Training-Free Zero-Shot Semantic Segmentation with LLM Refinement

Yuantian Huang[1,2]
huang_yuantian@cyberagent.co.jp
Satoshi Iizuka[2]
iizuka@cs.tsukuba.ac.jp
Kazuhiro Fukui[2]
kfukui@cs.tsukuba.ac.jp

[1] CyberAgent, Inc. Tokyo, Japan
[2] University of Tsukuba
Tsukuba, Japan

Abstract

Semantic segmentation models are predominantly based on supervised or unsupervised learning methodologies, which require substantial effort in annotation or training. In this study, we present a novel framework that leverages multiple pre-trained foundational models for semantic segmentation tasks on previously unseen images, eliminating the need for additional training. Our framework utilizes image recognition models to transform an input image into textual information. This text information is then used to engage an advanced Large Language Model (LLM) to predict the presence of specific classes within the given image. The labels predicted by the LLM are subsequently processed through an open-set detection and segmentation model to generate our ultimate outcomes. To ensure that the class information is precisely aligned with the intended context, we incorporate both a pre-refinement and a post-refinement procedure utilizing the LLM. The segmentation model is further modified to accept both bounding boxes and point prompts, resulting in higher accuracy than original usage that only accepts bounding boxes as input. Our proposed framework accomplishes training-free zero-shot semantic segmentation, requiring only the input image and customizable target classes for different scenarios as inputs. Experiments indicate that the proposed framework demonstrates the capacity to execute semantic segmentation effectively across various datasets. Notably, our results surpass those of existing unsupervised models despite the absence of any training procedure.

1 Introduction

Semantic Segmentation is a computer vision task that categorizes each pixel within an image into a semantic class. Traditional approaches are predominantly based on either supervised or unsupervised learning methodologies. However, the effectiveness of these segmentation methods heavily depends on the availability of extensive human-annotated data for training these models. In contrast, our proposed frameworks leverage a set of pre-trained models to avoid the need for training from annotated datasets. Furthermore, additional LLM refinement processes enable our method to adapt to various tasks with minimal effort.

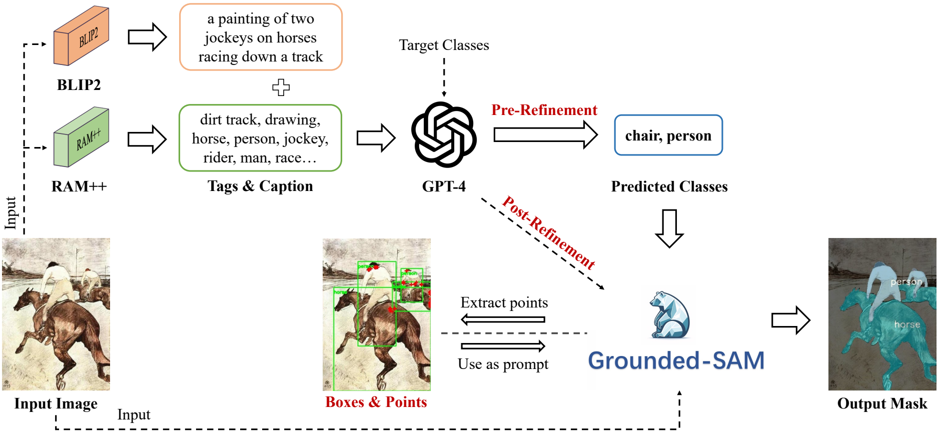


Figure 1: **Overview of our proposed framework.** The input image is processed by image recognition models to convert it into textual information, which is subsequently refined by GPT-4. The system prompt for GPT-4 is automatically generated based on the target classes. Following a pre-refinement process, relevant labels are extracted and input into the detection and segmentation model, ensuring the final output mask conforms to the target classes.

Our framework begins with image recognition models to transform an input image into textual information. We adopt the Recognize Anything Plus Model (RAM++) [14], an open-set image tagging model, to extract relevant tags from the input image. Additionally, we utilize the BILP-2 [20], a leading image caption model to generate a text caption from the input image. After combining these two pieces of textual information, they are then used to engage an advanced Large Language Model (LLM), GPT-4 [24], to predict the presence of specific classes within the given image.

