

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

Jiarui Xu^{1*} Sifei Liu^{2†} Arash Vahdat^{2†} Wonmin Byeon²
 Xiaolong Wang¹ Shalini De Mello²
¹UC San Diego ²NVIDIA

Project Page: <https://jerryxu.net/ODISE/>



Figure 1. We learn open-vocabulary panoptic segmentation with the internal representation of text-to-image diffusion models. K-Means clustering of the diffusion model’s internal representation shows semantically differentiated and localized information wherein objects are well grouped together (middle figure). We leverage these dense and rich diffusion features to perform open-vocabulary panoptic segmentation (right figure).

Abstract

We present *ODISE*: *O*pen-vocabulary *D*iffusion-based *S*egmenta*t*ion, which unifies pre-trained text-image diffusion and discriminative models to perform open-vocabulary panoptic segmentation. Text-to-image diffusion models have the remarkable ability to generate high-quality images with diverse open-vocabulary language descriptions. This demonstrates that their internal representation space is highly correlated with open concepts in the real world. Text-image discriminative models like CLIP, on the other hand, are good at classifying images into open-vocabulary labels. We leverage the frozen internal representations of both these models to perform panoptic segmentation of any category in the wild. Our approach outperforms the previous state of the art by significant margins on both open-vocabulary panoptic and semantic segmentation tasks. In particular, with COCO training only, our method achieves 23.4 PQ and 30.0 mIoU on the ADE20K dataset, with 8.3 PQ and 7.9 mIoU absolute improvement over the previous state of the art. We open-source our code and models at <https://github.com/NVlabs/ODISE>.

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

1. Introduction

Humans look at the world and can recognize limitless categories. Given the scene presented in Fig. 1, besides identifying every vehicle as a “truck”, we immediately understand that one of them is a pickup truck requiring a trailer to move another truck. To reproduce an intelligence with such a fine-grained and unbounded understanding, the problem of open-vocabulary recognition [40, 62, 83, 96] has recently attracted a lot of attention in computer vision. However, very few works are able to provide a unified framework that parses all object instances and scene semantics at the same time, i.e., panoptic segmentation.

Most current approaches for open-vocabulary recognition rely on the excellent generalization ability of text-image discriminative models [33, 62] trained with Internet-scale data. While such pre-trained models are good at classifying individual object proposals or pixels, they are not necessarily optimal for performing scene-level structural understanding. Indeed, it has been shown that CLIP [62] often confuses the spatial relations between objects [74]. We hypothesize that the lack of spatial and relational understanding in text-image discriminative models is a bottleneck for open-vocabulary panoptic segmentation.

On the other hand, text-to-image generation using diffusion models trained on Internet-scale data [1, 64, 66, 67],

[97] has recently revolutionized the field of image synthesis. It offers unprecedented image quality, generalizability, composition-ability and, semantic control via the input text. An interesting observation is that to condition the image generation process on the provided text, diffusion models compute cross-attention between the text’s embedding and their internal visual representation. This design implies the plausibility of the internal representation of diffusion models being well-differentiated and correlated to high/mid-level semantic concepts that can be described by language. As a proof-of-concept, in Fig. 1 (center), we visualize the results of clustering a diffusion model’s internal features for the image on the left. While not perfect, the discovered groups are indeed semantically distinct and localized. Motivated by this finding, we ask the question of whether Internet-scale text-to-image diffusion models can be exploited to create universal open-vocabulary panoptic segmentation learner for any concept in the wild?

To this end, we propose *ODISE*: Open-vocabulary DIffusion-based panoptic SEmgentation (pronounced *o-disee*), a model that leverages both large-scale text-image diffusion and discriminative models to perform state-of-the-art panoptic segmentation of any category in the wild. An overview of our approach is illustrated in Fig. 2. At a high-level it contains a pre-trained frozen text-to-image diffusion model into which we input an image and its caption and extract the diffusion model’s internal features for them. With these features as input, our mask generator produces panoptic masks of all possible concepts in the image. We train the mask generator with annotated masks available from a training set. A mask classification module then categorizes each mask into one of many open-vocabulary categories by associating each predicted mask’s diffusion features with text embeddings of several object category names. We train this classification module with either mask category labels or image-level captions from the training dataset. Once trained, we perform open-vocabulary panoptic inference with both the text-image diffusion and discriminative models to classify a predicted mask. On many different benchmark datasets and across several open-vocabulary recognition tasks, ODISE achieves state-of-the-art accuracy outperforming the existing baselines by large margins.

Our contributions are the following:

- To the best of our knowledge, ODISE is the first work to explore large-scale text-to-image diffusion models for open-vocabulary segmentation tasks.
- We propose a novel pipeline to effectively leverage both text-image diffusion and discriminative models to perform open-vocabulary panoptic segmentation.
- We significantly advance the field forward by outperforming all existing baselines on many open-vocabulary recognition tasks, and thus establish a new state of the art in this space.

2. Related Work

Panoptic Segmentation. Panoptic segmentation [39] is a fundamental vision task that encompasses both instance and semantic segmentation. However, previous works [5, 9–11, 39, 42, 47, 48, 65, 79, 86, 91] follow a closed closed-vocabulary assumption and only recognize categories present in the training set. They are hence limited in segmenting things/stuff present in finite-sized vocabularies, which are much smaller than the typical vocabularies that we use to describe the real world.

Open-Vocabulary Segmentation. Most prior works on open-vocabulary segmentation either perform object detection with instance segmentation alone [18, 23, 25, 44, 55, 87, 88, 93, 96] or open-vocabulary semantic segmentation alone [23, 40, 83, 95]. In contrast, we propose a novel unified framework for both open-vocabulary instance and semantic segmentation. Another distinction is that prior works only use large-scale models pre-trained for image discriminative tasks, e.g., image classification [30, 51] or image-text contrastive learning [33, 45, 57, 62]. The concurrent work MaskCLIP [16] also uses CLIP [62]. However, such discriminative models’ internal representations are sub-optimal for performing segmentation tasks versus those derived from image-to-text diffusion models as shown in our experiments.

