This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Repurposing SAM for User-Defined Semantics Aware Segmentation

Rohit Kundu[1], Sudipta Paul[2,*], Arindam Dutta[1], Amit K. Roy-Chowdhury[1]
[1]University of California, Riverside [2]Samsung Research America
{rkund006@, spaul007@, adutt020@, amitrc@ece.}ucr.edu

Abstract

The Segment Anything Model (SAM) excels at generating precise object masks from input prompts but lacks semantic awareness, failing to associate its generated masks with specific object categories. To address this limitation, we propose U-SAM, a novel framework that imbibes semantic awareness into SAM, enabling it to generate targeted masks for user-specified object categories. Given only object class names as input from the user, U-SAM provides pixel-level semantic annotations for images without requiring any labeled/unlabeled samples from the test data distribution. Our approach leverages synthetically generated or web crawled images to accumulate semantic information about the desired object classes. We then learn a mapping function between SAM's mask embeddings and object class labels, effectively enhancing SAM with granularity-specific semantic recognition capabilities. As a result, users can obtain meaningful and targeted segmentation masks for specific objects they request, rather than generic and unlabeled masks. We evaluate U-SAM on PASCAL VOC 2012 and MSCOCO-80, achieving significant $mIoU$ improvements of $+17.95\%$ and $+5.20\%$, respectively, over state-of-the-art methods. By transforming SAM into a semantically aware segmentation model, U-SAM offers a practical and flexible solution for pixel-level annotation across diverse and unseen domains in a resource-constrained environment. The project page is available at rohit-kundu.github.io/U-SAM.

1. Introduction

Semantic segmentation [2, 3, 6, 10, 17] has become a cornerstone of modern deep learning, enabling pixel-level understanding of images for applications in autonomous driving [7, 50], satellite imaging [27, 46], and robotics [13, 47]. However, its success heavily depends on large amounts of labeled data for supervised training [52, 57], which makes manual annotation both expensive and time-intensive. Labeling every pixel in an image to generate semantic masks is particularly laborious, creating a significant bottleneck.

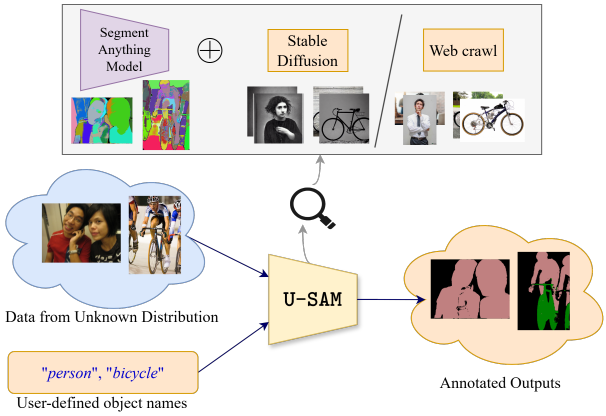


Figure 1. **Problem Overview**: We propose a novel and flexible pipeline, U-SAM, to tackle the challenging problem of generating pixel-level semantic annotations for user-specified object classes without any manual supervision. Given only a set of object class names, U-SAM can produce accurate semantic segmentation masks on any user-provided image data. By enhancing SAM with semantic region recognition capabilities and leveraging synthetic images generated by Stable Diffusion or web crawled images for the desired object classes, our approach addresses a highly complex segmentation problem with robust generalization.

The Segment Anything Model (SAM) [21] has emerged as a powerful foundation model for segmentation tasks, capable of generating precise object masks from input prompts. However, SAM's lack of semantic awareness restricts its utility in real-world applications where precise object identification is crucial. For instance, SAM can segment objects based on prompts like points or bounding boxes but cannot inherently distinguish between different object categories. This limitation means that SAM produces numerous unlabeled masks, requiring additional processing to identify specific objects.

To bridge this gap, we propose a novel approach that enhances SAM with the ability to generate semantic segmentation masks for user-specified object categories. Our method, *User-Defined Semantics Aware SAM* or U-SAM, requires only the names of the target object categories as input and outputs semantic segmentation masks for user-provided

*Work done while at UCR.

datasets during inference. Unlike existing methods that rely on labeled [14, 40] or unlabeled [15, 24, 58] data from the test dataset distribution, our approach eliminates the need for in-domain training data or prior knowledge of the test dataset distribution. This generalization allows users to provide images from any dataset, and our model reliably generates semantic masks tailored to the specified objects, such as "person" or "bicycle," as shown in Figure 1.

Notably, most weakly supervised segmentation methods [16, 37, 49], along with in-domain training data, also require object labels for every image. In contrast, our pipeline relaxes even this assumption by requiring only dataset-level object names. During inference, we do not know which images will contain which objects, making our problem setup more challenging than traditional unsupervised or weakly supervised learning and closer to zero-shot learning. This distinction underscores the novelty and complexity of our approach, which addresses a significant gap in current segmentation methodologies.

While generative models like Stable Diffusion [36] are capable of producing high-quality synthetic data, state-of-the-art semantic segmentation methods have yet to fully leverage such models for downstream tasks. So, to address this new segmentation challenge, we leverage synthetic images generated by Stable Diffusion [36] or web-crawled images corresponding to user-specified object categories. We then train a classifier head on these images to enhance SAM with semantic recognition capabilities. Once trained, U-SAM can perform semantic segmentation on any dataset—whether it is PASCAL VOC [11], MSCOCO-80 [29], or a custom user-created dataset—without requiring retraining or fine-tuning for each new test distribution. This capability makes U-SAM highly practical and versatile for real-world applications.

By transforming SAM [21] into a semantically aware model, we address a critical need in real-world applications where users require precise object identification without the burden of manual annotation or dataset-specific training. Our approach not only enhances SAM's capabilities but also offers a practical solution for pixel-level annotation across diverse and unseen domains.

In summary, our **contributions** are as follows:

- We propose U-SAM, a novel framework that enhances the SAM [21] by imbuing it with semantic awareness, allowing it to generate targeted semantic masks for user-specified object categories without manual supervision.
- Our approach eliminates the need for in-domain labeled/unlabeled training data from the test distribution, allowing U-SAM to operate without prior dataset knowledge. We leverage synthetic images generated by Stable Diffusion [36] or web crawled images based on user-provided object class names, creating a flexible pipeline that generalizes across diverse datasets without requiring retraining or fine-tuning for each new distribution.
- Extensive quantitative experiments demonstrate superior performance of U-SAM compared to state-of-the-art unsupervised ($+17.95\%$ on PASCAL VOC 2012 [11] and $+5.20\%$ on MSCOCO-80 [29]) and weakly supervised semantic segmentation methods ($+21.67\%$ on PASCAL VOC 2012 [11] and $+7.99\%$ on MSCOCO-80 [29]) trained using Stable Diffusion [36]-generated synthetic images or web crawled images.

