IOSD: Improved Open-vocabulary Segmentation with Diffusion Models

Xiang Wang
2100013146@stu.pku.edu.cn

Wei Zhang
2100013122@stu.pku.edu.cn

School of EECS, Peking University

Figure 1. **Mask and image predictions from IOSD**. We list some uncommon object categories, such as **Eiffel tower, red envelope and Lion dance**, and generate corresponding images and masks to assess and validate the model's quality.

Abstract

Recent efforts to exploit correlations between text and image using diffusion models have seen significant progress. In this paper, we propose ***Improved Open-vocabulary Segmentation(IOSD)*** *based on Grounded Diffusion, showing the interpretative capacity of latent image space for specific masks. Equipped with latest version of stable diffusion, we train the model on PASCAL and COCO classes, employing various splits of seen labels and showcasing its generalization capabilities to open-vocabulary prompts. Furthermore, we evaluate our generated masks with MMDetection, demonstrating its novelty in the realm of synthetic semantic image segmentation.*

1. Introduction

Segmentation of diverse images has long been a captivating field, garnering significant attention in academic research. For instance, Mask R-CNN[20] leverages a conventional CNN network for classification and introduces a dedicated mask generation network. Also recently, SAM[29] introduced a promptable segmentation task, where the objective is to generate a valid segmentation mask based on any segmentation prompt. While these models exhibit remarkable proficiency in segmenting everyday objects such as dogs and cats, they often struggle with rare or unrealistic objects. To elaborate, many segmentation models are trained on limited datasets like COCO and PASCAL, resulting in a natural bias towards common objects that are extensively featured during the training process as seen objects. There is an evident need for datasets encompassing

a broader array of categories to enhance the discriminative capabilities of segmentation and detection models.

On the other hand, text-to-image generative models have already achieved a robust correspondence between visual pixels and language, learned from extensive corpus of image-caption pairs. Building upon Latent Diffusion Models, Stable Diffusion[1, 18] introduces cross-attention layers into the model architecture, employs a stable training procedure, and establishes a correspondence between open-vocabulary text and image. With such correspondence, we are able to extract alignment capabilities between open-vocabulary text and images, enabling application to downstream tasks such as segmentation and detection—precisely what we aim to achieve.

To accomplish our objective, we draw inspiration from Grounded Diffusion[32] and enhance its segmentation module to align the visual characteristics of the latent space with a given prompt. To be specific, the original model design solely incorporated features from the last few Unet layers, derived from Stable Diffusion, and input them collectively into the fusion module. This design has two issues: (i) The transformer block takes image features all at once and input them into the model without considering changes during the image decoding process. (ii) It only extracts features from few final layers, failing to fully exploit the image representation capabilities of the diffusion model. Thus, we decide to make significant modifications by refining and reconstructing the visual encoder. Our aim is to extract latent features from each layer and input them sequentially into the model in the denoising order-from noise to the real image, anticipating a closer approximation to synthesis process. These representations now function as the key and value matrices within the fusion module, where we leverage a transformer architecture to continuously refining and fixing mask generation by recognizing differences and changes between these features, as shown in Fig.2.

During the assessment of our model's performance, we conduct a comprehensive comparison of segmentation results using the PASCAL VOC and COCO datasets against state-of-the-art models DAAM[52] and Grounded Diffusion[32]. Our model consistently outperforms these benchmarks across both seen and unseen classes. Additionally, we explore various training strategies, including altering the fixed prompt format and filtering ground truth masks with low confidence scores, all of which contribute to further enhancements in performance.

In summary, our contributions can be outlined as follows: (i) We introduce a novel model for the segmentation of synthetic images, proposing a paradigm for synthetic image segmentation and dataset construction. This framework facilitates research and evaluation in the field. (ii) We enhance the segmentation module to enable effective text-image alignment, allowing for the extraction of open-vocabulary text-to-image generability from Stable Diffusion. This enhancement paves the way for exploring applications in downstream tasks. (iii) We delve into the feasibility of knowledge transfer from large models and model compression. Our findings indicate that small models, specifically in our work referring to the segment module and fusion module, have the capability to acquire knowledge from larger models like Stable Diffusion. This highlights the potential for accelerating the inference process and reducing GPU memory usage.