Generative Models for Segmentation. There exist prior works, which are similar in spirit to ours in their use of image generative models, including GANs [3, 19, 35, 36, 98] or diffusion models [14, 17, 31, 34, 60, 69–73, 77] to perform semantic segmentation [2, 21, 41, 54, 76, 92]. They first train generative models on small-vocabulary datasets, e.g., cats [85], human faces [35] or ImageNet [13] and then with the help of few-shot hand-annotated examples per category, learn to classify the internal representations of the generative models into semantic regions. They either synthesize many images and their mask labels to train a separate segmentation network [41, 92]; or directly use the generative model to perform segmentation [2]. Among them, DDPM-Seg [2] shows the state-of-the-art accuracy. These prior works introduce the key idea that the internal representations of generative models may be sufficiently differentiated and correlated to mid/high-level visual semantic concepts and could be used for semantic segmentation. Our work is inspired by them, but it is also different in many respects. While previous works primarily focus on label-efficient semantic segmentation of small closed vocabularies, we, on the other hand, tackle open-vocabulary panoptic segmentation of many more and unseen categories in the wild.

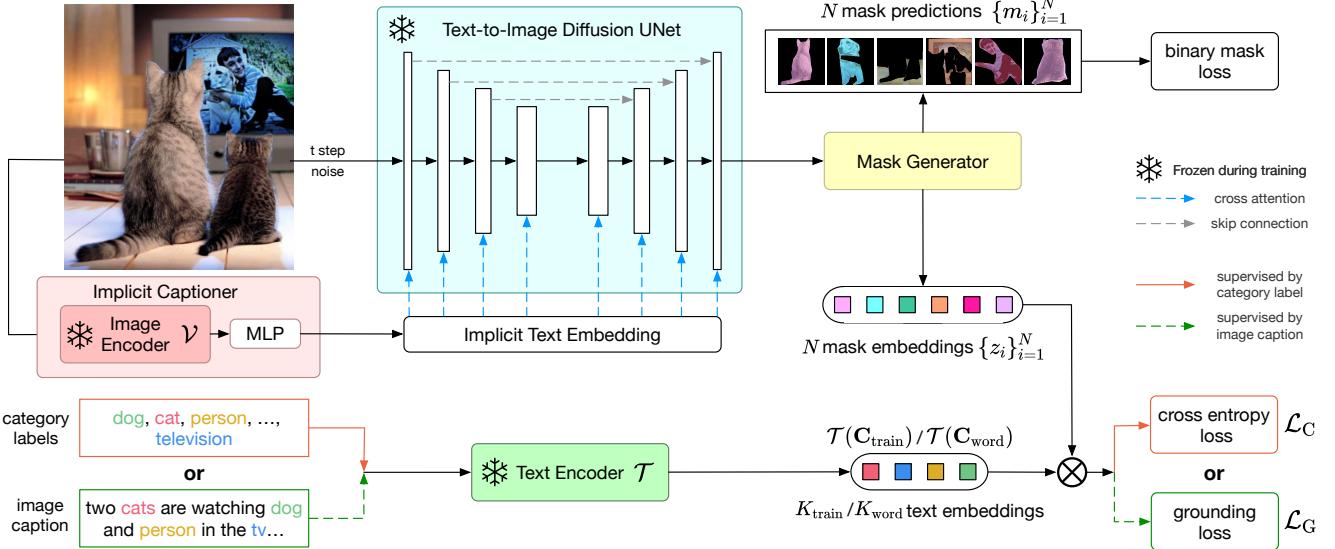


Figure 2. **ODISE Overview and Training Pipeline.** We first encode the input image into an implicit text embedding with an implicit captioner (image encoder \mathcal{V} and MLP). With the image and its implicit text embedding as input, we extract their diffusion features from a frozen text-to-image diffusion UNet (Sec 3.3). With the UNet’s features, a mask generator predicts class-agnostic binary masks and their associated mask embedding features (Sec 3.4). We perform a dot product between the mask embedding features and the text embeddings of training category names (red box) or the nouns of the image’s caption (green box) to categorize them. The similarity matrix for mask classification is supervised by either a cross entropy loss with ground truth category labels (red solid path), or via a grounding loss with the paired image captions (green dash path) (Sec 3.5).

3. Method

3.1. Problem Definition

Following [16, 39], we train a model with a set of base training categories $\mathbf{C}_{\text{train}}$, which may be different from the test categories, \mathbf{C}_{test} , i.e., $\mathbf{C}_{\text{train}} \neq \mathbf{C}_{\text{test}}$. \mathbf{C}_{test} may contain novel categories not seen during training. We assume that during training, the binary panoptic mask annotation for each category in an image is provided. Additionally, we also assume that either the category label of each mask or a text caption for the image is available. During testing, neither the category label nor the caption is available for any image, and only the names of the test categories \mathbf{C}_{test} are provided.

3.2. Method Overview

An overview of our method ODISE, for open-vocabulary panoptic segmentation of any category in the wild is shown in Fig. 2. At a high-level, it contains a text-to-image diffusion model into which we input an image and its caption and extract the diffusion model’s internal features for them (Sec 3.3). With these extracted features as input, and the provided training mask annotations, we train a mask generator to generate panoptic masks of all possible categories in the image (Sec 3.4). Using the provided training images’ category labels or text captions, we also train an open-vocabulary mask classification module. It uses each predicted mask’s diffusion features along with a text encoder’s embeddings of the training category names to classify a

mask (Sec 3.5). Once trained, we perform open-vocabulary panoptic inference with both the text-image diffusion and discriminative models (Sec 3.6 and Fig. 3). In the following sections, we describe each of these components.

3.3. Text-to-Image Diffusion Model

We first provide a brief overview of text-to-image diffusion models and then describe how we extract features from them for panoptic segmentation.

Background A text-to-image diffusion model can generate high-quality images from provided input text prompts. It is trained with millions of image-text pairs crawled from the Internet [59, 64, 67]. The text is encoded into a text embedding with a pre-trained text encoder, e.g., T5 [63] or CLIP [62]. Before being input into the diffusion network, an image is distorted by adding some level of Gaussian noise to it. The diffusion network is trained to undo the distortion given the noisy input and its paired text embedding. During inference, the model takes image-shaped pure Gaussian noise and the text embedding of a user provided description as input, and progressively de-noises it to a realistic image via several iterations of inference.