2. Related Work

Efficient Annotation: Efficient annotation methods aim to reduce the time and cost associated with the annotation process while maintaining or improving annotation quality. There are few works [5, 28] that use human-in-the-loop approaches to generate image-level annotation efficiently. Interactive object segmentation methods [4, 30] propose coarse segmentation masks which are then iteratively corrected by collecting point annotations (object vs background) from a human annotator. However, to the best of our knowledge, the research area for developing semantic segmentation annotation framework without access to any manual supervision remains unexplored.

Unsupervised Semantic Segmentation: Unsupervised semantic segmentation [8, 12, 33, 39, 44, 55, 56] involves the segmentation of an image into meaningful regions without using labeled training data. Researchers have investigated diverse approaches to tackle this challenge, incorporating methods from unsupervised learning [20, 41, 54], clustering [1, 19, 53], and self-supervised learning [18, 38, 45]. MaskContrast [42] stands out as an approach that performs unsupervised semantic segmentation within a contrastive learning framework. It assigns labels to the predicted clusters using Hungarian matching [22], optimizing test-time performance metrics by associating labels with clusters that maximize segmentation accuracy. ACSeg [26] employs a pre-trained ViT model to mine "concepts" from the pixel representation space of the unlabeled training images. Unlike previous methods, ACSeg [26] does not predefine the number of clusters an image will be partitioned into. Instead, they adaptively conceptualize on different images due to varying complexity in individual images.

However, these existing unsupervised methods assume access to a large unlabeled training dataset. This contrasts with our problem formulation, where we assume a scenario without access to such unlabeled data. In our approach, we operate under the assumption that we lack prior knowledge of the specific data distribution the user requires for annotation. We perform on-the-fly inference with our U-SAM model directly on the user's data, eliminating the need for pre-existing knowledge of the data distribution.

SAM Variants: Several methods that build on SAM's capabilities have been developed recently. SemanticSAM [25]

is a fully supervised method that segments and recognizes open-set objects at any granularity, and is trained on a combination of the SA-1B dataset [21] with other panoptic and part-segmentation datasets. In contrast, we neither have access to in-distribution training data, nor pixel-level annotations or image-level object labels. Our only requirement is the user-defined dataset-level object labels, meaning we lack information on which objects are present in each image. RegionSpot [48] leverages the SAM model and CLIP's [34] Vision-Language feature embeddings to perform object detection. However, the method is dependent on (unlabeled) box annotations for every image, necessitating a comprehensive annotation effort for each object category. Grounded-SAM [35] and LangSAM [32] are two similar open-source projects based on SAM which aim to segment and detect anything with natural language prompts. However, they are weakly supervised methods that require the object labels for every image, which are used to generate bounding box annotations using GroundingDINO [31], which in turn serves as the box prompts for SAM. In our work, we do not utilize such object-level annotations, which may be difficult to get in many applications.

3. Methodology

In this section, we define the problem statement (Sec. 3.1) and present an overview of our `U-SAM` framework (Sec. 3.2). We then briefly explain SAM's architecture (Sec. 3.3) and describe our strategy for collecting semantic information on different classes of interest (Sec. 3.4). Finally, we explain how this semantic information is linked with SAM's mask embeddings to develop our annotation system (Sec. 3.5) and outline the inference steps of the annotation system (Sec. 3.6).

3.1. Problem Statement

Consider we have a set of unlabeled N images $\mathcal{X} = \{x_i\}_{i=1}^{N}$ where, $x_i \in \mathbb{R}^{H\times W\times 3}$. Given a set of $C+1$ target classes $\mathcal{C} = \{c_i\}_{i=0}^{C}$ (C object classes and one background class), our objective is to generate semantic masks $\mathcal{U} = \{u_i\}_{i=1}^{N}$ for the unlabeled N images. Here, $u_i \in \mathcal{C}^{H\times W}$ and each image can contain variable number of objects. We also consider that apart from $\mathcal{X}$, we do not have access to more unlabeled images from the same distribution for unsupervised/self-supervised training.

3.2. Framework Overview

To generate semantic segmentation masks, the system needs to be able to distinguish different regions of an image and understand the semantics of those regions simultaneously. We leverage domain agnostic Segment Anything Model (SAM) [21] as the backbone of the system to distinguish the regions. Although SAM can generate masks at any granularity based on input prompts, it does not have the semantic understanding of the objects present in an image. To guide the semantic understanding, we utilize Stable Diffusion generated synthetic images/web crawled images of our classes of interest. Since, we know the image-level labels of the synthetic or web crawled images, we train a classifier head on top of SAM's mask decoder in a weakly-supervised setup. We learn a mapping for every generated mask of SAM by parsing the mask embedding and employing Multi-Layer Perceptron (MLP) layers to map it to class labels. We use multiple instance learning loss to learn the mapping function. Since SAM mask embeddings are domain agnostic, our learned classifier is also robust to distribution shifts and can be directly applicable to the unlabeled dataset of interest. Figure 2 provides a detailed architectural overview.

3.3. Segment Anything Model Architecture

Segment Anything Model (SAM) [21] is a foundation model for image segmentation task trained on a dataset consisting of 11M images and 1B masks. It can generate segmentation masks for images at varying granularity levels based on input prompts. It consists of three main parts:

Image Encoder: The input images are processed using a pre-trained Vision Transformer [9] modified to handle high resolution images of dimension $\mathbb{R}^{3\times 1024\times 1024}$. It is run once per image to obtain features of dimension $\mathbb{R}^{256\times 64\times 64}$ and can be applied prior to prompting the model.

Prompt Encoder: SAM takes sparse prompt inputs in the form of point or box annotations for mask inference. Alternatively, in SAM's "*automatic*" mode, users don't need to manually define points or bounding boxes in the image; instead, a grid of uniformly spaced points (say, 'd' number of total points) is generated throughout the image which are used as point annotations to prompt SAM. These point prompts are represented as positional embeddings summed with the learned embeddings (that indicates if a point is in the foreground or background) in SAM. Since, we do not have any prior information on where the object of interest is located, we use this "*automatic*" mode to generate all possible masks.

