of our training datasets, which are largely crawled from the Internet, we encounter a non-negligible amount of missing and incorrect labels. To mitigate this, we design an automatic data engine to generate additional tags and clean erroneous ones.

Generation. Our initial step involves training a baseline model using the captions and tags parsed from these captions, similar to the approach used in Tag2Text [10]. We then leverage this baseline model to supplement both the captions and tags, utilizing its generative and tagging capabilities, respectively. The original captions and tags, in conjunction with the generated captions, corresponding parsed tags, and generated tags, are merged to form a temporary dataset. This step significantly expands the number of tags in the 4M image dataset from 12 million to 39.8 million.

Cleaning. To address the issue of incorrect tags, we initially employ Grounding-Dino [29] to identify and crop regions corresponding to a specific category within all images. Subsequently, we cluster the regions from this category based on K-Means++ [4] and eliminate the tags associated with the outlier 10%. Simultaneously, we also remove tags without the prediction of this specific category using the baseline model. The motivation is that the precision of tagging models can be improved by predicting regions rather than whole images.

4. Experiment

4.1. Experimental Setting

Test Benchmarks. We conducted a comprehensive evaluation of the models on various popular benchmark datasets across different computer vision tasks, including classification, detection, and segmentation, as summarized in Table 1 For classification, we adopt the OpenImages V6 [14], which contains 9605 categories. However, due to issues of missing labels and incorrect annotations within the OpenImages dataset, we curated two high-quality subsets: OpenImages-common, comprising 214 well-annotated common categories, and OpenImages-rare, consisting of 200 categories not included in our label system for open-set experiments. Additionally, to facilitate better zero-shot evaluation, we employed an internal test set known as OPPO-common, which exhibits high annotation quality.

For detection and segmentation datasets, we selected the widely recognized COCO [16] and ADE20k [31, 32] datasets. In these datasets, we focused solely on semantic labels as image-level tagging ground-truth, disregarding bounding boxes and masks. It is important to note that ADE20k contains plenty of very small ground-truth annotations and ambiguous categories that deviate from mainstream concepts, *e.g., "buffet"*. Thus, we created a subset of ADE20k called ADE20k-clean by removing a few small targets and ambiguous categories.

Evaluation Metrics. To assess the performance of the models, we employed various evaluation metrics. Mean Average Precision (mAP) was used for reporting results in ablation experiments and comparisons with other classification models. For models where mAP was not available, we utilized Precision/Recall metrics and manually adjusted the threshold of different models to ensure comparability across evaluations.

Table 1. Details of test benchmarks.

Type	Dataset	#Category	#Image
	OPPO-common	200	44,606
Cls.	OpenImages-common [14]	214	57,224
	OpenImages-rare [14]	200	21,991
Det.	COCO-80 [16]	80	5,000
	COCO-133 [16]	133	5,000
Seg.	ADE20k [31, 32]	150	2,000
	ADE20k-clean [31, 32]	143	2,000

4.2. Comparison with SOTA Models

Comparison with Multi-Label Classification Models. We compare RAM with state-of-the-art (SOTA) models in multi-label classification, as show in Table 2. Generally, a generalist model typically lacks expertise in specific domains, whereas an expert model struggles to gen-

Table 2. Comparison with classification models in mAP. Cells marked with ✗ means unable to evaluate on such setting. Cell background color: Green means fully supervised learning; Blue means zero-shot performance; Yellow denotes that the model has seen the corresponding training images, but not the annotations. Notably, RAM's zero-shot generalization to OpenImages-common is superior to ML-Decoder's full supervision. RAM can also recognize categories in OpenImages-rare, even though it has not seen them during training.