Visual Representation Extraction The prevalent diffusion-based text-to-image generative models [59, 64, 66, 67] typically use a UNet architecture to learn the denoising process. As shown in the blue block in Fig. 2, the UNet consists of convolution blocks, upsampling and downsampling blocks, skip connections and attention

blocks, which perform cross-attention [78] between a text embedding and UNet features. At every step of the de-noising process, diffusion models use the text input to infer the de-noising direction of the noisy input image. Since the text is injected into the model via cross attention layers, it encourages visual features to be correlated to rich semantically meaningful text descriptions. Thus the feature maps output by the UNet blocks can be regarded as rich and dense features for panoptic segmentation.

Our method only requires a single forward pass of an input image through the diffusion model to extract its visual representation, as opposed to going through the entire multi-step generative diffusion process. Formally, given an input image-text pair (x, s) , we first sample a noisy image x_t at time step t as:

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

where t is the diffusion step we use, $\alpha_1, \dots, \alpha_T$ represent a pre-defined noise schedule where $\bar{\alpha}_t = \prod_{k=1}^t \alpha_k$, as defined in [31]. We encode the caption s with a pre-trained text encoder \mathcal{T} and extract the text-to-image diffusion UNet’s internal features f for the pair by feeding it into the UNet

$$f = \text{UNet}(x_t, \mathcal{T}(s)). \quad (2)$$

It is worth noting that the diffusion model’s visual representation f for x is dependent on its paired caption s . It can be extracted correctly when paired image-text data is available, e.g., during pre-training of the text-to-image diffusion model. However, it becomes problematic when we want to extract the visual representation of images without paired captions available, which is the common use case for our application. For an image without a caption, we could use an empty text as its caption input, but that is clearly sub-optimal, which we also show in our experiments. In what follows, we introduce a novel *Implicit Captioner* that we design to overcome the need for explicitly captioned image data. It also yields optimal downstream task performance.

Implicit Captioner Instead of using an off-the-shelf captioning network to generate captions, we train a network to generate an implicit text embedding from the input image itself. We then input this text embedding into the diffusion model directly. We name this module an implicit captioner. The red block in Fig. 2 shows the architecture of the implicit captioner. Specifically, to derive the implicit text embedding for an image, we leverage a pre-trained frozen image encoder \mathcal{V} , e.g., from CLIP [62] to encode the input image x into its embedding space. We further use a learned MLP to project the image embedding into an implicit text embedding, which we input into text-to-image diffusion UNet. During open-vocabulary panoptic segmentation training, the parameters of the image encoder and of the UNet are unchanged and we only fine-tune the parameters of the MLP.

Finally, the text-to-image diffusion model’s UNet along with the implicit captioner, together form ODISE’s feature extractor that computes the visual representation f for an input image x . Formally, we compute the visual representation f as:

$$\begin{aligned} f &= \text{UNet}(x_t, \text{ImplicitCaptioner}(x)) \\ &= \text{UNet}(x_t, \text{MLP} \circ \mathcal{V}(x)). \end{aligned} \quad (3)$$

3.4. Mask Generator

The mask generator takes the visual representation f as input and outputs N class-agnostic binary masks $\{m_i\}_{i=1}^N$ and their corresponding N mask embedding features $\{z_i\}_{i=1}^N$. The architecture of the mask generator is not restricted to a specific one. It can be any panoptic segmentation network capable of generating mask predictions of the whole image. We can instantiate our method with both bounding box-based [5, 38] and direct segmentation mask-based [9–11, 79] methods. While using bounding box-based methods like [5, 38], we can pool the ROI-Aligned [29] features of each predicted mask’s region to compute its mask embedding features. For segmentation mask-based methods like [9–11, 79], we can directly perform masked pooling on the final feature maps to compute the mask embedding features. Since our representation focuses on dense pixel-wise predictions, we use a direct segmentation-based architecture. Following [29], we supervise the predicted class-agnostic binary masks via a pixel-wise binary cross entropy loss along with their corresponding ground truth masks (treated as class-agnostic ones as well). Next, we describe how we classify each mask, represented by its mask embedding feature, into an open vocabulary.

3.5. Mask Classification

To assign each predicted binary mask a category label from an open vocabulary, we employ text-image discriminative models. These models [33, 57, 62], trained on Internet-scale image-text pairs, have shown strong open-vocabulary classification capabilities. They consist of an image encoder \mathcal{V} and a text encoder \mathcal{T} . Following prior work [23, 40], while training, we employ two commonly used supervision signals to learn to predict the category label of each predicted mask. Next, we describe how we unify these two training approaches in ODISE.

Category Label Supervision Here, we assume that during training we have access to each mask’s ground truth category label. Thus, the training procedure is similar to that of traditional closed-vocabulary training. Suppose that there are $K_{\text{train}} = |\mathbf{C}_{\text{train}}|$ categories in the training set. For each mask embedding feature z_i , we dub its corresponding known ground truth category as $y_i \in \mathbf{C}_{\text{train}}$. We encode the