Mask Decoder: SAM's mask decoder consists of a modified Transformer Decoder block [43] followed by a dynamic mask prediction heads. It maps the image and prompt embeddings along with an output token to a mask. A learned output token embedding which is analogous to the `[class]` token in [9], is first inserted into the set of prompt embeddings. This, along with the image embeddings are given as input to the transformer decoder that updates the output token embeddings (denoted by $m_{tokens} \in \mathbb{R}^{d\times 4\times 256}$). The image embeddings are upscaled by $4\times$ and a point-wise product is taken with the output of a 3-layer MLP whose inputs are m_{tokens}. This product is the mask prediction of SAM, and this is performed for all of the d prompts (point annotations) resulting in a set of d binary

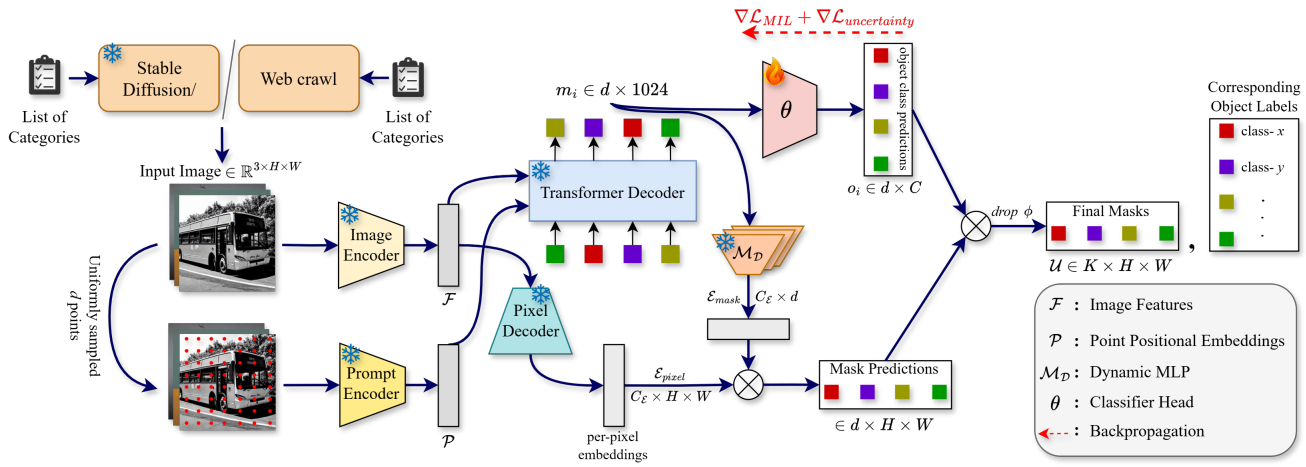


Figure 2. **Overview of the U-SAM architecture:** Given the list of user-defined target categories $\mathcal{C}$, we use Stable Diffusion [36] to generate a synthetic single-object image dataset which is encoded by SAM's [21] image encoder and a uniformly spaced grid of d points are generated across the image to prompt SAM. The image and point embeddings are passed into a transformer decoder, the output mask embeddings m_i (here, $m_i \in \mathbb{R}^{d\times 1024}$, corresponding to the d masks predicted by SAM) of which are used to train a classifier head θ to predict objects using a Multiple Instance Learning (MIL) setup and uncertainty losses (refer Sec. 3.5).

masks $\in \mathbb{Z}_2^{d\times 256\times 256}$ for each image.

3.4. Fetching Semantic Information

Although SAM [21] can predict masks by distinguishing different regions of the image, it does not have semantic understanding of the predicted masks. To enable SAM with recognition capability of user defined classes, we collect semantic information of those classes using synthetic images or web crawled images.

Synthetic Images: We use the Stable Diffusion [36] model which can generate hyper-realistic images given a text prompt. For each of the user-defined object classes in $\mathcal{C}$, we use the text prompt "a photo of a {c}" $\forall c \in \mathcal{C}$ and generate 200 single object images. This synthetic single-object image dataset serves as guide to enable semantic understanding for Segment Anything Model.

Web Crawled Images: Alternative to synthetic images, we can also use web crawled images for the same purpose. Through web crawling, we acquire single-object images for each user-defined object class in $\mathcal{C}$ by utilizing them as search queries on the web. The web crawled images serve as a valuable addition to the training data, providing diversity in backgrounds and object contexts. This approach enhances the model's ability to generalize across various visual scenarios and improves its discriminatory capacity for user-defined object classes.

3.5. Enabling Semantic Recognition

The d masks generated by SAM for each image contains semantic maps at all granularity and also includes background. Our objective is to identify mask(s) containing the desired object(s) specified by the user.

We introduce a classifier head, denoted as θ, to process SAM's flattened mask embeddings $m \in \mathbb{R}^{d\times 1024}$. The classifier head θ is a simple 4-layer MLP network. Its output $o_i = \theta(m_i)$, where $o_i \in \mathbb{R}^{d\times C}$, corresponds to the assigned object labels. In training θ with synthetic single-object data, each image has a 1-dimensional object label within a d-dimensional input space (d mask embeddings per image). This setup necessitates a weakly supervised learning approach, leading us to adopt multiple-instance learning.

Multiple Instance Learning (MIL): In MIL, the d samples (SAM's mask embeddings m_i for the i^{th} image) are grouped into 'positive' and 'negative' bags. 'Positive' bags contain at least one positive instance (a mask embedding $m_i^j \in \mathbb{R}^{1\times 1024}$ corresponding to the ground truth object label y_i), while 'negative' bags have no positive instances. The objective is to train a model that can classify the entire bag and distinguish each instance within it.