2. Related Work

Diffusion Models Diffusion model is a parameterized Markov chain that optimizes the lower variational bound on the likelihood function to generate samples matching real distribution. Ho et al. [22] first propose diffusion model and Dhariwal and Nichol[10] further show the potential of diffusion models, achieving better image sample quality compared with other generative models, i.e., GAN, on ImageNet dataset [9]. After that, denoising diffusion probabilistic models have achieved remarkable success on various tasks, and more and more researchers have turned their attention to diffusion models [1, 18, 21, 23, 26, 27, 31, 37–39, 47, 49, 51]. The unprecedented capabilities of generative AI have been unlocked by foundation models for visual computing, such as Stable Diffusion [2, 45, 50], Imagen [48], Midjourney, or DALL-E 2 [42] and DALL-E. In this paper, we choose stable diffusion XL-Turbo[50] as our frozen diffusion model.

Detection and Segmentation Object detection refers to the computer vision task of identifying and locating objects within an image or video frame. It has embraced methodologies of distinct properties—e.g., two-stage [16, 17, 19, 44] vs. one-stage [33, 34, 43], anchor-based [44] vs. anchor-free [12, 30, 53], and region-based [16, 17, 19, 44] vs. query-based (DETR) [3]. Segmentation involves partitioning images or video frames into multiple segments and objects. It has developed methodologies including CNN based models [6, 8, 40, 46], R-CNN based models [16, 20, 44], Attention-based models [7, 14, 24, 55, 57], and large models like SAM [29].

Transformers Since the Transformer[54] was proposed by Vaswani et al., Transformer has been widely used in various fields and has replaced domain-specific architectures across language[13, 15], vision [11, 35, 36, 56], and reinforcement learning [5, 25]. They have shown remarkable performance in almost all tasks including generation, regression, classification, detection, and so on. In this paper, we study the ability of transformers to fusion multimodal information, when used as the backbone of a fusion mod-