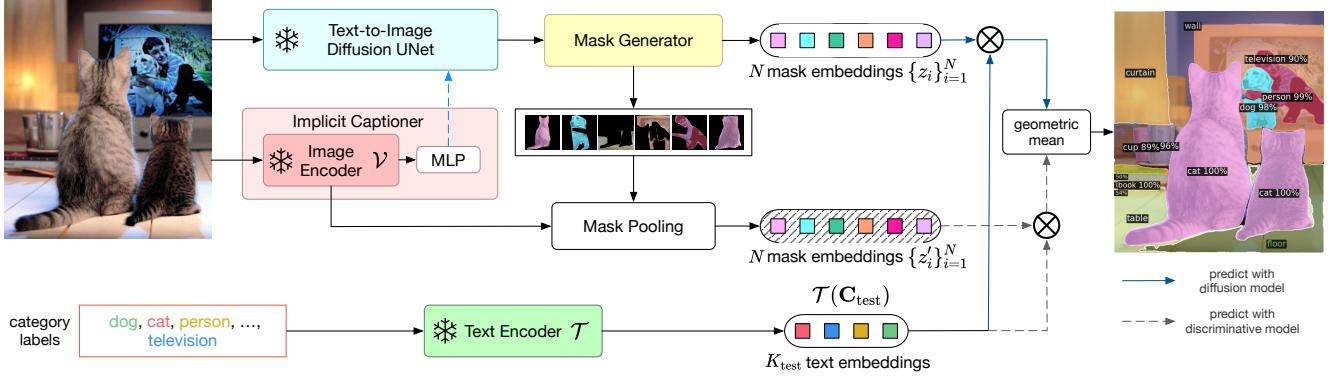


Figure 3. Open-Vocabulary Inference Pipeline. To classify each mask embedding into testing categories \mathbf{C}_{test} , we compute its similarity with the text encoder \mathcal{T} embedding of category names. Besides the mask embeddings from text-to-image diffusion model $\{z_i\}_{i=1}^N$, we also perform mask pooling on the features of image encoder \mathcal{V} from text-image discriminative model to get $\{z'_i\}_{i=1}^N$. We fuse the prediction of diffusion model (blue solid path) and discriminative model (grey dash path) with geometric mean.

names of all the categories in $\mathbf{C}_{\text{train}}$ with the frozen text encoder \mathcal{T} , and define the set of embeddings of all the training categories' names as

$$\mathcal{T}(\mathbf{C}_{\text{train}}) \triangleq [\mathcal{T}(c_1), \mathcal{T}(c_2), \dots, \mathcal{T}(c_{K_{\text{train}}})], \quad (4)$$

where the category name $c_k \in \mathbf{C}_{\text{train}}$. Then we compute the probability of the mask embedding feature z_i belonging to one of the K_{train} classes via a classification loss as:

$$\mathcal{L}_C = \frac{1}{N} \sum_i^N \text{CrossEntropy}(\mathbf{p}(z_i, \mathbf{C}_{\text{train}}), y_i), \quad (5)$$

$$\mathbf{p}(z_i, \mathbf{C}_{\text{train}}) = \text{Softmax}(z_i \cdot \mathcal{T}(\mathbf{C}_{\text{train}})/\tau), \quad (6)$$

where τ is a learnable temperature parameter.

Image Caption Supervision Here, we assume that we do not have any category labels associated with each annotated mask during training. Instead, we have access to a natural language caption for each image, and the model learns to classify the predicted mask embedding features using the image caption alone. To do so, we extract the nouns from each caption and treat them as the grounding category labels for their corresponding paired image. Following [23, 27, 89], we employ a grounding loss to supervise the prediction of the masks' category labels. Specifically, given the image-caption pair $(x^{(m)}, s^{(m)})$, suppose that there are K_{word} nouns extracted from $s^{(m)}$, denoted as $\mathbf{C}_{\text{word}} = \{w_k\}_{k=1}^{K_{\text{word}}}$. Suppose further that we sample B image-caption pairs $\{(x^{(m)}, s^{(m)})\}_{m=1}^B$ to form a batch. To compute the grounding loss, we compute the similarity between each image-caption pair as

$$g(x^{(m)}, s^{(m)}) = \frac{1}{K} \sum_{k=1}^K \sum_{i=1}^N \mathbf{p}(z_i, \mathbf{C}_{\text{word}})_k \cdot \langle z_i, \mathcal{T}(w_k) \rangle, \quad (7)$$

where z_i and $\mathcal{T}(w_k)$ are vectors of the same dimension and $\mathbf{p}(z_i, \mathbf{C}_{\text{word}})_k$ is the k -th element of the vector defined in Eq. 6 after Softmax. This similarity function encourages each noun to be grounded by one or a few masked regions of the image and avoids penalizing the regions that are not grounded by any word at all. Similar to the image-text contrastive loss in [33, 62], the grounding loss is defined by

$$\begin{aligned} \mathcal{L}_G = & -\frac{1}{B} \sum_{m=1}^B \log \frac{\exp(g(x^{(m)}, s^{(m)})/\tau)}{\sum_{n=1}^B \exp(g(x^{(m)}, s^{(n)})/\tau)} \\ & -\frac{1}{B} \sum_{m=1}^B \log \frac{\exp(g(x^{(m)}, s^{(m)})/\tau)}{\sum_{n=1}^B \exp(g(x^{(n)}, s^{(m)})/\tau)}, \end{aligned} \quad (8)$$

where τ is a learnable temperature parameter. Finally, note that we train the entire ODISE model with either \mathcal{L}_C or \mathcal{L}_G , together with the class-agnostic binary mask loss. In our experiments, we explicitly state which of these two supervision signals (label or caption) we use for training ODISE when comparing to the relevant prior works.

3.6. Open-Vocabulary Inference

During inference (Fig. 3), the set of names of the test categories \mathbf{C}_{test} is available. The test categories may be different from the training ones. Additionally, no caption/labels are available for a test image. Hence we pass it through the implicit captioner to obtain its implicit caption; input the two into the diffusion model to obtain the UNet's features; and use the mask generator to predict all possible binary masks of semantic categories in the image. To classify each predicted mask m_i into one of the test categories, we compute $\mathbf{p}(z_i, \mathbf{C}_{\text{test}})$ defined in Eq. 6 using ODISE and finally predict the category with the maximum probability.

In our experiments, we found that the internal representation of the diffusion model is spatially well-differentiated to produce many plausible masks for objects instances.