To calculate the loss for each bag, denoted as the classifier output 'o_i' in our problem, we aim to express each d-dimensional mask embedding through a single confidence score for each category. For a set of mask embeddings associated with a given image, the activation score for a specific category is computed as the average of the k-max activations across the number of masks (d) for that category. In our case, the dimension $d = 100$ remains constant for all inputs, and we set $k = \max\left(1, \lceil \frac{d}{a} \rceil\right)$, where a is a design parameter. Thus, our class-wise confidence scores for the j^{th} class of the i^{th} input can be represented as,

$$s_i^j = \frac{1}{k} \max_{\substack{\mathcal{O}\subset o_i[:,j] \\ |\mathcal{O}|=k}} \sum_{l=1}^{k} \mathcal{O}_l. \tag{1}$$

Subsequently, a softmax non-linearity is employed to derive the probability mass function across all categories as, $p_i^j = \frac{\exp(s_i^j)}{\sum_{j=1}^{C}\exp(s_j^i)}$. It is necessary to compare this probability mass function with the actual distribution of labels for each image to calculate the MIL loss. Since each image (during test-time) can have multiple objects in it, we represent the labels as a multi-hot vector, where 1 occurs if the object is present, else 0. We then normalize the ground truth label vector transforming it into a valid pmf. The MIL loss is subsequently computed as the cross-entropy between the predicted pmf, denoted as $\mathbf{p_i}$, and the normalized ground truth ($y_i = \left[y_i^1, \ldots, y_i^C\right]^T$) for N images and C object categories as,

$$\mathcal{L}_{MIL} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} -y_i^j \log(p_i^j). \tag{2}$$

Uncertainty Distillation: To further improve the discriminative ability of the object classifier, we employ an Uncertainty Distillation setup by leveraging the uncertainty information during training. The MIL trained classifier from the previous stage acts as the teacher model θ_t (with parameters frozen), and the student model θ_s is an untrained classifier having the same architecture as θ_t. The primary motivation behind uncertainty distillation lies in enhancing the model's ability to make well-calibrated predictions and acknowledge its own uncertainties in challenging scenarios.

For the mask embeddings of the i^{th} image, m_i, we compute the teacher logits as $o_i^t = \theta_t(m_i)$ and the teacher predictions as $\hat{y}_i^t = argmax(p_i^t)$, where $p_i^t = softmax(o_i^t) \in \mathbb{R}^{d\times C}$ are the object-label prediction probability scores for all the d masks. We calculate the entropy (uncertainty) in the teacher logits as $H_i^t = -\sum_{j=1}^{C} p_{ij}^t \cdot \log(p_{ij}^t)$. The goal is to minimize the reliance on (1) high entropy predictions, denoted by $H_i^{t,high}$, which are obtained by thresholding H_i^t (such that, $H_i^{t,high} = H_i^t > threshold$ and $H_i^{t,low} = H_i^t < threshold$), and (2) low entropy ***incorrect*** predictions, denoted as $H_i^{t,low,incorr} = H_i^{t,low} \cap \{\hat{y}_i^t \neq y_i\}$, where y_i is the ground truth object label. Thus, we obtain a set of bad predictions as $B_i = H_i^{t,high} \cup H_i^{t,low,incorr}$. Here, B_i represents the indices of the masks out of all the d masks predicted by SAM for the i^{th} image, where the classifier head predicts a label with highly uncertainty, or predicts a wrong label confidently.

Next, we obtain the student logits as $o_i^s = \theta_s(m_i)$ and compute the MIL loss, $\mathcal{L}_{MIL}$, as per the previous section. With the obtained bad teacher prediction indices, we compute the bad student logits as $o_i^{s,bad} = o_i^s[B_i, :]$. Finally, we compute the uncertainty loss (to be minimized) as,

$$\mathcal{L}_{uncertainty} = \frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} \log\left(\frac{\exp(o_{ij}^{s,bad})}{\sum_{j=1}^{C}\exp(o_{ij}^{s,bad})}\right). \tag{3}$$

Overall Objective: The net loss used to guide the student model is a weighted sum of the MIL and uncertainty losses as, $\mathcal{L}_s = \lambda_1 \cdot \mathcal{L}_{MIL} + \lambda_2 \cdot \mathcal{L}_{uncertainty}$. The trained student model θ_s is used for inference on the test data.

3.6. Inference

For the images in the test set, we obtain the mask embeddings m_i from SAM in the same way as during training on synthetic / web crawled images. These embeddings are passed through our trained classifier head θ_s to acquire object-class confidence scores for each mask within an image. By applying a predefined threshold to the confidence scores, we selectively retain masks whose object labels have been predicted with high probability. Among the selected subset masks for each image, certain masks may be redundant. We use Non-Maximal Suppression (NMS) to eliminate those overlapping masks and ensure the retention of masks only corresponding to distinct objects.

4. Experiments

SAM's automatic mask generation mode uniformly places a grid of d points across the image, and for `U-SAM`, we set $d = 100$. We extracted and saved all d predicted masks from SAM, with the mask embeddings ($m_i \in \mathbb{R}^{d\times 1024}$). This one-pass extraction process ensures efficient storage for subsequent use in our workflow. We generated 200 synthetic single object images per class using the StableDiffuson-v1-4 [36] model. We first train the classifier head (teacher) θ_t, with the MIL loss (L_{MIL}) as the sole objective with batch size 64, learning rate 0.001 optimized with the Adam optimizer. With these same hyperparameters, we train the student model θ_s with the objective $\mathcal{L}_s = \lambda_1 \cdot \mathcal{L}_{MIL} + \lambda_2 \cdot \mathcal{L}_{uncertainty}$, where we set $\lambda_1 = 1$ and $\lambda_2 = 0.15$. Our quantitative evaluation of `U-SAM` relies on two metrics: mean Intersection over Union (mIoU) and mean Average Precision at 50% IoU threshold (mAP_{50}). For our entire workflow, we utilized a single NVIDIA GeForce RTX 3090 GPU. We perform our experiments on two popular benchmarks: PASCAL VOC [11] and the MSCOCO-80 [29] datasets.

Quantitative Results: `U-SAM` is trained on synthetic/web crawled data to recognize object semantics and applied for inference on the *val* sets of PASCAL VOC and MSCOCO-80 datasets. While unsupervised semantic segmentation methods are a common choice for annotation without manual supervision, state-of-the-art (SOTA) unsupervised methods typically rely on large-scale unlabeled datasets from the same distribution as the test set [8, 51, 59]. Given our problem setup that we lack access to such large-scale unlabeled data from the same distribution, a fair comparison with existing SOTA unsupervised methods is challenging. To establish a fair baseline for comparison, we setup two unsupervised methods, Leopart [59] and TransFGU

Table 1. **SOTA Comparison on PASCAL VOC** *val* data. U-SAM significantly outperforms state-of-the-art methods like TransFGU [51] and Leopart [59] (unsupervised) and ACR [23] (weakly supervised) when these methods are trained on synthetic or web crawled images. This underscores U-SAM's robustness in scenarios lacking in-distribution training data. *US* and *WSS* denote Unsupervised and Weakly-Supervised Segmentation, respectively.