The predicted classes are subsequently processed through an open-set semantic segmentation model, Grounded-SAM [29], to generate our ultimate outcomes. Grounded-SAM uses Grounding DINO [21] as an open-set object detector to combine with the segment anything model (SAM) [17]. Despite its superior performance in open-set semantic segmentation tasks, adapting this model to tasks with closed-set semantic segmentation proves challenging. To bridge the gap between extracted textual information and predefined semantic classes, ensuring that the class information aligns precisely with the intended context, we implement two refinement processes that comprise both a pre-refinement and a post-refinement procedure using a GPT-4 model.

By leveraging these robust pre-trained models, our proposed framework is able to perform high-accuracy semantic segmentation tasks without any preliminary training. Apart from the input image, our framework only requires target classes for a closed-set setting to adapt to previously unseen scenarios, including realistic scenes or artistic images. An overview of our proposed framework is shown in Fig. 1. The experiments demonstrate that our proposed framework surpasses existing unsupervised methods. We offer both qualitative and quantitative comparisons between our method and baseline methods. Additionally, the ablation study confirms the effectiveness of each component in our framework.

This paper presents the following contributions:

- A novel framework that is able to perform semantic segmentation tasks without the requirement for annotated datasets and training.

Table 1: **Comparison of Semantic Segmentation Methods.**

Methods	Paired Dataset Unrequired	Training Unrequired	Customizable Target Classes
Supervised	×	×	×
Unsupervised	✓	×	×
Zero-Shot Transfer	✓	×	×
Training-Free Zero-Shot (ours)	✓	✓	✓

- Two refinement processes convert textual information into predefined semantic classes to improve the overall accuracy.
- Jointly use points and boxes as prompts to improve the segmentation accuracy.
- Customizable semantic classes could be easily specified based on prompt templates, allowing adaptation to various unseen scenarios.

2 Related Work

2.1 Semantic Segmentation

Semantic Segmentation is an essential task in the field of computer vision, involving the categorization of image pixels into semantic classes. Previous approaches are mainly based on Convolutional Neural Networks (CNNs), with significant contributions from extended works including UNet [30], DeepLab series [5, 6, 7] and, more recently, Transformer-based [9, 31, 36]. However, most advanced models are mainly supervised methods, necessitating extensive human annotation efforts. There have been attempts with weakly-supervised and unsupervised models [8, 16, 34, 35, 39] that require limited or no annotations, though they typically yield lower accuracy. Zero-shot transfer learning methods [1, 40] train on seen labels and then apply shared knowledge to segment unseen labels. More recently, the surge in Vision-Language Models has promoted the advent of potent open-set approaches [17, 18, 33, 37, 41, 42] are continually emerging. These models exhibit a solid capability to segment unseen images across varied scenarios. However, closed-set problems are still mainstream in real-world applications. This paper focuses on leveraging robust foundational models to efficiently perform closed-set semantic segmentation without any training process or additional data. A comprehensive comparison of various methodologies is summarized in Table 1.

2.2 Foundational Models

Vision-language Models combines computer vision and natural language processing capabilities. A classic framework including an image encoder, a text encoder, and a methodology to leverage embedding from the two encoders. Recent progress start with CLIP [28], which learns from generic visual-textual representations to perform great potential in a wide set of tasks by leveraging pre-trained knowledge. Extended works including BLIP [19] and BLIP2 [20]. Specifically, other than image caption model, RAM [38] and RAM++ [14] server as image tagging models that only output relevant tags from an input image. This ability revokes a substantial potential for connecting segmentation models. This study uses RAM++ and BLIP2 models as our image recognition components.

Large Language Models The significance of Large Language Models (LLMs) is growing not only as a research area but also in our everyday lives. The explosion in popularity of LLMs began with the GPT family [23, 24, 26, 27], showcasing their impressive capabilities in understanding and generating text. However, a limitation arises from the fact that recent powerful GPTs are closed-sourced, which restricts their reproducibility and transparency. In contrast, models from the LLama series [22, 32] offer open-sourced alternatives comparable to the capabilities of GPT models. While our primary focus in this work is leveraging GPT-4 for its superior performance, we also include a comparison with Llama-3 in our ablation study to provide a broader perspective on available LLM technologies.

Segmentation and Detection Models. SAM [17] offers a novel approach to accurately segmenting unseen images in a zero-shot manner, which has been further enhanced in terms of accuracy [15] and other aspects [18, 29]. In addition to segmentation, there is a promising potential in open-set object detection models, such as DINO [3], DINOv2 [12, 25], and Grounding-DINO [21]. Recently, Grounded-SAM [29] integrates RAM, Grounding-DINO, and SAM to establish a comprehensive pipeline for open-set semantic segmentation without the need for any training. However, one of the main limitations of this model is that the output from open-set detection models is not well-organized and can sometimes even detract from the intended objective. To overcome this problem, we leverage a GPT-4 model to effectively bridge the gap between open-set and closed-set tasks. This enables our proposed framework to sustain high performance while becoming more manageable, customizable, and versatile across different scenarios. An additional improvement involves utilizing the points prompt alongside the box prompt for better segmentation performance.