However, its object classification ability can be further enhanced by combining it once again with a text-image discriminative model, e.g., CLIP [62], especially for open-vocabularies. To this end, here we leverage a text-image discriminative model’s image encoder \mathcal{V} to further classify each predicted masked region of the original input image into one of the test categories. Specifically, as Fig. 3 illustrates, given an input image x , we first encode it into a feature map with the image encoder \mathcal{V} of a text-image discriminative model. Then for a mask m_i , predicted by ODISE for image x , we pool all the features at the output of the image encoder $\mathcal{V}(x)$ that fall inside the predicted mask m_i to compute a mask pooled image feature for it

$$z'_i = \text{MaskPooling}(\mathcal{V}(x), m_i). \quad (9)$$

We use $\mathbf{p}(z'_i, \mathbf{C}_{\text{test}})$ from Eq.6 to compute the final classification probabilities from the text-image discriminative model. Finally, we take the geometric mean of the category predictions from the diffusion and discriminative models as the final classification prediction,

$$\mathbf{p}_{\text{final}}(z_i, \mathbf{C}_{\text{test}}) \propto \mathbf{p}(z_i, \mathbf{C}_{\text{test}})^\lambda \mathbf{p}(z'_i, \mathbf{C}_{\text{test}})^{(1-\lambda)}, \quad (10)$$

where $\lambda \in [0, 1]$ is a fixed balancing factor. We find that pooling the masked features is more efficient and yet as effective as the alternative approach proposed in [15, 25], which crops each of the N predicted masked region’s bounding box from the original image and encodes it separately with the image encoder \mathcal{V} (see details in the supplement).

4. Experiments

We first introduce our implementation details. Then we compare our results against the state of the art on open-vocabulary panoptic and semantic segmentation. Lastly, we present ablation studies to demonstrate the effectiveness of the components of our method.

4.1. Implementation Details

Architecture We use the stable diffusion [66] model pre-trained on a subset of the LAION [68] dataset as our text-to-image diffusion model. We extract feature maps from every three of its UNet blocks and, like FPN [49], resize them to create a feature pyramid. We set the time step used for the diffusion process to $t = 0$, by default. We use CLIP [62] as our text-image discriminative model and its corresponding image \mathcal{V} and text \mathcal{T} encoders everywhere. We choose Mask2Former [10] as the architecture of our mask generator, and generate $N = 100$ binary mask predictions.

Training Details We train ODISE for 90k iterations with images of size 1024^2 and use large scale jittering [22]. Our batch size is 64. For caption-supervised training, we set $K_{\text{word}} = 8$. We use the AdamW [53] optimizer with a learning rate 0.0001 and a weight decay of 0.05. We use the

COCO dataset [50] as our training set. We utilize its provided panoptic mask annotations as the supervision signal for the binary mask loss. For training with image captions, for each image we randomly select one caption from the COCO dataset’s caption [7] annotations.

Inference and Evaluation We evaluate ODISE on ADE20K [94] for open-vocabulary panoptic, instance and semantic segmentation; and the Pascal datasets [20, 56] for semantic segmentation. We also provide the results ODISE for open-vocabulary object detection and open-world instance segmentation in the supplement. We use only a single checkpoint of ODISE for mask prediction on all tasks on all datasets. For panoptic segmentation, we report the panoptic quality (PQ) [39], mean average precision (mAP) on the “thing” categories, and the mean intersection over union (mIoU) metrics (additional SQ and RQ metrics are in the supplement). In panoptic segmentation annotations [39], the “thing” classes are countable objects like people, animals, etc. and the “stuff” classes are amorphous regions like sky, grass, etc. Since we train ODISE with panoptic mask annotations, we can directly infer both instance and semantic segmentation labels with it. When evaluating for panoptic segmentation, we use the panoptic test categories as \mathbf{C}_{test} , and directly classify each predicted mask into the test category with the highest probability. For semantic segmentation, we merge all masks assigned to the same “thing” category into a single one and output it as the predicted mask.

Speed and Model Size ODISE has 28.1M trainable parameters (only 1.8% of the full model) and 1,493.8M frozen parameters. It performs inference for an image (1024^2) at 1.26 FPS on an NVIDIA V100 GPU and uses 11.9 GB memory.

4.2. Comparison with State of the Art

Open-Vocabulary Panoptic Segmentation For open-vocabulary panoptic segmentation, we train ODISE on COCO [50] and test on ADE20K [94]. We report results in Table 1. ODISE outperforms the concurrent work MaskCLIP [16] by 8.3 PQ on ADE20K. Besides the PQ metric, our approach also surpasses MaskCLIP [16] at open-vocabulary instance segmentation on ADE20K, with 8.4 gains in the mAP metric. The qualitative results can be found in Fig. 4 and more in the supplement.

Open-Vocabulary Semantic Segmentation We show a comparison of ODISE to previous work on open-vocabulary semantic segmentation in Table 2. Following the experiment in [23], we evaluate mIoU on 5 semantic segmentation datasets: (a) A-150 with 150 common classes and (b) A-847 with all the 847 classes of ADE20K [94], (c) PC-59 with 59 common classes and (d) PC-459 with full 459 classes of Pascal Context [56], and (e) the classic Pascal

Method	Supervision			ADE20K			COCO		
	label	mask	caption	PQ	mAP	mIoU	PQ	mAP	mIoU
MaskCLIP [16]	✓	✓		15.1	6.0	23.7	-	-	-
ODISE (Ours)	✓	✓		22.6	14.4	29.9	55.4	46.0	65.2
ODISE (Ours)		✓	✓	23.4	13.9	28.7	45.6	38.4	52.4

Table 1. Open-vocabulary panoptic segmentation performance.

Method	Training Dataset	Supervision			mIoU					
		label	mask	caption	A-847	PC-459	A-150	PC-59	PAS-21	COCO
SPNet [82]	Pascal VOC	✓	✓		-	-	-	24.3	18.3	-
ZS3Net [4]	Pascal VOC	✓	✓		-	-	-	19.4	38.3	-
LSeg [40]	Pascal VOC	✓	✓		-	-	-	-	47.4	-
SimBaseline [84]	COCO	✓	✓		-	-	15.3	-	74.5	-
ZegFormer [15]	COCO	✓	✓		-	-	16.4	-	73.3	-
LSeg+ [23]	COCO	✓	✓		3.8	7.8	18.0	46.5	-	55.1
MaskCLIP [16]	COCO	✓	✓		8.2	10.0	23.7	45.9	-	-
ODISE (Ours)	COCO	✓	✓		11.1	14.5	29.9	57.3	84.6	65.2
GroupViT [83]	GCC+YFCC			✓	4.3	4.9	10.6	25.9	50.7	21.1
OpenSeg [23]	COCO		✓	✓	6.3	9.0	21.1	42.1	-	36.1
ODISE (Ours)	COCO		✓	✓	11.0	13.8	28.7	55.3	82.7	52.4

Table 2. Open-vocabulary semantic segmentation performance.