Method	Training Data		mIoU	mAP_{50}
	Synthetic	Web Crawl		
Leopart [59] *(US)*	✓		7.21%	-
TransFGU [51] *(US)*	✓		2.05%	-
ACR [23] *(WSS)*	✓		3.49%	-
U-SAM	✓		**25.16%** (+17.95%)	**49.06%**
Leopart [59] *(US)*		✓	6.79%	-
TransFGU [51] *(US)*		✓	2.45%	-
ACR [23] *(WSS)*		✓	13.42%	-
U-SAM		✓	**22.42%** (+9.00%)	**45.59%**

Table 2. **SOTA Comparison on MSCOCO-80** *val* data. We observe that U-SAM significantly outperforms state-of-the-art methods like TransFGU [51] and Leopart [59] (unsupervised) and ACR [23] (weakly supervised) when trained on synthetic or web crawled images. This highlights the reliance of these methods on having the train and test data from the same distribution. Notably, they struggle when the train and test distributions mismatch, even with identical class spaces. *US* and *WSS* denote Unsupervised and Weakly-Supervised Segmentation, respectively.

Method	Training Data		mIoU	mAP_{50}
	Synthetic	Web Crawl		
Leopart [59] (US)	✓		3.84%	-
TransFGU [51] (US)	✓		0.95%	-
ACR [23] (WSS)	✓		0.61%	-
U-SAM	✓		**8.60%** (+4.76%)	**30.90%**
Leopart [59] (US)		✓	3.81%	-
TransFGU [51] (US)		✓	1.02%	-
ACR [23] (WSS)		✓	2.10%	-
U-SAM		✓	**9.01%** (+5.20%)	**31.80%**

[51], and one weakly supervised method, ACR [23] (which achieved high results on in-domain evaluations), trained on our synthetic/web crawled data and evaluated on the PASCAL VOC and MSCOCO-80 datasets in Tables 1 and 2. *This demonstrates that these existing methods do not work well in unseen domains and require access to data from the same distribution as the test set in order to train or adapt their models.* U-SAM does not have this limitation, and thus could be applied across a large variety of applications.

Tables 1 and 2 present performance comparisons between U-SAM and SOTA trained on synthetic or web crawled data and evaluated on the PASCAL VOC and MSCOCO-80 datasets, respectively. On PASCAL VOC, U-SAM demonstrates competitive performance compared to unsupervised methods like Leopart [59] and TransFGU [51], as well as the weakly supervised method ACR [23], when these methods are trained on synthetic/web crawled data. This highlights that these SOTA methods are heavily reliant on the test data belonging to the same distribution as the training data, a requirement that U-SAM does not have, showcasing its robustness across diverse datasets.

Table 3. Comparison of U-SAM performance with the SAM+CLIP [34] method on the PASCAL VOC and MSCOCO-80 datasets. *SD*: Training set is Stable Diffusion generated synthetic images; *WC*: Training set is web crawled images.

PASCAL VOC			MSCOCO-80		
Method	**mIoU**	**mAP_{50}**	**Method**	**mIoU**	**mAP_{50}**
SAM+CLIP	19.17%	42.45%	SAM+CLIP	7.91%	29.93%
U-SAM *(SD)*	**25.16%**	**49.06%**	U-SAM *(SD)*	8.60%	30.90%
U-SAM *(WC)*	22.42%	45.59%	**U-SAM *(WC)***	**9.01%**	**31.80%**

On the MSCOCO-80 dataset, which features diverse object classes and complex scenes, all methods face significant challenges. Despite these challenges, U-SAM maintains a competitive edge when compared to Leopart [59], TransFGU [51], and ACR [23] trained on synthetic/web crawled data. This underscores that these methods are heavily dependent on the alignment between the train and test data distributions, a limitation that U-SAM avoids. The results on MSCOCO-80, as shown in Table 2, reveal a trade-off between localization precision and segmentation accuracy. A high mAP_{50} score indicates proficiency in object detection and localization, while a low mIoU score suggests difficulties in precisely delineating object boundaries or capturing fine-grained details.

Discussion: It is essential to highlight a key aspect of our methodology: our training data originates from a small set of synthetic (200 images per class) or web crawled data ($\sim$ 45 images per class) which are single-object images, and thus distinctly different from the test data distribution. Unlike existing methods, which often rely on detailed annotations for each image, *our method only requires knowledge of the user-defined class names to be segmented across the entire dataset*. The ability to apply an existing method to a new dataset, as achieved by U-SAM, without any knowledge of the distribution of the data in this dataset is a highly desired feature as it enhances generalizability well beyond what unsupervised domain adaptation can do.

Furthermore, the effectiveness of our method is evident in its ability to yield competitive results regardless of the data source, be it synthetic or web crawled images. When TransFGU [51], Leopart [59] and ACR [23] are trained on our synthetic or web crawled datasets, the performance drops significantly compared to their performance on in-distribution training data in both PASCAL VOC and MSCOCO-80 datasets, emphasizing the unique challenges posed by our problem setting. On the other hand, U-SAM achieves +17.95% and +5.17% mIoU, compared to Leopart [59]. This flexibility highlights the robustness of our approach, enabling applications in scenarios where extensive manual annotations are difficult or impractical to obtain.