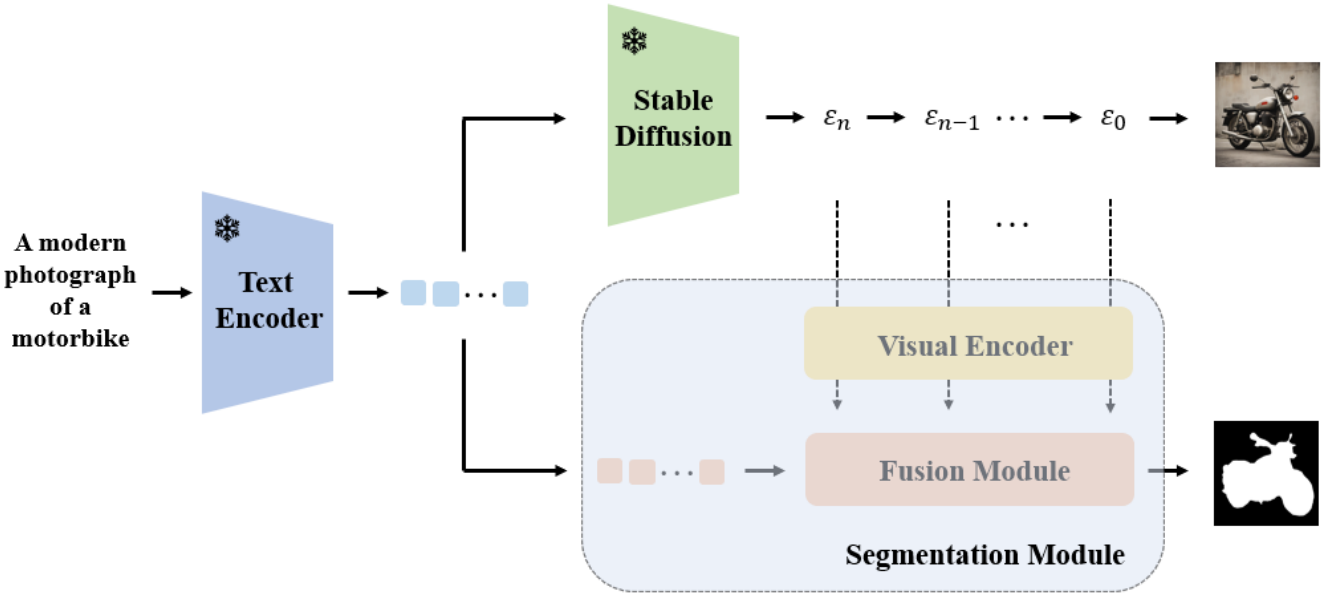

Figure 2. **Overview of IOSD framework.** We have two pipelines: one for image generation and another for mask generation. The output from Diffusion UNet layer is passed through the visual encoder, generating keys and values for fusion module. The query for each fusion layer is obtained from the text-embedding vector. We extract class-associated tensors and exclude descriptive-text tensors to obtain class embedding.

ule to predict the segmentation of query target. Attention-based models, Encoder-decoder based models, Multiscale and pyramid network based models, R-CNN based models (for instance segmentation)

Grounded Diffusion[32] Grounded Diffusion serves as our foundational work, laying the groundwork for our research endeavor. It introduces a groundbreaking concept, steering the well-established Stable Diffusion towards grounded generation. This involves the segmentation of visual entities explicitly described in a given text prompt while simultaneously generating an image. The model employs a visual encoder to concatenate latent features sharing same spatial resolutions. Additionally, it leverages Transformers[54] as a fusion module, providing valuable insights that inspire our own investigation.

3. Method

Model Structure Overview As Fig. 2 demonstrated, our model consists of three parts, pre-trained frozen text encoder ϕ_{clip}, pre-trained frozen diffusion model ϕ_{sdxl}, and segmentation model ϕ_{seg}. The pipeline can be formulated as

$$(\mathcal{I}, \mathcal{F}) = \phi_{sdxl}(\phi_{clip}(\mathcal{X})) \quad (1)$$

$$\mathcal{M} = \phi_{seg}(\phi_{clip}(\mathcal{Q}), \mathcal{F}) \quad (2)$$

where $\mathcal{X}$ represents prompt for diffusion to generate image $\mathcal{I}$, $\mathcal{F}$ represents intermediate features from diffusion model, and $\mathcal{M}$ represents predicted binary mask corresponding to query $\mathcal{Q}$ on generated image.

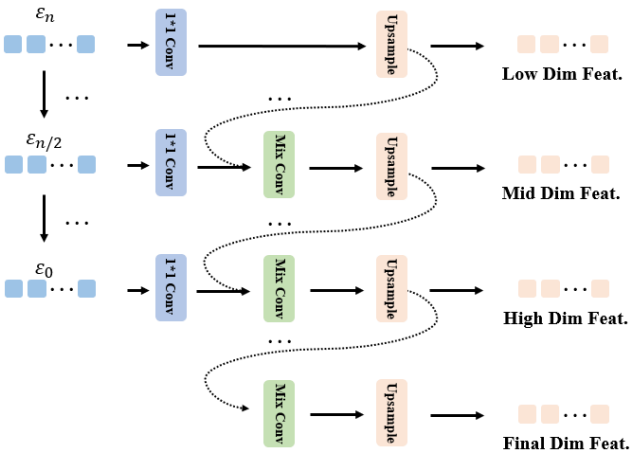

Figure 3. **The architecture of visual encoder.** With features at different resolutions, we pass them through convolution layers and upsample layers. These features then serve as key and value states, guiding the fusion module to focus on the differences in each resolution's latent space and generate masks accordingly.

