Supplementary to Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models

Jiarui Xu[1*] Sifei Liu[2†] Arash Vahdat[2†] Wonmin Byeon[2]
Xiaolong Wang[1] Shalini De Mello[2]
[1]UC San Diego [2]NVIDIA

In this supplement we provide additional implementation details; quantitative experimental and qualitative visual results.

1. Implementation Details

We open-source our code and models at https://github.com/NVlabs/ODISE.

Training We train ODISE for 90k iterations with images of size 1024^2 and use large scale jittering [7] with random scales between $[0.1-2.0]$ as data augmentation. We use 32 NVIDIA V100 GPUs with 2 images per GPU with an effective batch size is 64. We use the AdamW [14] optimizer with a learning rate 0.0001 and a weight decay of 0.05. We use a step learning rate schedule and reduce the learning rate by a factor of 10 at 81k and 86k iterations. We set the balancing factor between the diffusion and discriminative models to $\lambda = 0.65$ for all tasks. Following [2–4], we use Hungarian matching to match the predicted masks to the ground-truth ones. We compute the training losses between the matched pairs.

Open-Vocabulary Inference An object can often be described by more than one possible description, e.g., the dog category could be described by "dog" or "puppy". We use the same prompt engineering strategy as in [8] to create an ensemble of text prompts for each test category and predict the category with the maximum probability.

Speed and Model Size It takes 5.3 days to train ODISE for 90k iterations on the COCO dataset. It has 28.1M trainable parameters (only 1.8% of the full model) and 1,493.8M frozen parameters (including Stable Diffusion and CLIP). It performs single image inference at 1.26 FPS on an NVIDIA V100 GPU and uses 11.9 GB memory with an image of size 1024^2. We also replace the bounding box cropping proposed in [10] that runs at 0.38 FPS, with mask feature pooling described in Section 3.6 of the main paper. Mask pooling yields a 3x speedup, while maintaining similar PQ on ADE20K: 23.4 for mask pooling versus 23.7 for bounding box cropping.

[*]Jiarui Xu was an intern at NVIDIA during the project. † equal contribution.

2. Experiments

2.1. Comparison with State of the Art

Open-Vocabulary Panoptic Segmentation Besides panoptic quality (PQ), we additionally report the detailed metrics of segmentation quality (SQ) and recognition quality (RQ) for ODISE and MaskCLIP [6] on both the thing (Th) and stuff (St) categories of the ADE20K dataset in Table 2.1. Here, all models were trained on COCO. ODISE outperforms MaskCLIP [6] w.r.t. all metrics.

Method	PQ	PQ^{Th}	PQ^{St}	SQ	SQ^{Th}	SQ^{St}	RQ	RQ^{Th}	RQ^{St}
MaskCLIP	15.1	13.5	18.3	70.5	70.0	71.4	19.2	17.5	22.7
ODISE (Ours)	**23.4**	**21.9**	**26.6**	**78.1**	**77.7**	**78.8**	**28.3**	**26.6**	**31.6**

Table 2.1. **Detailed panoptic segmentation metrics on ADE20K.** ODISE outperforms MaskCLIP [6] w.r.t. all metrics.

	Cityscapes			Mapillary Vistas		
Method	PQ	SQ	RQ	PQ	SQ	RQ
CLIP(H)	18.5	69.4	24.2	11.7	60.5	15.1
ODISE (Ours)	**23.9**	**75.3**	**29.0**	**14.2**	**61.0**	**17.2**

Table 2.2. **Results of panontic segmentation on Cityscapes and Mapillary Vistas.** ODISE outperforms CLIP(H) by a large margin on both datasets.

We also evaluate ODISE trained on COCO on the Cityscapes [5] and Mapillary Vistas [15] datasets in Table 2.2. Since the source code for MaskCLIP [6] is not publicly available, we regard ODISE's implementation with CLIP(H) features (from Table 3 of the main paper) as a close proxy to MaskCLIP and compare against it (Table 2.2). Here too, ODISE, which is based on diffusion features, outperforms its CLIP(H) variants by large margins. Note that in this experiment, we use the original text labels provided with the respective test datasets and didn't carefully select the category names for computing the text embedding. Hence, the results could be further improved

if categories like "terrain" are converted into more detailed descriptions.