VOC dataset [20] with 20 foreground classes and 1 background class (PAS-21). For a fair comparison to prior work, we train ODISE with either category or image caption labels. ODISE outperforms the existing state-of-the-art methods on open-vocabulary semantic segmentation [16, 23] by a large margin: by 7.6 mIoU on A-150, 4.7 mIoU on A-847, 4.8 mIoU on PC-459 with caption supervision; and by 6.2 mIoU on A-150, 4.5 mIoU on PC-459 with category label supervision, versus the next best method. Notably, it achieves this despite using supervision from panoptic mask annotations, which is noted to be suboptimal for semantic segmentation [10].

We provide comparisons to the state of the art for additional open-vocabulary tasks of object detection and discovery in the supplement.

4.3. Ablation Study

To demonstrate the contribution of each component of our method, we conduct an extensive ablation study. For faster experimentation, we train ODISE with 512^2 resolution images and use image caption supervision everywhere.

Visual Representations We compare the internal representation of text-to-image diffusion models to those of other state-of-the-art pre-trained discriminative and generative models. We evaluate various discriminative models trained with full label, text or self-supervision. In all experiments we freeze the weights of the pre-trained models and use exactly the same training hyperparameters and mask generator as in our method. For each supervision category we select the best-performing and largest publicly available discriminative models. We observe from Table 3 that ODISE outperforms all other models in terms of PQ on

Model	Training Data	ADE20K			COCO		
		PQ	mAP	mIoU	PQ	mAP	mIoU
Pre-trained with class labels							
DeiT-v3(H) [75]	IN-21k	21.4	11.4	28.0	41.4	29.2	52.3
Swin(H) [51]	IN-21k	20.9	10.7	27.7	42.4	31.6	54.0
ConvNeXt(H) [52]	IN-21k	21.0	11.0	27.8	43.1	33.1	54.3
MViT(H) [46]	IN-21k	21.1	11.6	28.1	44.0	36.3	54.5
LDM [66]	IN-1k	20.7	10.9	26.5	41.7	35.3	50.6
Pre-trained with self-supervision							
MoCo-v3(H) [8]	IN-1k	19.3	9.6	25.8	37.1	26.8	47.1
DINO(B) [6]	IN-1k	20.6	10.5	26.3	39.5	29.8	49.5
MAE(H) [28]	IN-1k	21.5	10.9	27.6	37.9	31.6	46.3
BEiT-v2(H) [61]	IN-21k	21.4	11.4	28.0	41.4	29.2	52.3
Pre-trained with text							
CLIP(L) [62]	WIT	20.4	9.6	27.0	40.6	26.7	52.1
CLIP(H) [62]	LAION	21.2	10.8	28.1	41.0	27.9	52.1
ODISE	LAION	23.3	13.0	29.2	44.2	38.3	53.8

Table 3. Comparison with the state-of-the-art visual representations. B, L, H in the parentheses denote the model’s size.

both datasets. To offset any potential bias arising from the larger size of the LAION dataset (2B image-caption pairs) with which the stable diffusion model is trained, versus the smaller datasets used to train the discriminative models, we also compare to CLIP(H) [32, 62]. It is trained on an equal-sized LAION [68] dataset. Despite both models being trained on the same data, our diffusion-based method outperforms CLIP(H) by a large margin on all metrics. This demonstrates that the diffusion model’s internal representation is indeed superior for open-vocabulary segmentation than that of discriminative pre-trained models.

The recent DDPMSEg [2] model is somewhat related to our model. Besides us, it is the only prior work that uses



Figure 4. **Qualitative Visualization** on COCO (first 2 rows) and ADE20K (last row) validation and test sets. To demonstrate open-vocabulary recognition capability, we merge category names of LVIS, COCO and ADE20K together and perform open-vocabulary inference with $\sim 1.5k$ classes directly. “Bamboo”, “swimming pool”, “conveyer belt”, “chandelier”, “booth”, “stool”, “column”, “pool table”, “bannister”, etc., are novel categories from LVIS/ADE20K that are not annotated in COCO. ODISE shows plausible open-vocabulary panoptic results. The supplement contains more visual results.

Captioner	ADE20K			COCO		
	PQ	mAP	mIoU	PQ	mAP	mIoU
(a) Empty	21.8	11.8	27.3	43.5	37.0	52.3
(b) Heuristic [90]	22.2	12.1	28.1	44.0	36.3	53.3
(c) BLIP [43]	22.3	12.4	28.2	44.1	37.1	53.6
(d) Implicit	23.3	13.0	29.2	44.2	38.3	53.8

Table 4. **Ablation results of different caption generators.**

diffusion models and obtains state-of-the-art performance on label-efficient segmentation learning. Since DDPMSSeg relies on category specific diffusion models it is not designed for open-world panoptic segmentation. Hence its direct comparison against our approach is not feasible. As an alternative, we compare against the internal representations of a class-conditioned generative model [66] trained on more categories from ImageNet [13] (LDM row in Table 3). Not surprisingly, we find that despite both generative models being diffusion-based, our approach of using a model trained on Internet-scale data is more effective at generalizing to open-vocabulary categories.

Captioning Generators As discussed in Sec. 3.3, the internal features of a text-to-image diffusion model are dependent on the embedding of the input caption. To derive the optimal set of features for our downstream task, we introduce a novel implicit captioning module to directly generate implicit text embeddings from an image. This module also facilitates inference on images sans paired captions at

test time. Here, we construct several baselines to show the effectiveness of our implicit captioning module. The results are shown in Table 4. The various alternatives that we compare are: providing an empty string to the text encoder for any given image, such that the text embedding for all images is fixed (row (a)); employing two different off-the-shelf image captioning networks to generate an explicit caption for each image on-the-fly (rows (b) and (c)), where (c) [43] is trained on the COCO caption dataset, while (b) [90] is not; and our proposed implicit captioning module (row (d)). Overall, we find that using an explicit/implicit caption is better than using empty text. Furthermore, (c) improves over (b) on COCO but has similar PQ on ADE20K. It may be because the pre-trained BLIP [43] model does not see ADE20K’s image distribution during training and hence it cannot output high-quality captions for it. Lastly, since our implicit captioning module derives its caption from a text-image discriminative model trained on Internet-scale data, it is able to generalize best among all variants compared.