3 Proposed Framework

3.1 Framework Architecture

Our framework consists of three sub-components: a) Image recognition models, which include a Recognize Anything Plus Model (RAM++) [14] and a BLIP-2 [20] model. These models process input images to generate a list of tags and a caption that reflects the textual information present in the input image. b) An advanced GPT-4 [24] model is employed for a pre-refinement process, which process the textual information into predicted classes for the segmentation model. The system automatically generates prompts based on predefined target classes specific to different datasets. c) a pre-trained open-set segmentation model Grounded-SAM [29] that is able to detect and segment certain classes in the images based on predicted classes. A post-refinement process is applied during the detection phase using the same GPT-4 model. An overview of the model is shown in Fig. 1.

3.2 Image Recognition Models

We utilize two image recognition models in our framework. The primary model is RAM++ [14], an enhanced version of the original RAM [38]. Unlike typical image recognition models that generate descriptive caption text for images, RAM++ produces a set of tags containing elements within the image. These tags are more advantageous for semantic segmentation tasks compared to traditional caption text. Moreover, we believe that contextual information plays a significant role in aiding Large Language Models (LLMs) to better understand the context of an image. Thus, we adopt an additional image caption model, BLIP-2, to provide context for the image. The effectiveness of the additional BLIP-2 model is verified through an ablation study discussed in the following section.

3.3 Refinement using LLM

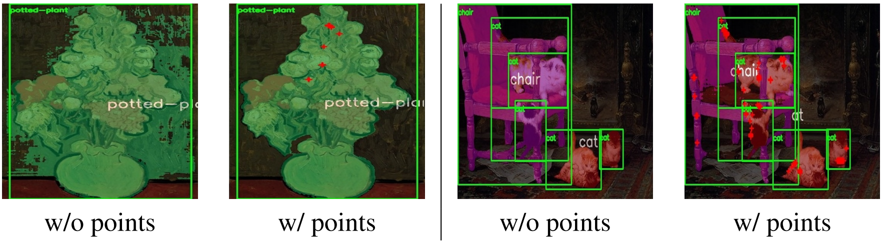
We utilize an advanced Large Language Model (LLM), GPT-4 [24], to perform two refinement processes to enhance the overall accuracy.

3.3.1 Pre-Refinement

A pre-refinement process is conducted to convert the textual information generated by image recognition models into specific classes that will be inputted into the segmentation model. This is essential to bridge the gap between target classes and detected text. In open-set models such as RAM++, BLIP2, and Grounding-DINO, outputs might include labels "girl, woman, man, boy, body, human, person...". However, in a closed-set task where the only required category is "person", other labels may compromise overall accuracy and cannot be utilized in the subsequent segmentation model. However, with the help of our pre-refinement process, all similar labels can be consolidated into a single target label, effectively bridging the gap between open-set and closed-set scenarios. Our system prompt for GPT-4 during the process is automatically generated based on the following template:

> Task Description:
> - You will receive a list of caption tags accompanied by a caption text and must assign appropriate labels from a predefined label list: L.
>
> Instructions:
> Step 1. Visualize the scene suggested by the input caption tags and text.
> Step 2. Analyze each term within the overall scene to predict relevant labels from the predefined list, ensuring no term is overlooked.
> Step 3. Now forget the input list and focus on the scene as a whole, expanding upon the labels to include any contextually relevant labels that complete the scene or setting.
> Step 4. Compile all identified labels into a comma-separated list, adhering strictly to the specified format.
>
> Contextually Relevant Tips:
> - Equivalencies include converting "girl, man" to "person" and "flower, vase" to "potted plant", while "bicycle, motorcycle" suggest "rider".
> - An outdoor scene may include labels like "sky", "tree", "clouds", "terrain".
> - An urban scene may imply "bus", "bicycle", "road", "sidewalk", "building", "pole", "traffic-light", "traffic-sign".
>
> Output:
> - Do not output any explanations other than the final label list.
> - The final output should strictly adhere to the specified format: label1, label2, ... labeln

Notably, L indicates a predefined list of semantic classes. For instance, in the context of the DRAM dataset, this list includes: *"background, bird, boat, bottle, cat, chair, cow, dog, horse, person, potted-plant, sheep"*. Additional examples are optional and could be automatically generated if an annotated segmentation dataset is available. Overall accuracy will improve according to our experiments, Our experiments indicate that overall accuracy would improve with these additions; however, since this paper concentrates on a zero-shot scenario, further details are relegated to the supplementary materials.



w/o points w/ points w/o points w/ points

Figure 2: **Qualitative Comparison**. With and without point prompts.