3.1. Segmentation Module

Segmentation module contains two main components, visual encoder and fusion module, which merge to complete text and image alignment. In the visual encoder, we collect

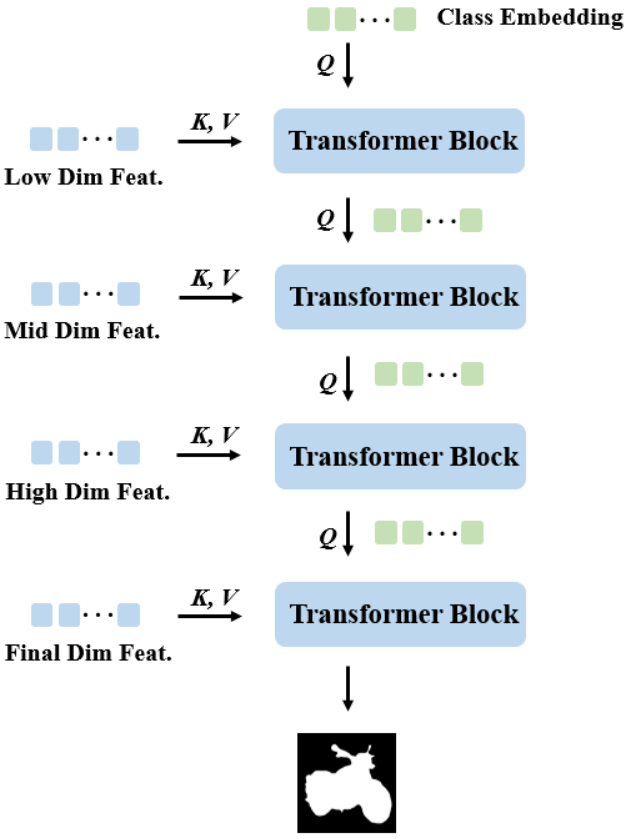

Figure 4. **The architecture of fusion module.** The features obtained from the visual encoder are used as key and value inputs to the transformer block. We also utilize the text prompt to obtain class embedding as queries, continuously updating and refining the mask generation.

the output of each layer in the Unet structure of the Diffusion module and categorize these features based on their size as shown in Fig.3. In contrast to Grounded Diffusion, we have opted not to concatenate and mix features of different sizes before passing them through the encoder. Instead, we choose to process these sizes separately. The features of each size are combined with the output of the previous size, and after passing through their own Convolutional and Upsampling layers, we obtain features for the particular size. Ultimately, we obtain features of four different sizes: low dimension, medium dimension, high dimension, and final dimension.

Once we obtain features of different dimensions, we feed these features into the fusion module sequentially. We employ a transformer architecture as the foundation of the fusion module, utilizing these features as key and value, with the output of the previous layer serving as the query (the initial layer's key comes from the class embedding vector), as shown in FIg.4. The rationale behind this design is to enable the fusion module to progressively focus on the distinctions among these features of different sizes during iterations: From low to high dimensions, the features are approaching the real image step by step, enabling the network to attend to differences and correct mask errors in the previous layer's output.

3.2. Diffusion Models

The efficacy of generated imagery is intrinsically tied to diffusion models, which not only influence the performance of pre-trained detectors but also propagate errors into the segmentation module when generating images with low quality. The caliber of these images is pivotal in crafting high-quality synthetic datasets for subsequent tasks. To strike a balance between efficiency and output quality, we choose SDXL-Turb[50] from StabilityAI.

3.3. Filter

The paramount importance of superior training data in ensuring model stability and proficiency is well-established. Li et al. highlighted the detrimental impact of pre-trained detector malfunctions on segmentation modules, advocating for the exclusion of failure cases of the pre-trained detector. Our investigations reveal that merely filtering out failure cases of the pre-trained detector is inadequate. To safeguard ground truth mask integrity, it is crucial to also dismiss instances where the pre-trained detector exhibits low confidence. Consequently, we devised a refined filter to eliminate low-quality or mismatched masks.

3.4. Prompt Engineering

Prompt Engineering: The domain of prompt engineering has undergone extensive exploration, consistently demonstrating its efficacy in augmenting large diffusion model capabilities. Nevertheless, constructing voluminous prompts for each training class and selecting one at random as the input $\mathcal{X}$ destabilizes initial segment module training, attributable to the excessive complexity for nascent models. Conversely, employing a rudimentary template for prompt generation risks segment module overfitting. To mitigate these issues, we introduce a balanced approach by incorporating a negative class into a fixed prompt structure. Our template, `"A cinematic shot of class1 and class2, portrayed naturally and realistically."` integrates a randomly selected class from the training class split as either class1 or class2 with a 50% probability, while the other is the query target. The 'negative class' is distinct from the 'query class', and notably, these classes are not observed concurrently in real-world scenarios which will enhance the generalization ability. This structured template ensures consistent training progression, while the introduction of a randomly selected negative class bolsters the model's performance in zero-shot open-vocabulary segmentation tasks.