Finally, to additionally verify the effectiveness of ODISE, we also swap the training and evaluation datasets, i.e., we train on ADE20K and evaluated on COCO, and report the results in Table 2.3. Here too, we regard the variant of ODISE with CLIP(H) features as a proxy to MaskCLIP [6] and compare against it. ODISE outperforms its CLIP(H) variant by a large margin.

	COCO			ADE20K		
Method	PQ	SQ	RQ	PQ	SQ	RQ
CLIP(H)	20.7	72.6	26.5	25.7	72.3	32.1
ODISE (Ours)	**25.0**	**79.4**	**30.4**	**31.4**	**77.9**	**36.9**

Table 2.3. **Results of swapped training on ADE20K and testing on COCO.** ODISE outperforms CLIP(H) by a large margin on both datasets.

Open-Vocabulary Object Detection We also evaluate ODISE for the task of open-vocabulary object detection on the LVIS [11] dataset (Table 2.4). By regarding all categories to belong to "things", we directly evaluate on LVIS's object detection labels, which contain annotations for 1203 fine-grained categories for COCO [13] images. For this task, we measure mAP_r, which denotes the mAP score on 337 rare categories only. We evaluate ODISE trained with both types of supervision: mask category labels or image captions. ODISE outperforms MaskCLIP [6] by a large margin w.r.t. both mAP and mAP_r. Note that the validation split of LVIS [11] has overlapping images with COCO [13]'s training split, but the category labels of LVIS are only available during inference.

	Supervision			LVIS	
Method	label	mask	caption	mAP	mAP_r
MaskCLIP [6]	✓	✓		8.4	-
ODISE (Ours)	✓	✓		15.4	19.4
ODISE (Ours)		✓	✓	**17.1**	**21.1**

Table 2.4. **Open-Vocabulary Object Detection.** mAP_r denotes the mAP score for 337 rare categories only. ODISE surpasses MaskCLIP by a large margin, both with category label and caption during training.

Open-World Instance Segmentation The task of open-world instance segmentation aims at discovering at test time, all plausible instance masks that may be present in an image in a class-agnostic manner. We also evaluate ODISE in for this task. Following [17], we report the average recall of 100 mask proposals (AR@100) on the UVO [18] and ADE20K [19] datasets. As reported in Table 2.5, here too we outperform the existing state of the art [17] by 14.3 points on UVO and 9.3 points on ADE20K. It demonstrates that with the internal representation of pre-trained text-to-image diffusion models it is plausible to discover open-world instances.

	AR@100		
Method	UVO	ADE20K	COCO
LDET [12]	42.6	-	-
GGN [17]	43.4	21.0	-
ODISE (Ours)	**57.7**	**30.3**	**56.6**

Table 2.5. **Open-world Instance Segmentation.** ODISE outperforms GGN on discovering open-world instances on both the UVO and ADE20K datasets.

2.2. Ablation Study

Visual Representations In Fig. 2.1 we show k-means clustering of the text-to-image diffusion model's and CLIP's frozen internal features; diffusion features are much more semantically differentiated. Quantitative comparisons of ODISE and its CLIP(H) variant in Table 3 of the main paper and Table 2.2 and Table 2.3 further substantiate diffusion features' superiority over those of CLIP's.

Figure 2.1. **K-mean clustering of text-to-image diffusion and CLIP models' internal features**. The internal features of the diffusion model are much more semantically differentiated than those from CLIP.

Open-vocabulary Inference Pipelines For final open-vocabulary classifciation, we fuse class prediction from the diffusion and discriminative models. We report their individual performance in Table 2.6. Individually the diffusion approach performs better on both the ADE20K and COCO datasets than the discriminative only approach. Nevertheless, fusing both together results in higher accuracy on both datasets. Finally, note that even without fusion, our diffusion-only method already surpasses the existing MaskCLIP method (in Table 1 of the main paper).

