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

Jiarui Xu1*  Sifei Liu2†  Arash Vahdat2†  Wonmin Byeon2
Xiaolong Wang1  Shalini De Mello2
1UC San Diego  2NVIDIA

Abstract

We present ODISE: Open-vocabulary DIffusion-based panoptic SEgmentation, 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.

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 [36, 57, 76, 89] 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 [30, 57] 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 [57] often confuses the spatial relations between objects [69]. 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, 59, 61, 62,
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 SEgmentation (pronounced o-di-se), 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 [35] is a fundamental vision task that encompasses both instance and semantic segmentation. However, previous works [5, 9–11, 35, 38, 43, 44, 60, 74, 79, 84] 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 [17, 22, 23, 40, 51, 80, 81, 86, 89] or open-vocabulary semantic segmentation alone [22, 36, 76, 88]. 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 [27, 47] or image-text contrastive learning [30, 41, 53, 57]. The concurrent work MaskCLIP [15] also uses CLIP [57]. 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, 18, 32, 33, 91] or diffusion models [13, 16, 28, 31, 55, 64–68, 72] to perform semantic segmentation [2, 20, 37, 50, 71, 85]. They first train generative models on small-vocabulary datasets, e.g., cats [78], human faces [32] or ImageNet [12] 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 [37, 85]; 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.
3. Method

3.1. Problem Definition

Following [15, 35], we train a model with a set of base training categories $C_{\text{train}}$, which may be different from the test categories, $C_{\text{test}}$, i.e., $C_{\text{train}} \neq C_{\text{test}}$. $C_{\text{test}}$ 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 $C_{\text{test}}$ 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 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).

3.3. 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 [54, 59, 62]. The text is encoded into a text embedding with a pre-trained text encoder, e.g., T5 [58] or CLIP [57]. 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 [54, 59, 61, 62] 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 [73] 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:

\[
x_t \triangleq \sqrt{\alpha_t} x + \sqrt{1 - \alpha_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),
\]

where \(t\) is the diffusion step we use, \(\alpha_1, \ldots, \alpha_T\) represent a pre-defined noise schedule where \(\alpha_t = \prod_{k=1}^t \alpha_k\), as defined in [28]. 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

\[
f = \text{UNet}(x_t, \mathcal{T}(s)).
\]

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 [57] 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:

\[
f = \text{UNet}(x_t, \text{ImplicitCaptioner}(x)) = \text{UNet}(x_t, \text{MLP} \circ \mathcal{V}(x)).
\]

### 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, 34] and direct segmentation mask-based [9–11, 74] methods. While using bounding box-based methods like [5, 34], we can pool the ROI-Aligned [26] features of each predicted mask’s region to compute its mask embedding features. For segmentation mask-based methods like [9–11, 74], 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 [26], 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 [30, 53, 57], 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 [22, 36], 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 \(K_{\text{train}} = |\mathcal{C}_{\text{train}}|\) categories in the training set. For each mask embedding feature \(z_i\), we dub its corresponding known ground truth category as \(y_i \in \mathcal{C}_{\text{train}}\). We encode the names of all the categories in \(\mathcal{C}_{\text{train}}\) with the frozen text encoder \(\mathcal{T}\), and define the set of embeddings of all the training
categories’ names as
\[ \mathcal{T} (C_{\text{train}}) = \{ \mathcal{T}(c_1), \mathcal{T}(c_2), \ldots, \mathcal{T}(c_{K_{\text{train}}}) \}, \]  
(4)
where the category name \( c_k \in C_{\text{train}} \). Then we compute the probability of the mask embedding feature \( z_i \) belonging to one of the \( K_{\text{train}} \) classes via a classification loss as:
\[ \mathcal{L}_C = \frac{1}{N} \sum_i^N \text{CrossEntropy} (p(z_i, C_{\text{train}}), y_i), \]  
(5)
\[ p(z_i, C_{\text{train}}) = \text{Softmax}(z_i \cdot \mathcal{T}(C_{\text{train}})/\tau), \]  
(6)
where \( \tau \) 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 [22, 24, 82], 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 \( K_{\text{word}} \) nouns extracted from \( s^{(m)} \), denoted as \( C_{\text{word}} = \{ w_k \}_{k=1}^{K_{\text{word}}} \). 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
\[ g(x^{(m)}, s^{(m)}) = \frac{1}{K} \sum_{k=1}^K \sum_i^N p(z_i, C_{\text{word}})[k](z_i, \mathcal{T}(w_k)), \]  
(7)
where \( z_i \) and \( \mathcal{T}(w_k) \) are vectors of the same dimension and \( p(z_i, C_{\text{word}})[k] \) is the \( k \)-th element of the vector defined in Eq. 6 using 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 [30, 57], the grounding loss is defined by
\[ \mathcal{L}_G = -\frac{1}{B} \sum_{m=1}^B \log \frac{\exp(g(x^{(m)}, s^{(m)})/\tau)}{\sum_{m=1}^B \exp(g(x^{(m)}, s^{(m)})/\tau)} - \frac{1}{B} \sum_{m=1}^B \log \frac{\exp(g(x^{(m)}, s^{(m)})/\tau)}{\sum_{n=1}^B \exp(g(x^{(n)}, s^{(m)})/\tau)}, \]  
(8)
where \( \tau \) 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.
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 $V$ 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 $V(x)$ that fall inside the predicted mask $m_i$ to compute a mask pooled image feature for it