Algorithm 1 Training For Segment Model ϕ_{seg}

Given class split set $\mathcal{S}$
Initialize ϕ_{seg} randomly.
Apply Adam[28]:
for each iteration **do**
    Uniformly sample i from $\mathcal{S}$ as query $\mathcal{Q}$
    Construct prompt $\mathcal{X}$ with a fixed template.
    Apply CLIP[41] ϕ_{clip} to get prompt condition $\hat{\mathcal{X}}$ and query embedding $\hat{\mathcal{Q}}$.
    Apply diffusion model ϕ_{sdxl} to get generated image $\mathcal{I}$ and intermediate features $\mathcal{F}$.
    Apply ϕ_{mmdet} to get mask as the ground truth $\hat{\mathcal{M}}$
    Apply filter ϕ_{filter} to decide abandon $\hat{\mathcal{M}}$ or not.
    if abandoned **then** continue.
    end if
    Apply a segmentation module ϕ_{seg} to get prediction mask $\mathcal{M}$.
    Calculate loss $\mathcal{L}$ donated by BCE loss between $\mathcal{M}$ and $\hat{\mathcal{M}}$
end for
return ϕ_{seg}

4. Experiment

4.1. Class Split

Referring to Grounded Diffusion[32] original paper, we found that the class 'bench' appears in both seen and unseen in class-split3 of COCO, and it appears twice in class-split1 or class-split2, which is out of our expectation. We manually deleted this class once in all class-splits to correct this mistake. For more details, you can refer to the appendix of the original paper.

4.2. Evaluation

We trained our model independently in all class splits of PASCAL and COCO with Algorithm 1. See Sec. 4.1 for more information of class splits. According to the number of classes, we set different iterations to make sure training converges. We trained 3000 iterations on 1 x RTX4090 for every class split on PASCAL, while we trained 12000 iterations on 1 x RTX4090 for every class split on COCO. After training, we evaluate models with mIoU score on every class split. For each class, we generate 50 images and apply mmdet[4] to get the segmentation mask as ground truth. The results are listed in Tab. 1. Some visualizations of open vocabulary task are listed in Fig. 5.

Following Li et al., we take DDAM[52] as a baseline, while the other is G.D.[32]. While G.D. defeats DAAM, our model defeats G.D. in every class split on both PASCAL and COCO. Additionally, our model outperforms G.D. by a significant margin on COCO. The capability of our model is not fully demonstrated because PASCAL only contains twenty simple categories, for which G.D. is powerful enough.

Table 1. **Quantitative results**. We evaluate models with mIoU score on three class splits, for each class, we generate 50 images and apply mmdet[4] to get segmentation mask as ground truth. Our setting of training is DSF, see Tab. 2 for details.

model	class split	PASCAL		COCO	
		seen↑	unseen↑	seen↑	unseen↑
DAAM[52]	Split1	61.66	75.63	62.25	55.56
	Split2	65.75	59.25	60.08	65.55
	Split3	67.11	53.82	62.81	52.48
	Average	64.84	62.90	61.71	57.76
GD[32]	Split1	90.16	83.19	83.35	76.81
	Split2	90.08	86.19	82.83	74.93
	Split3	90.67	79.86	84.85	67.89
	Average	90.30	83.08	83.68	73.21
Ours	Split1	92.54	84.76	90.79	88.04
	Split2	92.18	88.18	90.81	79.34
	Split3	93.29	80.71	91.49	83.03
	Average	**92.67**	**84.55**	**91.03**	**83.47**

4.3. Ablation Study

We conduct an ablation study on PASCAL split1 to justify key design choices. For convenience, we simplify our designs as D, S, F, and P, which corresponds in turn to Sec. 3.2, Sec. 3.1, Sec. 3.3, and Sec. 3.4.