Additional baseline: We incorporated the CLIP [34] model, renowned for its zero-shot performance, as an ad-

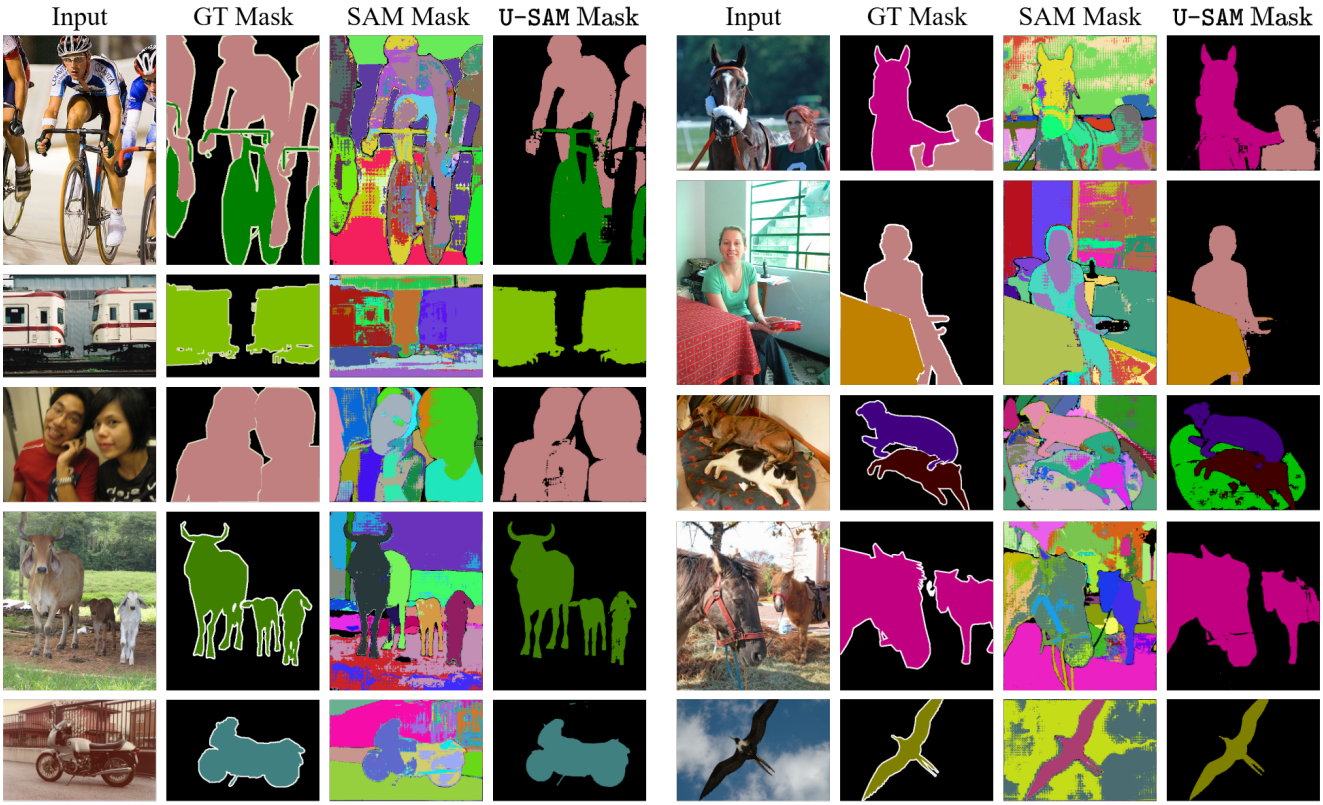


Figure 3. **Qualitative results:** The original image, the ground truth (GT) mask, all of the SAM generated masks overlayed on top of each other, and U-SAM predicted masks are shown respectively in the four columns. For the GT mask and U-SAM predicted masks, the colors indicate the class labels, while random colors were used for the SAM masks since it does not provide class-labels.

ditional baseline to compare U-SAM's performance. The baseline setup for CLIP involves computing the similarity between images and a set of natural language words/phrases (object labels in our case), leveraging its shared Vision-Language embedding space. Utilizing the d masks predicted by SAM for each image, we generate d pooled images, by multiplying the masks to the original image, with each corresponding to a distinct object category. Subsequently, we calculate the similarity between all d pooled images and each of the C user-defined object classes. Similar to our U-SAM post-processing, we set a similarity threshold at 70% and apply Non-Maximal Suppression to retain only unique objects while eliminating redundant masks. The results obtained by this SAM+CLIP [34] method compared to our U-SAM model are reported in Table 3. Unlike SAM+CLIP [34], which lacks granularity matching and relies on CLIP's training data, U-SAM uses web-crawled data, making it more versatile across diverse datasets like PASCAL VOC [11] and MSCOCO-80 [29].

Qualitative Results: In Figure 3, we showcase qualitative results obtained by U-SAM. In each example, SAM generates multiple masks at various granularities, including background masks. For instance, in the first example featuring a person on a bicycle, SAM produces separate masks for the bike helmet, the person's arms, and legs. In contrast, U-SAM provides masks only for the objects requested by the user, demonstrating targeted segmentation.

The second example illustrates two disconnected objects of the same class, while the third example shows two objects of the same class with a connected contour. In all cases, U-SAM successfully filters out the desired mask with the correct object label. Unlike SAM, which segments numerous parts of the image—both small and large objects—U-SAM focuses on the user-specified classes. For example, in the third example (left set), SAM segments the eyes and face separately, but U-SAM outputs only the person mask, as it aligns with the user-defined class. This demonstrates that U-SAM achieves the desired granularity level.

Furthermore, in the third example of the right set in Figure 3, the ground truth mask lacks annotation for the "sofa" object. However, U-SAM correctly identifies and outputs the sofa mask, as it falls within the user-requested object class space. This highlights U-SAM's ability to provide accurate and relevant segmentation masks, which might have even been missed by the ground truth masks.

Changed Granularity: We modified the class definitions for the PASCAL VOC dataset to adjust the granularity

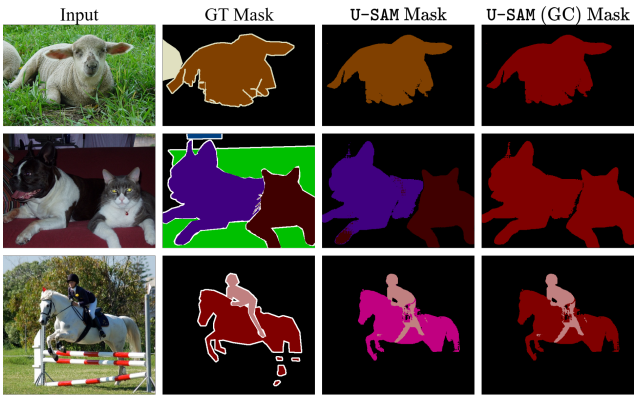


Figure 4. **Qualitative results with changed granularity:** The third column shows the U-SAM predictions on PASCAL classes, and the fourth column shows the predictions when the granularity level has been changed to categorize "dog", "cat" and "sheep" classes as a single "animals" class. The colors represent the different class labels: Brown: "sheep", Violet: "dog", Dark Brown: "cat", Red: "animals", Pink: "person". *GC*: Granularity Changed.