3.3.2 Post-Refinement

One main limitation of the Grounded-SAM model is that the detection results from the Grounding-DINO component may be similar to but not necessarily identical to the desired classes. For instance, the model could identify an object as a "table" instead of a "dining table," even the target object is the "dining table." The process may enhance the overall accuracy in such a scenario. Our system prompt for GPT-4 during the post-refinement process is automatically generated based on the following template:

> Task Description:
> You will receive a specific phrase and must assign an appropriate label from the predefined label list: L.
> Please adhere to the following rules:
> - Select and return only one relevant label from the predefined label list that corresponds to the given phrase.
> - Do not include any additional information or context beyond the label itself.
> - Format is purely the label itself, without any additional punctuation or formatting.

Similar to the system prompt of the pre-refinement process, L indicates a predefined list of semantic classes.

3.4 Segmentation

Based on the outputs from Grounded-SAM, we observed that sometimes, the bounding box encompasses a larger area than the actual object. Considering that the original SAM model fundamentally supports multiple prompt inputs, we propose not only using bounding box prompts in Grounded-SAM but also incorporating both boxes and points as prompts for segmentation. Examples are shown in the Figure 2.

These points are extracted as the top-n probability points from the detection results of the Grounding-DINO model. We also explore how varying the number of points affects overall accuracy. Our experiments suggest that using $n_p = 20$ points as prompts may be optimal. For more details, Please check the supplementary material. The ablation study in the following section will further demonstrate the effectiveness of point prompts in the segmentation process.

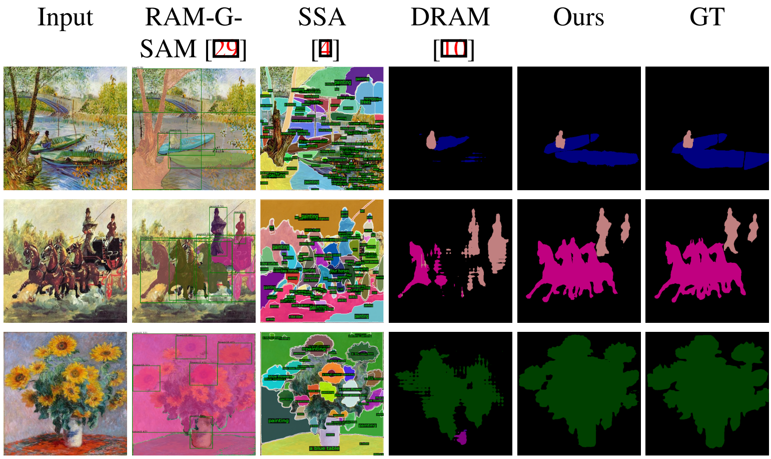


Figure 3: **Qualitative comparison**. We compare our approach against existing approaches.