Diffusion Time Steps We study, which diffusion step(s) are most effective for extracting features from, similarly to

model		ADE20K			COCO		
diffusion	discriminative	PQ	mAP	mIoU	PQ	mAP	mIoU
	✓	15.0	9.6	17.5	26.5	23.5	23.6
✓		20.1	10.3	24.4	42.3	37.8	52.0
✓	✓	**23.3**	**13.0**	**29.2**	**44.2**	**38.3**	**53.8**

Table 2.6. **Ablation results of fusing class predictions of diffusion and discriminative models.**

	ADE20K			COCO		
time step	PQ	mAP	mIoU	PQ	mAP	mIoU
0	**23.3**	**13.0**	29.2	**44.2**	**38.3**	**53.8**
100	22.8	12.5	29.3	43.2	36.4	52.3
200	21.5	11.9	28.0	42.4	35.1	51.7
500	20.9	11.1	27.0	38.2	31.1	47.6
0+100+200	23.1	12.9	**29.7**	43.7	37.4	53.0
learnable	22.8	12.9	29.2	44.0	37.5	53.4

Table 2.7. **Ablation results of different diffusion time steps.** 0+100+200 denotes the concatenation of the features at time steps 0, 100, and 200.

DDPMSeg [1]. The noise process is defined as

$$x_t \triangleq \sqrt{\bar{\alpha}_t}x + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{1}$$

The larger the t value is, the larger the noise distortion added to the input image is. In stable diffusion [16] there are a total of 1000 diffusion steps. From Table 2.7, we observe that all metrics decrease as t increases and the best results are for t=0 (our final value). Concatenating 3 time steps, 0, 100, 200, yields a similar accuracy to t=0 only, but is 3× slower in terms of inference time and hence is not preferred. We also train our model with t as a learnable parameter, and find that many random training runs all converge to a value close to zero, further validating our optimal choice of t=0. A plausible explanation for this observed phenomenon could be that the highest-quality features for the downstream task of segmentation may be derived from the least noisy input image at $t = 0$.

3. Qualitative Results

To demonstrate the open-vocabulary recognition capabilities of ODISE, we merge the category names from LVIS [11], COCO [13], ADE20K [19] together and perform open-vocabulary inference with $\sim$1.5k test classes. We only train ODISE on COCO's [13] training dataset and evaluate open-vocabulary panoptic inference on ADE20K [19] and Ego4D [9]. The qualitative results on COCO's [13] validation dataset, ADE20K [19] and Ego4D [9] are shown in Fig. 2.2, Fig. 2.3 and Fig. 2.4, respectively. Most categories, e.g., "police cruiser", "flag", "conveyor belt", "chandelier", "aquarium", "grocery bag", "power shovel", *etc.*, are novel categories from LVIS [11] or ADE20K [19] that are not annotated in COCO [13]. It is worth noting that Ego4D [9] is a video dataset, which consists of diverse ego-centric videos. Despite the large domain gap between the testing dataset Ego4D [9] and our training dataset COCO [13], ODISE still outputs good-quality plausible panoptic segmentation results on Ego4D's novel categories.

4. Limitations and Future Work

In the current datasets, the category definitions are sometimes ambiguous and non-exclusive, e.g., in ADE20K, "tower" is often mis-classified as "building". Although this could be mitigated by prompt and ensemble engineering, how category definitions affect evaluation accuracy, would be interesting to analyze in the future.

5. Ethics Concerns

The text-to-image diffusion model that we use is pre-trained with web-crawled image-text pairs collected by previous works. Despite applying filtering, there may still be potential bias in its internal representation.

References

[1] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 3

[2] 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. 1

[3] 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. 1

[4] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 1

[5] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[6] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary panoptic segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022. 1, 2

[7] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2918–2928, 2021. 1

Figure 2.2. **Qualitative visualization of open-vocabulary panoptic segmentation results on COCO**.

Figure 2.3. **Qualitative visualization of open-vocabulary panoptic segmentation results on ADE20K**.

Figure 2.4. **Qualitative visualization of open-vocabulary panoptic segmentation results on Ego4D**.

[8] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021. 1

[9] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 3

[10] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 1

[11] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 3

[12] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters*, 7(2):5453–5460, 2022. 2

[13] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 3

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[15] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 1

[16] 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. 3

[17] Weiyao Wang, Matt Feiszli, Heng Wang, Jitendra Malik, and Du Tran. Open-world instance segmentation: Exploiting pseudo ground truth from learned pairwise affinity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4422–4432, 2022. 2

[18] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10776–10785, 2021. 2

[19] 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(3):302–321, 2019. 2, 3