of user-defined classes. Specifically, we introduced some superset classes that encompass several PASCAL VOC classes together. The revised class definitions and the corresponding PASCAL VOC classes falling into these categories are as follows:

1. animals: "*bird*", "*cat*", "*cow*", "*dog*", "*horse*", "*sheep*"
2. furniture: "*chair*", "*diningtable*", "*sofa*"
3. household items: "*bottle*", "*pottedplant*", "*tvmonitor*"
4. person: "*person*"
5. transportation: "*aeroplane*", "*bicycle*", "*boat*", "*bus*", "*car*", "*motorbike*", "*train*"

We applied the U-SAM model to the PASCAL VOC dataset with the changed granularities. For each of the five new classes, we generate 200 synthetic images using the Stable Diffusion model, extract SAM mask embeddings m_i for each image, and train the classifier head θ. This updated granularity configuration yielded an mAP_{50} score of 61.54%. Some of the visual results obtained are shown in Figure 4. U-SAM has been able to detect these classes in both granularity levels- one where the typical PASCAL VOC classes were used, and when the granularity level was changed by altering the definitions of the desired classes.

Ablation Study: In our ablation study (Table 4), we investigated the impact of introducing uncertainty distillation alongside the MIL loss in our classifier head (θ) training framework. The primary objective was to understand the performance implications of leveraging uncertainty information during training, particularly in challenging cases.

Our experiments yielded compelling results, showcasing significant performance improvements, particularly in the PASCAL VOC dataset (with a $+5\%$ increase in mIoU and $+6.59\%$ in mAP_{50}) when leveraging web crawled data for training. The strategic incorporation of Uncertainty Distillation alongside the MIL loss played a pivotal role in enhancing the model's discriminative capabilities. Notably, this improvement is pronounced in the case of MSCOCO-80, where the utilization of web crawled training data addresses the challenges posed by the dataset's diversity and complexity. The model's adept handling of prediction uncertainties proves crucial in navigating the intricacies of real-world scenarios presented by the MSCOCO-80 dataset.

Table 4. **Ablation study results:** Uncertainty Distillation significantly improves performance in both synthetic and web crawled data training paradigms for the PASCAL VOC and MSCOCO-80 datasets. MIL: Multiple Instance Learning.

Method	Training Data		Test Data	mIoU	mAP
	Synthetic	Web crawled			
MIL only	✓		PASCAL VOC	22.42%	42.47%
MIL + Distillation	✓			**25.16%** (+2.53%)	**49.06%** (+6.56%)
MIL only		✓		17.42%	39.86%
MIL + Distillation		✓		**22.42%** (+5.00%)	**45.59%** (+5.73%)
MIL only	✓		MSCOCO-80	8.01%	30.89%
MIL + Distillation	✓			**8.60%** (+0.59%)	**30.97%** (+0.08%)
MIL only		✓		8.57%	27.57%
MIL + Distillation		✓		**9.01%** (+0.44%)	**31.81%** (+4.24%)

A key observation is the model's improved robustness in scenarios where traditional methods might struggle, showcasing the practical relevance of incorporating Uncertainty Distillation. The ability to make well-calibrated predictions and acknowledge uncertainties translated into a more effective and reliable segmentation framework.

5. Conclusion

U-SAM represents a significant advancement in semantic segmentation by transforming the SAM model into a semantically aware framework. SAM, while powerful in generating precise object masks, lacked the ability to recognize and label specific object categories. Our approach addresses this limitation by integrating a classifier head into SAM, trained using synthetic or web-crawled images based solely on user-provided object class names. This enhancement enables U-SAM to generate targeted semantic masks for user-defined object categories without relying on in-domain training data or image-level class labels. Instead, our pipeline requires only dataset-level class names, making it highly efficient and flexible for real-world applications. By automating pixel-level annotations and eliminating the need for manual labeling, U-SAM offers a reliable solution for semantic segmentation, capable of generalizing across diverse datasets without retraining or fine-tuning. This paves the way for more accessible and adaptable semantic segmentation tools, aligning with the evolving needs of users who require precise object identification without the burden of extensive data preparation.

Acknowledgments: This research is partially supported by US National Science Foundation grant CMMI-2326309.

References

[1] Amit Aflalo, Shai Bagon, Tamar Kashti, and Yonina Eldar. Deepcut: Unsupervised segmentation using graph neural networks clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–41, 2023. 2

[2] Hritam Basak, Rohit Kundu, and Ram Sarkar. Mfsnet: A multi focus segmentation network for skin lesion segmentation. *Pattern Recognition*, 128:108673, 2022. 1

[3] Hritam Basak, Soumitri Chattopadhyay, Rohit Kundu, Sayan Nag, and Rammohan Mallipeddi. Ideal: Improved dense local contrastive learning for semi-supervised medical image segmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[4] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11700–11709, 2019. 2

[5] Steve Branson, Grant Van Horn, and Pietro Perona. Lean crowdsourcing: Combining humans and machines in an online system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7474–7483, 2017. 2

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 1

[7] Chen Chen, Chenyu Wang, Bin Liu, Ci He, Li Cong, and Shaohua Wan. Edge intelligence empowered vehicle detection and image segmentation for autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 24(11): 13023–13034, 2023. 1

[8] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16794–16804, 2021. 2, 5

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[10] Arindam Dutta, Rohit Lal, Dripta S Raychaudhuri, Calvin-Khang Ta, and Amit K Roy-Chowdhury. Poise: Pose guided human silhouette extraction under occlusions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6153–6163, 2024. 1

[11] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. 2, 5, 7

[12] Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 2

[13] Liang Gong, Zhiyu Yang, Yihang Yao, Binhao Chen, Wenjie Wang, Xiaofeng Du, Yidong He, and Chengliang Liu. An integrated in situ image acquisition and annotation scheme for instance segmentation models in open scenes with a human–robot interaction approach. *IEEE Transactions on Human-Machine Systems*, 2023. 1

[14] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, 35:1140–1156, 2022. 2

[15] Xiaoqing Guo, Chen Yang, Baopu Li, and Yixuan Yuan. Metacorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3927–3936, 2021. 2

[16] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *Advances in Neural Information Processing Systems*, 36:30726–30737, 2023. 2