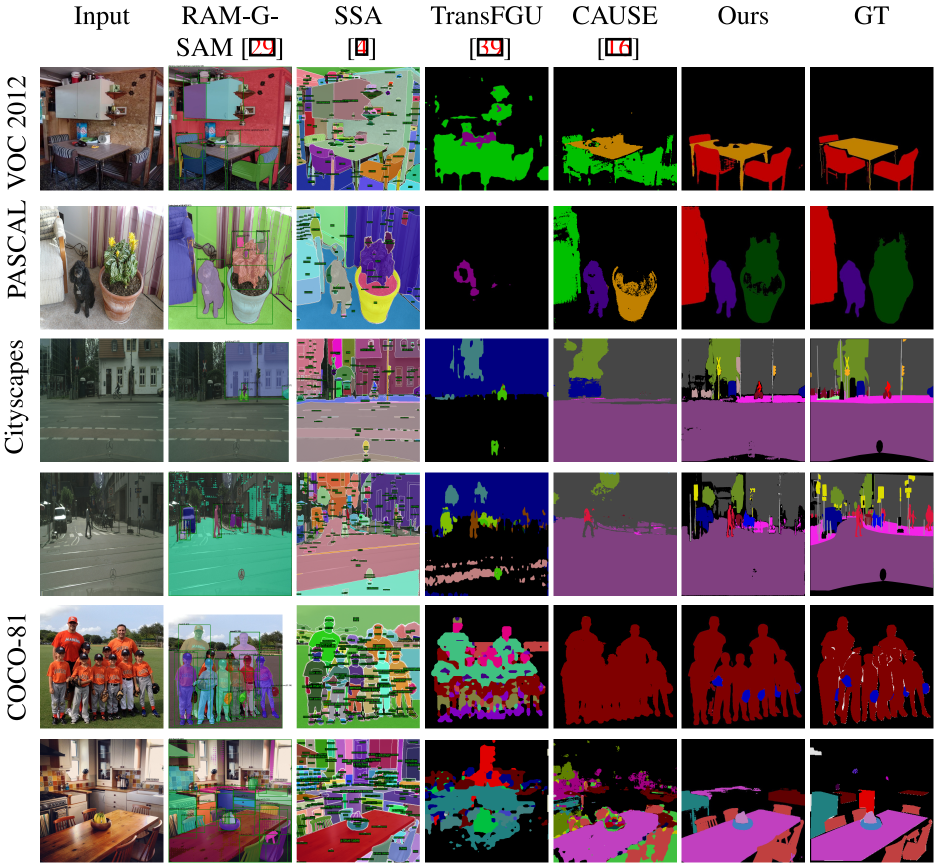


Figure 4: **Qualitative comparison**. Notably, on the Cityscapes dataset, the CAUSE model used a crop version from the original image since it only accepts square input.

Table 2: **Quantitative Comparison.** The upper three rows of the baseline models are trained under unsupervised learning, while both RAM-Grounded-SAM and our proposed model do not require any training.

Methods	DRAM	PASCAL VOC 2012	Cityscapes	COCO-81
Unsupervised Learning				
DRAM [10]	45.75	/	/	/
TransFGU [39]	/	37.15	16.83	12.69
CAUSE [16]	/	53.3	28.00	21.2
Training-free				
RAM-Grounded-SAM [29]	32.80	25.48	11.40	14.32
Ours	**62.01**	**63.57**	**34.36**	**37.45**

Table 3: **Ablation Study.**

Methods	DRAM	PASCAL VOC 2012	Cityscapes	COCO-81
Our Full Model	**62.01**	**63.57**	**34.36**	**37.45**
w/o BLIP-2	61.30	60.42	32.11	36.47
w/o Pre-Refinement	37.88	52.64	29.35	15.15
w/o Post-Refinement	61.28	61.72	28.53	35.94
w/o Points Prompt	60.81	62.71	32.57	36.01
w/o all elements (RAM-Grounded-SAM [29])	32.80	25.48	11.40	14.32
GPT-4 → Llama-3-8b	55.09	51.35	23.63	33.40
Input GT Labels	70.29	67.62	41.06	45.66

4 Experiment

4.1 Datasets

The datasets used in our experiments are as follows:

- **DRAM [10]** (Diverse Realism in Art Movements) containing 11 classes across four art movements, including Realism, Impressionism, Post-Impressionism, and Expressionism with unseen artworks. It has 5677 images for training and 718 images for validation.
- **PASCAL VOC 2012 [13]** comprises 1,464 images for training, 1,449 images for validation, and includes 20 semantic classes.
- **Cityscapes [11]** consists of 30 semantic classes and 4,500 images for training and 500 for validation.
- **COCO-81** is a subset of COCO-Stuff [2] that comprises 80 object categories and one background category. It consists of 118,000 images for training and 5,000 images for validation.

4.2 Baselines

We select two leading models in the area of unsupervised semantic segmentation, TransFGU [39] and CAUSE [16] as our baseline models. Specifically, we employ the model proposed in the original paper of the DRAM [10] dataset as the baseline model for the DRAM dataset.

For training-free methods, we consider RAM-Grounded-SAM [29] and SSA-engine [4] as our baseline model, both are extended work from SAM. These powerful open-set methods enable the SAM model of semantic segmentation tasks. However, as discussed in Section 3.3.1. Open-set methods tend to output numerous similar labels rather than pinpointing the most relevant target label. This results in extremely lower accuracy in closed-set scenarios, limiting its practical usage. Specifically, RAM-Grounded-SAM lacks a closed-set option, while the closed-set method of SSA requires a pre-trained supervised segmentation model, which does not meet our training-free setting. Although direct quantitative comparisons with open-set methods are not provided due to the aforementioned constraints, we include the performance of RAM-Grounded-SAM in our ablation study for a comprehensive analysis.