Diffusion Model As we expected, a more powerful diffusion model leads to a more powerful segment module. The two key factors are image quality, which is naive, and prompt alignment. The segment module takes in both text embedding and intermediate features from the diffusion model. The more aligned between generated image and the prompt, the smaller the gap between text embedding and intermediate features, and, consequently, the more accurate the segment module will be.

Segment Module Based on replacing the pre-trained diffusion model, optimizing the style module brings improvement. Comparing D & DS and DF & DSF, we can observe steady improvement. On the premise that the mIoU score reaches 90, it is very difficult to improve, because the model is required to be able to segment the details of the object with extreme accuracy, not just the outline. It verified that our pyramid-based design is really helpful to segment model to segment from outline to details step by step and get higher accuracy.

Filter As shown in Tab. 2, adopting a filter will bring great improvement on seen classes, and the filter speeds up the

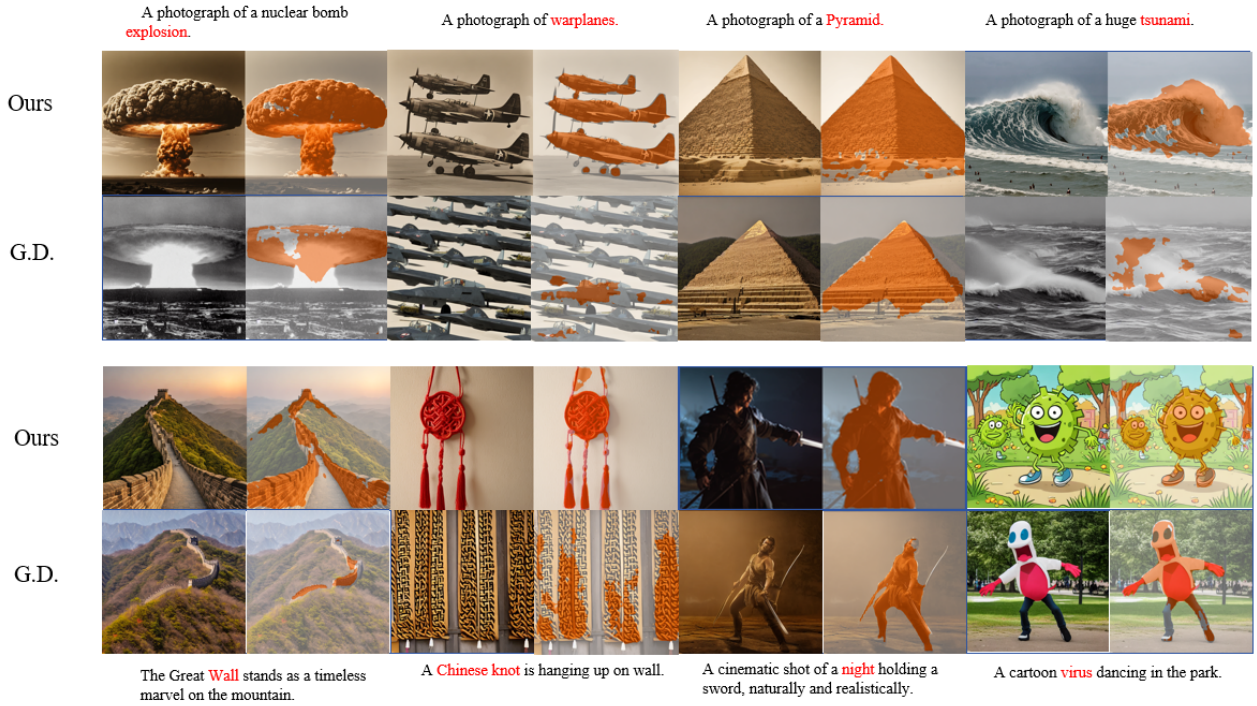

Figure 5. **Compare with G.D.** Our model gets better performance in open-vocabulary segmentation tasks.

convergence. This phenomenon verified that training data with higher quality will contribute to steadier and faster training as well as more powerful models. which is worth noticing, the performance drops slightly when we adopt a filter. We believe that the cause is the hyperparameters setting of the filter is too high so that the quality of training data is too high to generalize to harder cases.