Methods	Tags‡	Multi-label Classification			Detection	Segmentation		
		OPPO-common	OpenImages-common	OpenImages-rare	COCO-80	COCO-133	ADE20k	ADE20k-clean
ML-Decoder [23]	33.9M	82.4†	85.8	79.5	72.8†	✗	✗	✗
MKT [8]	0.6M	78.2	77.8	63.5	62.9	51.0	37.1	38.4
Tag2Text-4M [10]	11.4M	83.0	82.9	✗	78.3†	66.9†	✗	✗
Tag2Text-14M [10]	33.6M	85.4	83.4	✗	78.2†	67.1†	✗	✗
RAM-4M	39.3M	85.6	86.0	66.7	79.0	68.3	51.5	53.2
RAM-14M	119.9M	86.9	86.5	69.2	80.6	69.4	55.4	56.9

† A few categories that are not supported by the model are excluded when calculating mAP.
‡ The total number of common tags that co-occur in the training set and the top-10k parsed tags.

Table 3. Comparison with detection, segmentation and vision-language models in Precision/Recall. Cells marked with ✱ means poor performance in large-sized categories, or long inference time due to the high image resolution, *e.g.,* 1024 for ODISE. Notably, RAM outperforms CLIP and BLIP with large margins in common categories.

Methods	Backbone	Multi-label Classification			Detection	Segmentation		
		OPPO-common	OpenImages-common	OpenImages-rare	COCO-80	COCO-133	ADE20k	ADE20k-clean
Grounding-DINO [17]	Swin-B	✱	✱	✱	83.1 / 86.9	66.4 / 48.3	34.3 / 24.7	35.6 / 26.0
ODISE [26]	Diffusion-v3	✱	✱	✱	78.5 / 85.9	71.1 / 80.2	47.4 / 48.0	48.2 / 50.3
SEEM [33]	FocalNet-L	✗	✗	✗	75.7 / 67.8	71.8 / 61.0	✗	✗
CLIP-400M [22]	ViT-B	76.6 / 54.1	77.9 / 52.9	67.5 / 46.5	64.0 / 38.7	47.8 / 36.4	30.3 / 5.3	31.0 / 5.5
BLIP-129M [15]	ViT-B	76.7 / 57.5	78.6 / 55.1	65.2 / 46.5	67.0 / 39.0	53.8 / 34.6	28.5 / 8.8	29.1 / 9.3
Tag2Text-4M [10]	Swin-B	76.6 / 74.8	75.9 / 71.9	✗	80.5 / 66.1†	71.2 / 54.0†	✗	✗
Tag2Text-14M [10]	Swin-B	77.9 / 79.4	76.4 / 73.3	✗	80.1 / 64.5†	71.2 / 53.2†	✗	✗
RAM-4M	Swin-B	78.4 / 75.2	79.2 / 73.7	53.9 / 48.4	81.8 / 66.1	74.3 / 54.0	47.0 / 47.6	47.8 / 50.3
RAM-14M	Swin-L	78.8 / 79.4	80.3 / 75.7	53.8 / 54.3	82.9 / 66.4	74.3 / 54.1	53.2 / 50.0	53.7 / 52.2

† A few categories that are not supported by the model are excluded when calculating precision and recall.

eralize beyond its specialized field. Specifically, the supervised expert model ML-Decoder [23] excels in its designated domain of expertise, OpenImages, but faces challenges in generalizing to other domains and unseen categories. MKT [8] is a generalist model in tagging by transferring the knowledge from CLIP, fails to achieve satisfactory accuracy across all domains. Tag2Text [10] is powerful at zero-shot tagging, but it lacks the ability to handle open-set scenarios.

RAM exhibits impressive tagging abilities, showcasing impressive accuracy and broad coverage. Particularly noteworthy is the performance of RAM-4M, which surpasses ML-Decoder on the OpenImages-common dataset. While ML-Decoder relies on 9 million annotated images from OpenImages, our RAM-4M achieves higher accuracy with a training set of 4 million annotation-free image-text data. This improvement is attributed to the utilization of 39.3 million common tags derived from the 4 million images, outperforming ML-Decoder trained with only 33.9 million common tags from 9 million images. Moreover, RAM can recognize any common category by leveraging a vast range of 6,400+ seen common categories, coupled with its open-vocabulary ability.