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1

[18] Wenbin He, Suphanut Jamonnak, Liang Gou, and Liu Ren. Clip-s4: Language-guided self-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11207–11216, 2023. 2

[19] Wei Huang, Chang Chen, Yong Li, Jiacheng Li, Cheng Li, Fenglong Song, Youliang Yan, and Zhiwei Xiong. Style projected clustering for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3061–3071, 2023. 2

[20] Tsung-Wei Ke, Jyh-Jing Hwang, Yunhui Guo, Xudong Wang, and Stella X Yu. Unsupervised hierarchical semantic segmentation with multiview cosegmentation and clustering transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2571–2581, 2022. 2

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 2, 3, 4

[22] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 2

[23] Hyeokjun Kweon, Sung-Hoon Yoon, and Kuk-Jin Yoon. Weakly supervised semantic segmentation via adversarial

learning of classifier and reconstructor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11329–11339, 2023. 6

[24] Suhyeon Lee, Junhyuk Hyun, Hongje Seong, and Euntai Kim. Unsupervised domain adaptation for semantic segmentation by content transfer. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 8306–8315, 2021. 2

[25] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023. 2

[26] Kehan Li, Zhennan Wang, Zesen Cheng, Runyi Yu, Yian Zhao, Guoli Song, Chang Liu, Li Yuan, and Jie Chen. Acseg: Adaptive conceptualization for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7162–7172, 2023. 2

[27] Xin Li, Feng Xu, Fan Liu, Yao Tong, Xin Lyu, and Jun Zhou. Semantic segmentation of remote sensing images by interactive representation refinement and geometric prior-guided inference. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 1

[28] Yuan-Hong Liao, Amlan Kar, and Sanja Fidler. Towards good practices for efficiently annotating large-scale image classification datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4350–4359, 2021. 2

[29] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 5, 7

[30] Huan Ling, Jun Gao, Amlan Kar, Wenzheng Chen, and Sanja Fidler. Fast interactive object annotation with curve-gcn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5257–5266, 2019. 2

[31] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[32] Luca Medeiros. lang-segment-anything. https://github.com/luca-medeiros/lang-segment-anything, 2023. 3

[33] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8364–8375, 2022. 2

[34] 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. 3, 6, 7

[35] IDEA Research. Grounded segment anything. https://github.com/IDEA-Research/Grounded-Segment-Anything, 2023. 3

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4, 5

[37] Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2023. 2

[38] Linus Scheibenreif, Joëlle Hanna, Michael Mommert, and Damian Borth. Self-supervised vision transformers for land-cover segmentation and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1422–1431, 2022. 2

[39] Hyun Seok Seong, WonJun Moon, SuBeen Lee, and Jae-Pil Heo. Leveraging hidden positives for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19540–19549, 2023. 2

[40] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 2

[41] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3554–3563, 2024. 2

[42] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10052–10062, 2021. 2

[43] 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. 3

[44] Antonin Vobecky, David Hurych, Oriane Siméoni, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, and Josef Sivic. Drive&segment: Unsupervised semantic segmentation of urban scenes via cross-modal distillation. In *European Conference on Computer Vision*, pages 478–495. Springer, 2022. 2

[45] Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *IEEE transactions on pattern analysis and machine intelligence*, 45(12):15790–15801, 2023. 2

[46] Jining Yan, Jingwei Liu, Dong Liang, Yi Wang, Jun Li, and Lizhe Wang. Semantic segmentation of land cover in urban areas by fusing multi-source satellite image time series. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 1

[47] Qingqing Yan, Shu Li, Chengju Liu, Ming Liu, and Qijun Chen. Roboseg: Real-time semantic segmentation on

computationally constrained robots. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(3):1567–1577, 2020. 1

[48] Haosen Yang, Chuofan Ma, Bin Wen, Yi Jiang, Zehuan Yuan, and Xiatian Zhu. Recognize any regions. *arXiv preprint arXiv:2311.01373*, 2023. 3

[49] Xiaobo Yang and Xiaojin Gong. Foundation model assisted weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 523–532, 2024. 2

[50] Shanliang Yao, Runwei Guan, Xiaoyu Huang, Zhuoxiao Li, Xiangyu Sha, Yong Yue, Eng Gee Lim, Hyungjoon Seo, Ka Lok Man, Xiaohui Zhu, et al. Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review. *IEEE Transactions on Intelligent Vehicles*, 2023. 1

[51] Zhaoyuan Yin, Pichao Wang, Fan Wang, Xianzhe Xu, Hanling Zhang, Hao Li, and Rong Jin. Transfgu: a top-down approach to fine-grained unsupervised semantic segmentation. In *European conference on computer vision*, pages 73–89. Springer, 2022. 5, 6

[52] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 1

[53] Qihang Yu, Huiyu Wang, Dahun Kim, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Cmt-deeplab: Clustering mask transformers for panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2560–2570, 2022. 2

[54] Ziqi Yu, Botao Zhao, Yipin Zhang, Shengjie Zhang, Xiang Chen, Haibo Yang, Tingying Peng, and Xiao-Yong Zhang. Cross-grained contrastive representation for unsupervised lesion segmentation in medical images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2347–2354, 2023. 2

[55] Andrii Zadaianchuk, Matthaeus Kleindessner, Yi Zhu, Francesco Locatello, and Thomas Brox. Unsupervised semantic segmentation with self-supervised object-centric representations. *arXiv preprint arXiv:2207.05027*, 2022. 2

[56] Daoan Zhang, Chenming Li, Haoquan Li, Wenjian Huang, Lingyun Huang, and Jianguo Zhang. Rethinking alignment and uniformity in unsupervised image semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11192–11200, 2023. 2

[57] Jielu Zhang, Zhongliang Zhou, Gengchen Mai, Mengxuan Hu, Zihan Guan, Sheng Li, and Lan Mu. Text2seg: Remote sensing image semantic segmentation via text-guided visual foundation models. *arXiv preprint arXiv:2304.10597*, 2023. 1

[58] Xingchen Zhao, Niluthpol Chowdhury Mithun, Abhinav Rajvanshi, Han-Pang Chiu, and Supun Samarasekera. Unsupervised domain adaptation for semantic segmentation with pseudo label self-refinement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2399–2409, 2024. 2

[59] Adrian Ziegler and Yuki M Asano. Self-supervised learning of object parts for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14502–14511, 2022. 5, 6