Prompt Engneering We expect prompt engineering to improve the capability of open-vocabulary segmentation. However, the experiment indicates that it not only slows down the coverage speed but also leads to a significant decrease in mIoU score. But when we inference using the model trained in the DSFP setting, we find better performance in the open-vocabulary segmentation task. It inspires us that the mIoU score on normal objects is not accurate enough to evaluate the ability to open-vocabulary segmentation tasks.

Table 2. **Ablation Study**. We conduct an ablation study on PASCAL split1. D: adopt up-to-date diffusion model SDXL-Turbo[50] from StablityAI. S: redesign segment module. F: adopt filter to remove error from mmdet[4]. P: Prompt Engineering, build prompt with both query class and a negative class which is different from query class. imp: improvement comparing G.D.[32]. iter: converge speed, how many iterations are needed to get converge.

Setting	mIoU↑		imp.↑		iter.↓
	seen	unseen	seen	unseen	
G.D.[32]	90.16	83.19	0	0	∼ 5000
D	91.35	84.34	1.19	1.15	∼ 5000
DS	92.09	**84.85**	1.97	**1.66**	∼ 5000
DF	92.20	84.22	2.04	1.03	∼ **3000**
DSF	**92.54**	84.76	**2.38**	1.57	∼ **3000**
DSFP	81.17	73.85	-8.99	-9.34	∼ 10000

5. Conclusion

In this course project, we choose to enhance the work of Grounded Diffusion and proposed the IOSD model for the detection of open-vocabulary objects. We introduces our own innovative ideas and improvements for the segmentation module in Grounded Diffusion, which achieves significant results. Additionally, throughout the experiments, we identify shortcomings in the text prompts processing and ground truth selection of the original work. To address these issues, we refined the training process by employing prompt engineering and filtering masks based on confidence scores. With such extensive experimentation, the robustness of our proposed methods has been established and guaranteed.

5.1. limitation

Regarding the shortcomings of this work, the primary concern lies in the limitations of knowledge distillation performance. In our experiments, IOSD demonstrates impressive results in single-object segmentation tasks. However, when it comes to multi-object segmentation (e.g., simultaneous segmentation of cats and dogs in single image), the model's discriminative ability is less satisfactory. We attribute this limitation to insufficient training resources, preventing the model from fully acquiring the comprehensive knowledge embedded in stable diffusion. Additionally, due to time constraints, we had to forgo attempts to construct a complete dataset, although we believe this would be a meaningful endeavor.

Nevertheless, we remain optimistic that addressing these challenges is feasible with an expanded dataset and increased computational resources. We propose that further enhancement could effectively overcome these limitations in the near future.

5.2. future work

As we proposed in our presentation, We believe that this work can also be extended to the domain of video segmentation and detection—simply by adapting stable diffusion to the task of generating videos based on textual prompts. Fortunately, Stable Video Diffusion has already achieved this capability. We look forward to the potential applications of this work and have intentions to explore segmentation and detection in the context of video generation in the future.

References

[1] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34: 17981–17993, 2021. 2

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5, 6

[5] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34: 15084–15097, 2021. 2

[6] 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. 2

[7] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016. 2

[8] 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. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[12] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019. 2

[13] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020. 2

[14] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention net-

work for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019. 2

[15] Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183. IEEE, 2020. 2

[16] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[17] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2

[18] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9): 1904–1916, 2015. 2

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[23] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23 (1):2249–2281, 2022. 2

[24] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019. 2

[25] Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34:1273–1286, 2021. 2

[26] Gwanghyun Kim and Jong Chul Ye. Diffusionclip: Text-guided image manipulation using diffusion models. 2021. 2

[27] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 2

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 2

[30] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. 2

[31] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 2

[32] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7667–7676, 2023. 2, 3, 4, 5, 6

[33] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[34] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016. 2

[35] 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. 2

[36] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 2

[37] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Re-

paint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 2

[38] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[39] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2

[40] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 2

[41] 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. 5

[42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 2

[43] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[46] 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. 2

[47] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[49] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 2

[50] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023. 2, 4, 6

[51] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[52] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*, 2022. 2, 5

[53] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 2

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[55] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1857–1866, 2018. 2

[56] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022. 2

[57] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 267–283, 2018. 2