Comparison with Detection and Segmentation Models. The comparison in Table 3 reveals that supervised detection and segmentation models excel in specific domains such as COCO datasets, which encompass a limited number of categories. However, these models face challenges when it comes to recognizing a larger number of categories. On the one hand, they take much more computational overheads as they requires more complex network and larger input image sizes for extra localization task. Especially, ODISE [26] takes long inference time due to its adoption of the diffusion model and large input image resolution. On the other hand, the scalability of training data for detection and segmentation is limited, resulting in poor generalization performance for these models. Although Grounding-DINO [17] serve as a generalist model, it struggles to achieve satisfactory

Table 4. Ablation study of RAM model based on Tag2Text baselines. *"Seen Categories"* refers to the number of training categories.*"Captioning"* refers to the joint training of captioning and tagging tasks. *"Textual Queries"* refers to using a text encoder to generate label queries possessing semantic information. *"Distillation"* refers to image feature distillation using CLIP's image encoder.

Case	Seen Categories	Captioning	Textual Queries	Distillation	OPPO -common	OpenImages -common	-rare
Tag2Text	3,429				80.60	83.52	✗
	3,429	✓			81.37	84.04	✗
(a)	3,429	✓	✓		81.22	84.09	60.99
(b)	3,429	✓	✓	✓	81.70	84.16	61.88
(c)	**6,449**	✓	✓	✓	80.27	83.09	63.54

Table 5. Ablation study of data engine. *"Parsing"* means the training tags parsed from the captions. *"Generation"* means the supplementation of captions and tags. *"Cleaning"* refers to data cleaning. *"Fine-tuning"* refers to fine-tuning the pre-trained model with COCO.

Backbone	Pre-train #Images	#Tags	Parsing	Generation	Cleaning	Fine-tuning	OPPO -common	OpenImages -common	-rare
Swin-Base	4M	12.0M	✓				80.27	83.09	63.54
	4M	41.7M	✓	✓			82.50	84.27	67.17
	4M	39.8M	✓	✓	✓		82.83	84.94	66.88
	4M	39.8M	✓	✓	✓	✓	85.56	86.01	66.74
	14M	121.5M	✓	✓	✓		83.52	85.39	68.54
	14M	121.5M	✓	✓	✓	✓	86.47	86.50	68.79
Swin-Large	14M	121.5M	✓	✓	✓		83.26	84.94	68.60
	14M	121.5M	✓	✓	✓	✓	86.92	86.46	69.21

performance for large-sized categories. In contrast, RAM demonstrates impressive open-set ability, surpassing existing detection and segmentation models. RAM showcases its capability to generalize across a broader range of categories, providing a robust solution for the challenges faced by conventional detection and segmentation models.

Compared with Vision-Language Models. Despite the open-set recognition capabilities of CLIP [22] and BLIP [15], these models suffer from subpar accuracy. Furthermore, their interpretability is limited, as they rely on cosine similarity computations of dense embeddings for image-text pairs. In contrast, RAM exhibits superior performance, surpassing CLIP and BLIP by a significant margin, with accuracy increases of over 20% observed across almost all datasets. However, it is worth noting that RAM performs slightly worse than CLIP and BLIP in the case of OpenImages-rare dataset. We attribute this discrepancy to the smaller training dataset utilized for RAM and the relatively less emphasis placed on rare classes during training.

4.3. Model Ablation Study

In Table 4, we study the impact of various model improvements to RAM based on Tag2Text [10] and make the following key observations. *1)* The training integration of captioning and tagging can promote the tagging ability. *2)* The open-set recognition capability can be achieved

through textual queries by CLIP [22], but has little impact on the seen categories in training. *3)* The expansion of the label system introduces minimal impact on existing categories, which can be attributed to the additional categories increases the difficulty of model training. However, this expansion concurrently enhances the model's coverage and enhances the open-set ability of unseen categories.