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

We use $p(z'_i, C_{\text{test}})$ 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,

$$p_{\text{final}}(z_i, C_{\text{test}}) \propto p(z_i, C_{\text{test}})^\lambda p(z'_i, C_{\text{test}})^{(1-\lambda)},$$

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 [14, 23], which crops each of the $N$ predicted masked region’s bounding box from the original image and encodes it separately with the image encoder $V$ (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 [61] model pretrained on a subset of the LAION [63] dataset as our text-to-image diffusion model. We extract feature maps from every three of its UNet blocks and, like FPN [45], 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 [57] as our text-image discriminative model and its corresponding image $V$ and text $T$ 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 [21]. Our batch size is 64. For caption-supervised training, we set $K_{\text{word}} = 8$. We use the AdamW [49] optimizer with a learning rate 0.0001 and a weight decay of 0.05. We use the COCO dataset [46] 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 [87] for open-vocabulary panoptic, instance and semantic segmentation; and the Pascal datasets [19, 52] 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) [35], 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 [35], 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 $C_{\text{test}}$, 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 [46] and test on ADE20K [87]. We report results in Table 1 (and more in the supplement). ODISE outperforms the concurrent work MaskCLIP [15] by 8.3 PQ on ADE20K. Besides the PQ metric, our approach also surpasses MaskCLIP [15] at open-vocabulary instance segmentation on ADE20K, with 8.4 points gain in mAP. ODISE’s 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 [22], 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 [87], (c) PC-59 with 59 common classes and (d) PC-459 with full 459 classes of Pascal Context [52] and (e) the classic Pascal VOC dataset [19] 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 [15, 22] by a large margin: 7.6 mIoU on A-150, 4.7 mIoU on A-847,
Table 1. Open-vocabulary panoptic segmentation performance.

<table>
<thead>
<tr>
<th>Method</th>
<th>ADE20K Supervision</th>
<th>COCO Supervision</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaskCLIP [15]</td>
<td>✓ ✓</td>
<td>- -</td>
</tr>
<tr>
<td>ODISE (Ours)</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
</tr>
</tbody>
</table>

22.6 14.4 29.9 55.4 46.0 65.2

Table 2. Open-vocabulary semantic segmentation performance.

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].

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 512² resolution images and use image caption supervision everywhere. Specifically, we evaluate different visual representations; caption generators; open-vocabulary inference pipelines and diffusion time-step(s) used to extract features (the latter two ablations are in the supplement).

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 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) [29, 57], which is trained on an equal-sized LAION [63] dataset. Our diffusion-based method outperforms CLIP(H) on all metrics. In the supplement (Fig 3.4), we show k-means clustering of frozen diffusion and CLIP features and find that the diffusion features are much more semantically differentiated. This demonstrates that the diffusion model’s internal representation is indeed superior for open-vocabulary segmentation that 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 diffusion models and obtains state-of-the-art performance on label-efficient segmentation learning. Since DDPMSeg relies on category specific diffusion models it is not de-
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”, “conveyor 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.

<table>
<thead>
<tr>
<th>Captioner</th>
<th>ADE20K PQ</th>
<th>ADE20K mAP</th>
<th>ADE20K mIoU</th>
<th>COCO PQ</th>
<th>COCO mAP</th>
<th>COCO mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Empty</td>
<td>21.8</td>
<td>11.8</td>
<td>27.3</td>
<td>43.5</td>
<td>37.0</td>
<td>52.3</td>
</tr>
<tr>
<td>(b) Heuristic [83]</td>
<td>22.2</td>
<td>12.1</td>
<td>28.1</td>
<td>44.0</td>
<td>36.3</td>
<td>53.3</td>
</tr>
<tr>
<td>(c) BLIP [39]</td>
<td>22.3</td>
<td>12.4</td>
<td>28.2</td>
<td>44.1</td>
<td>37.1</td>
<td>53.6</td>
</tr>
<tr>
<td>(d) Implicit</td>
<td><strong>23.3</strong></td>
<td><strong>13.0</strong></td>
<td><strong>29.2</strong></td>
<td><strong>44.2</strong></td>
<td><strong>38.3</strong></td>
<td><strong>53.8</strong></td>
</tr>
</tbody>
</table>

Table 4. Ablation results of different caption 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 an implicit text embedding 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 (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) [39] is trained on the COCO caption dataset, while (b) [83] 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 [39] model does not see ADE20K’s image distribution during training. 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.

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 large generative models in open-vocabulary segmentation tasks and establishes a new state of the art. We demonstrate that diffusion models are not only capable of generating plausible images but also of learning rich semantic representations. Our work opens a new direction for how to effectively leverage the internal representation of large diffusion 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