Diffusion Time Steps We also study which diffusion step(s) are most effective for extracting features from, similarly to DDPMSSeg [2]. The noise process is defined in Eq.1. The larger the t value is, the larger the noise distortion added to the input image is. In stable diffusion [66] there are a 1000 total time steps. From Table 5, all metrics decrease as t increases and the best results are for $t=0$ (our final value).

time step	ADE20K			COCO		
	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 5. **Ablation results of different diffusion time steps.** 0+100+200 denotes the concatenation of the features at time steps 0, 100, and 200.

model diffusion	discriminative	ADE20K			COCO		
		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 6. **Ablation results of fusing class predictions of diffusion and discriminative models.**

Concatenating 3 time steps, 0, 100, 200, yields a similar accuracy to $t=0$ only, but is $3\times$ slower. 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 $t=0$.

Mask Classifiers For final open-vocabulary classification (Fig. 3), we fuse class prediction from the diffusion and discriminative models. We report their individual performance in Table 6. Individually the diffusion approach performs better on both datasets than the discriminative only approach. Nevertheless, fusing both together results in higher values on both ADE20K and COCO. Finally, note that even without fusion, our diffusion-only method already surpasses existing methods (see Tables 1, 2).

5. Conclusion

We take the first step in leveraging the frozen internal representation of large-scale text-to-image diffusion models for downstream recognition tasks. ODISE shows the great potential of text-to-image generation models in open-vocabulary segmentation tasks and establishes a new state of the art. This work demonstrates that text-to-image diffusion models are not only capable of generating plausible image but also of learning rich semantic representations. It opens up a new direction for how to effectively leverage the internal representation of text-to-image models for other tasks as well in the future.

Acknowledgements. We thank Golnaz Ghiasi for providing the prompt engineering labels for evaluation. Prof. Xiaolong Wang’s laboratory was supported, in part, by NSF CCF-2112665 (TILOS), NSF CAREER Award IIS-2240014, DARPA LwLL, Amazon Research Award, Adobe Data Science Research Award, and Qualcomm Innovation Fellowship.

References

- [1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 1
- [2] Dmitry Baranichuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 2, 7, 8
- [3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2
- [4] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 7
- [5] 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, 13
- [6] 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 IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 7
- [7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 6
- [8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 7
- [9] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12475–12485, 2020. 2, 4
- [10] 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. 2, 4, 6, 7, 13
- [11] 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. 2, 4, 13
- [12] 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. 13

- [13] 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](#) [8](#)
- [14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. [2](#)
- [15] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. [6](#) [7](#)
- [16] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary panoptic segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022. [2](#) [3](#) [6](#) [7](#) [13](#) [14](#)
- [17] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped langevin diffusion. *arXiv preprint arXiv:2112.07068*, 2021. [2](#)
- [18] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. [2](#)
- [19] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. [2](#)
- [20] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. [6](#) [7](#)
- [21] Danil Galeev, Konstantin Sofiiuk, Danila Rukhovich, Mikhail Romanov, Olga Barinova, and Anton Konushin. Learning high-resolution domain-specific representations with a gan generator. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 108–118. Springer, 2021. [2](#)
- [22] 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. [6](#) [13](#)
- [23] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021. [2](#) [4](#) [5](#) [6](#) [7](#) [13](#)
- [24] 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. [14](#)
- [25] 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. [2](#) [6](#) [13](#)
- [26] 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. [14](#)
- [27] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. [5](#)
- [28] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. [7](#)
- [29] 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. [4](#)
- [30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [2](#)
- [31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. [2](#) [4](#)
- [32] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below. [7](#)
- [33] Chao Jia, Yinfai 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. [1](#) [2](#) [4](#) [5](#)
- [34] Alexia Jolicoeur-Martineau, Rémi Piché-Taillefer, Rémi Tachet des Combes, and Ioannis Mitliagkas. Adversarial score matching and improved sampling for image generation. *arXiv preprint arXiv:2009.05475*, 2020. [2](#)
- [35] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. [2](#)
- [36] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. [2](#)
- [37] 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. [14](#)
- [38] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019. [4](#)

- [39] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019. [2](#), [3](#), [6](#)
- [40] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. [1](#), [2](#), [4](#), [7](#)
- [41] Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21330–21340, 2022. [2](#)
- [42] Feng Li, Hao Zhang, Shilong Liu, Lei Zhang, Lionel M Ni, Heung-Yeung Shum, et al. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *arXiv preprint arXiv:2206.02777*, 2022. [2](#)
- [43] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. [8](#)
- [44] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. [2](#)
- [45] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. [2](#)
- [46] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. [7](#)
- [47] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 214–223, 2021. [2](#)
- [48] Zhiqi Li, Wenhui Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, Ping Luo, and Tong Lu. Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1280–1289, 2022. [2](#)
- [49] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. [6](#)
- [50] 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. [6](#), [14](#), [15](#)
- [51] 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](#), [7](#)
- [52] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. [7](#)
- [53] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. [6](#), [13](#)
- [54] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sredit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. [2](#)
- [55] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection with vision transformers. *arXiv preprint arXiv:2205.06230*, 2022. [2](#)
- [56] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. [6](#)
- [57] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021. [2](#), [4](#)
- [58] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kotschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. [13](#)
- [59] 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. [3](#)
- [60] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. [2](#)
- [61] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. [7](#)
- [62] 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. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

- [63] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020. 3
- [64] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 3
- [65] Jiawei Ren, Cunjun Yu, Zhongang Cai, Mingyuan Zhang, Chongsong Chen, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Refine: Prediction fusion network for panoptic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2477–2485, 2021. 2
- [66] 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. 1, 3, 6, 7, 8
- [67] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 3
- [68] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 6, 7
- [69] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2
- [70] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2
- [71] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 2
- [72] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020. 2
- [73] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2
- [74] Sanjay Subramanian, Will Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022. 1
- [75] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118*, 2022. 7
- [76] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4475–4485, 2021. 2
- [77] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021. 2
- [78] 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. 4
- [79] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5463–5474, 2021. 2, 4
- [80] 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. 14
- [81] 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. 14
- [82] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, 2019. 7
- [83] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic Segmentation Emerges from Text Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 1, 2, 7
- [84] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757*, 2021. 7
- [85] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 2
- [86] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means mask transformer. In *European Conference on Computer Vision*, pages 288–307. Springer, 2022. 2
- [87] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. *arXiv preprint arXiv:2203.11876*, 2022. 2
- [88] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 2
- [89] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions.