4.3 Quantitative Comparison

We use Mean Intersection over Union (mIoU) as the metric to evaluate segmentation accuracy. As shown in the Table 2. Our proposed method outperforms all unsupervised methods without any training process or additional information about the datasets, where our only inputs are the image and target label list.

4.4 Visual Comparison

We provide a qualitative comparison of results in Fig. 3 and Fig. 4. The results demonstrate that our proposed model significantly outperforms the unsupervised baseline models without any training process. While the open-set baseline models appear to perform adequately in segmentation tasks at first glance, a closer examination reveals several issues we previously discussed. For instance, in the third row in Fig. 3 and the second row in Fig. 4, both the RAM-Grounded-SAM and SSA models excessively segment the plant, introducing non-essential labels and negatively impacting the accuracy of the target label. Specifically, the SSA model seems to overly detect labels across every scene.

4.5 Ablation Study

We additionally conduct an ablation study to verify the effectiveness of each element. The results are shown in Table 3, which indicates the performance declines upon removing each element from our complete model. Notably, the performance when removing our pre-refinement process significantly drops because of the gap between target labels and detected labels, as discussed in Section 3.3.1. When our full model excludes all adapted elements, it obviously results in lower performance. This is equivalent to RAM-Grounded-SAM. For reference, we also present results from replacing the GPT-4 model with the Llama-3-8b [22] model, one of the best open-source LLMs. Additionally, we provide results for the scenario where the ground truth labels are inputted into the segmentation model for references. It represents the theoretical peak performance achievable by our pre-refinement process. As shown in the table, our results closely match the outcomes from ground truth labels.

5 Conclusion

We present a novel framework for training-free zero-shot semantic segmentation tasks. By leveraging a series of pre-trained models along with additional LLM refinement processes, our proposed framework achieves comparable accuracy to existing methods. This is accomplished without the requirement for training and avoids heavy dependence on annotated semantic datasets of previous supervised and unsupervised learning approaches. Limitations of our proposed method include reduced performance with a large number of semantic classes and limited accuracy in background object segmentation.

References

[1] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE, 2018.

[3] 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 International Conference on Computer Vision (ICCV)*, 2021.

[4] Jiaqi Chen, Zeyu Yang, and Li Zhang. Semantic segment anything. https://github.com/fudan-zvg/Semantic-Segment-Anything, 2023.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. arxiv. *arXiv preprint arXiv:1706.05587*, 5, 2017.

[8] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4288–4298, 2022.

[9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[10] Nadav Cohen, Yael Newman, and Ariel Shamir. Semantic segmentation in art paintings. *Computer Graphics Forum*, 41(2):261–275, 2022. doi: 10.1111/cgf.14473.

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2): 303–338, June 2010.

[14] Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. Open-set image tagging with multi-grained text supervision, 2023.

[15] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023.

[16] Junho Kim, Byung-Kwan Lee, and Yong Man Ro. Causal unsupervised semantic segmentation. *arXiv preprint arXiv:2310.07379*, 2023.

[17] 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.

[18] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023.

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[21] 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.

[22] Meta. Introducing meta llama 3. 2024. URL https://ai.meta.com/blog/meta-llama-3/.

[23] OpenAI. Introducing chatgpt. 2023. URL https://openai.com/index/chatgpt/.

[24] OpenAI. Gpt-4 technical report, 2024.

[25] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[26] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[27] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[28] 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.

[29] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[31] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021.

[32] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[33] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1130–1140, 2023.

[34] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2314–2320, 2016.

[35] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2314–2320, 2017. doi: 10.1109/TPAMI.2016.2636150.

[36] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[37] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36, 2024.

[38] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, Yandong Guo, and Lei Zhang. Recognize anything: A strong image tagging model, 2023.

[39] Yin Zhaoyun, Wang Pichao, Wang Fan, Xu Xianzhe, Zhang Hanling, Li Hao, and Jin Rong. Transfgu: A top-down approach to fine-grained unsupervised semantic segmentation. In *European Conference on Computer Vision*, pages 73–89. Springer, 2022.

[40] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023.

[41] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Curran Associates Inc., 2024.

[42] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024.