4.4. Data Engine Ablation Study

We present an ablation study of the data engine in Table 5. The findings are summarized as follows: *1)* Adding more tags from 12.0M to 41.7M significantly improves model performance across all test sets, indicating the severe missing label problem in original datasets. *2)* Further cleaning the tags of some categories results in a slight increase in performance on the OPPO-common and OpenImages-common test sets. Limited by the inference speed of Grounding-Dino, we only conduct cleaning process for 534 categories. *3)* Scaling up the training images from 4M to 14M brings remarkable improvements across all test sets. *4)* Employing a larger backbone network leads to a slight improvement on OpenImages-rare and even slightly inferior performance on common categories. We attribute this phenomenon to our insufficient resources available for conducting hyper-parameter search. *5)* Fine-tuning with tags parsed from the COCO Caption dataset [16] demonstrates remarkable increases in performance on the OPPO-common and

OpenImages-common test sets. The COCO Caption dataset provides five descriptive sentences for each image, offering a comprehensive description that approximates a complete set of tag labels.

5. Conclusion

We present the Recognize Anything Model (RAM), a strong foundation model designed for image tagging, which heralds a novel paradigm in this field. RAM demonstrate the zero-shot ability to recognize any category with high accuracy, surpassing the performance of both fully supervised models and existing generalist approaches like CLIP and BLIP. RAM represents a considerable advancement for large-scale models in the field of computer vision, holding the potential to empower the recognition capabilities of any visual tasks or datasets.

There still exists room for further refinement of RAM. For example, scaling up the training dataset beyond 14 million images to better cover diverse domains, multiple rounds of data engine, increasing the backbone parameters to enhance the model capacity.

Limitations. Similar to CLIP, the current version of RAM efficiently recognizes common objects and scenes, yet struggles with abstract tasks like object counting. Moreover, zero-shot RAM's performance lags behind task-specific models in fine-grained classifications, such as differentiating between car models or identifying specific flower or bird species. It is also noteworthy that RAM is trained on open-source datasets and could potentially reflect dataset biases.

References

[1] Apple Developer. https://developer.apple.com/documentation/vision.

[2] Google Cloud vision API. https://cloud.google.com/vision.

[3] Microsoft Azure cognitive service. https://azure.microsoft.com/zh-cn/products/cognitive-services/vision-services/.

[4] David Arthur and Sergei Vassilvitskii. K-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2007.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pretraining to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[7] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[8] Sunan He, Taian Guo, Tao Dai, Ruizhi Qiao, Bo Ren, and Shu-Tao Xia. Open-vocabulary multi-label classification via multi-modal knowledge transfer. *CoRR*, abs/2207.01887, 2022.

[9] Xinyu Huang, Youcai Zhang, Ying Cheng, Weiwei Tian, Ruiwei Zhao, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Xiaobo Zhang. Idea: Increasing text diversity via online multi-label recognition for vision-language pre-training. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4573–4583, 2022.

[10] Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*, 2023.

[11] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[14] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *arXiv:2201.12086 [cs]*, Feb. 2022. arXiv: 2201.12086.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[17] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[18] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021.

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer:

Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[20] OpenAI. Gpt-4 technical report, 2023.

[21] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[23] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben Baruch, and Asaf Noy. Ml-decoder: Scalable and versatile classification head. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 32–41. IEEE, 2023.

[24] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, Xiaogang Wang, and Yu Qiao. Internimage: Exploring large-scale vision foundation models with deformable convolutions. *CoRR*, abs/2211.05778, 2022.

[25] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6609–6618, 2019.

[26] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. *arXiv preprint arXiv:2303.04803*, 2023.

[27] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. In *NeurIPS*, 2022.

[28] Yuhang Zang, W ei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 106–122. Springer, 2022.

[29] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *CoRR*, abs/2203.03605, 2022.

[30] Youcai Zhang, Yuhao Cheng, Xinyu Huang, Fei Wen, Rui Feng, Yaqian Li, and Yandong Guo. Simple and robust loss design for multi-label learning with missing labels. *arXiv preprint arXiv:2112.07368*, 2021.

[31] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[32] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.

[33] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *CoRR*, abs/2304.06718, 2023.