- In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 5
- [90] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Chormanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 8
- [91] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021. 2
- [92] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021. 2
- [93] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luwei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 2
- [94] 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. 6, 14
- [95] Chong Zhou, Chen Change Loy, and Bo Dai. Denseclip: Extract free dense labels from clip. *arXiv preprint arXiv:2112.01071*, 2021. 2
- [96] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. *arXiv preprint arXiv:2201.02605*, 2022. 1, 2
- [97] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2022. 1
- [98] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2

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

A. 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 [22] 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 [53] 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 [5, 10, 11], 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 [23] 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 [25] 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.

B. Experiments

B.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 [16] on both the thing (Th) and stuff (St) categories of the ADE20K dataset in Table B.1. Here, all models were trained on COCO. ODISE outperforms MaskCLIP [16] 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 B.1. Detailed panoptic segmentation metrics on ADE20K. ODISE outperforms MaskCLIP [16] w.r.t. all metrics.

We also evaluate ODISE trained on COCO on the Cityscapes [12] and Mapillary Vistas [58] datasets in Table B.2. Since the source code for MaskCLIP [16] is not publicly available, we regard ODISE’s implementation with CLIP(H) features (from Table 3 of the main paper)

Method	Cityscapes			Mapillary Vistas		
	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 B.2. **Results of panoptic segmentation on Cityscapes and Mapillary Vistas.** ODISE outperforms CLIP(H) by a large margin on both datasets.

as a close proxy to MaskCLIP and compare against it (Table B.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 B.3. Here too, we regard the variant of ODISE with CLIP(H) features as a proxy to MaskCLIP [16] and compare against it. ODISE outperforms its CLIP(H) variant by a large margin.

Method	COCO			ADE20K		
	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 B.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 [26] dataset (Table B.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 [50] 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 [16] by a large margin w.r.t. both mAP and mAP_r . Note that the validation split of LVIS [26] has overlapping images with COCO [50]'s training split, but the category labels of LVIS are only available during inference.

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 [80], we report the average recall of 100 mask proposals (AR@100) on the UVO [81] and ADE20K [94] datasets. As reported in Table B.5, here

Method	Supervision			LVIS	
	label	mask	caption	mAP	mAP_r
MaskCLIP [16]	✓	✓		8.4	-
ODISE (Ours)	✓	✓		15.4	19.4
ODISE (Ours)		✓	✓	17.1	21.1

Table B.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.

too we outperform the existing state of the art [80] 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.

Method	AR@100		
	UVO	ADE20K	COCO
LDET [37]	42.6	-	-
GGN [80]	43.4	21.0	-
ODISE (Ours)	57.7	30.3	56.6

Table B.5. **Open-world Instance Segmentation.** ODISE outperforms GGN on discovering open-world instances on both the UVO and ADE20K datasets.

B.2. Ablation Study

Visual Representations In Fig. B.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 B.2 and Table B.3 further substantiate diffusion features' superiority over those of CLIP's.

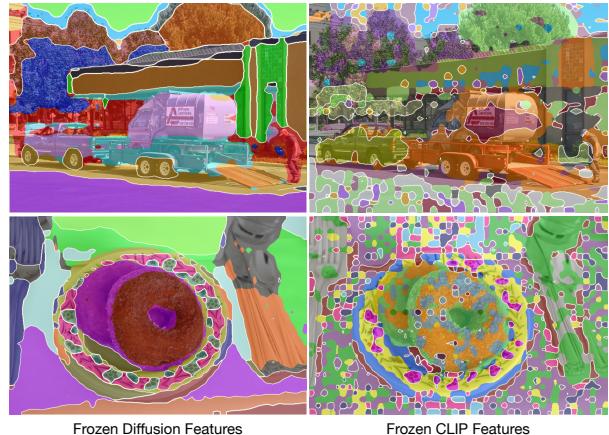


Figure B.1. **K-mean clustering of text-to-image diffusion and CLIP models' internal features.** The internal features of the diffusion model are much more semantically differentiated than those from CLIP.



Figure B.2. Qualitative visualization of open-vocabulary panoptic segmentation results on COCO.



Figure B.3. Qualitative visualization of open-vocabulary panoptic segmentation results on ADE20K.



Figure B.4. Qualitative visualization of open-vocabulary panoptic segmentation results on Ego4D.

C. Qualitative Results

To demonstrate the open-vocabulary recognition capabilities of ODISE, we merge the category names from LVIS [26], COCO [50], ADE20K [94] together and perform open-vocabulary inference with $\sim 1.5k$ test classes. We only train ODISE on COCO’s [50] training dataset and evaluate open-vocabulary panoptic inference on ADE20K [94] and Ego4D [24]. The qualitative results on COCO’s [50] validation dataset, ADE20K [94] and Ego4D [24] are shown in Fig. B.2, Fig. B.3 and Fig. B.4, respectively. Most categories, e.g., “police cruiser”, “flag”, “conveyor belt”, “chandelier”, “aquarium”, “grocery bag”, “power shovel”, etc., are novel categories from LVIS [26] or ADE20K [94] that are not annotated in COCO [50]. It is worth noting that Ego4D [24] is a video dataset, which consists of diverse ego-centric videos. Despite the large domain gap between the testing dataset Ego4D [24] and our training dataset COCO [50], ODISE still outputs good-quality plausible panoptic segmentation results on Ego4D’s novel categories.

D. 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.

E. 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.