Rethinking Evaluation Metrics of Open-Vocabulary Segmentation

Hao Zhou, Lu Qi, Tiancheng Shen, Hai Huang, Xu Yang, *Senior Member, IEEE*, Xiangtai Li, and Ming-Hsuan Yang, *Fellow, IEEE*

***Abstract*—This paper highlights a problem of evaluation metrics adopted in the open-vocabulary segmentation. The evaluation process relies heavily on closed-set metrics on zero-shot or cross-dataset pipelines without considering the similarity between predicted and ground truth categories. We first survey eleven similarity measurements between two categorical words using WordNet linguistics statistics, text embedding, or language models by comprehensive quantitative analysis and user study to tackle this issue. Based on those explored measurements, we design novel evaluation metrics, Open mIoU, Open AP, and Open PQ, tailored for three open-vocabulary segmentation tasks. We benchmark the proposed evaluation metrics on twelve open-vocabulary methods in three segmentation tasks. Despite the relative subjectivity of similarity distance, we demonstrate that our metrics can still well evaluate the open ability of the existing open-vocabulary segmentation methods. We hope our work can bring the community new thinking about evaluating model ability for open-vocabulary segmentation.**

***Index Terms*—Open-vocabulary segmentation, evaluation metrics, similarity measurements.**

Received 24 April 2024; revised 3 January 2025; accepted 15 April 2025. Date of publication 21 April 2025; date of current version 3 July 2025. This work was supported in part by the National Natural Science Foundation of China under Grant U21A20490, in part by Youth Innovation Promotion Association CAS under Grant 2022133, and in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) under Grant RS-2024-00457882, National AI Research Lab Project. Recommended for acceptance by J. Cai. *(Hao Zhou and Lu Qi contributed equally to this work.) (Corresponding authors: Hai Huang; Xu Yang.)*

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

Hao Zhou was with the Harbin Engineering University and the Institute of Automation, Chinese Academy of Sciences, Beijing 100045, China. He is now with the School of Computing and Information Technology, Great Bay Institute for Advanced Study/Great Bay University, Dongguan 523000, China, and also with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Beijing 100190, China.

Lu Qi is with Insta360 Research, Shenzhen 515100, China (e-mail: qqlu1992@gmail.com).

Tiancheng Shen and Ming-Hsuan Yang are with the University of California Merced, Merced, CA 95343 USA.

Hai Huang is with the National Key Laboratory of Autonomous Marine Vehicle Technology, Harbin Engineering University, Harbin 150000, China (e-mail: huanghai@hrbeu.edu.cn).

Xu Yang is with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China (e-mail: xu.yang@ia.ac.cn).

Xiangtai Li is with Nanyang Technological University,, Singapore 639798.

The source code will be available at https://github.com/qqlu/Entity.

Digital Object Identifier 10.1109/TPAMI.2025.3562930

I. Introduction

OPEN-VOCABULARY segmentation (OVS) [20], [32], [65], [84] refers to the task of segmenting images into several regions while assigning each region a potentially *unlimited or even open set* of object category. Compared with the typical segmentation task that defines a fixed set of object classes, OVS is more flexible and adaptable to new and unseen object classes, which can benefit various risk-aware real-world applications like autonomous driving [47] and robot-related tasks [12].

Recent advances in large vision language models (VLMs) [34], [51], such as CLIP [51], have signficant impacts tp the OVS field [20], [32], [84]. However, the widely adopted evaluation methodologies, which often employ closed-set metrics on zero-shot or cross-dataset evaluations, fail to capture the performance of OVS models because they do not consider the similarity between predicted and ground truth categories. For example, the true positive (TP) score for both cabinet and chest would be zero if the ground truth is a table, as shown in the first column of Fig. 1. This contradicts human perceptual intuition [2], [61] since we should not repudiate those predictions. This motivates us to rethink how to estimate model ability for open-vocabulary segmentation effectively.

To enable open-world evaluation, we first explore eleven heuristic similarity measurements between two categorical words that can be divided into WordNet methods, text-embedding models, and pre-trained language models. The WordNet [43] is a lexical database that organizes words into a hierarchical structure by linguistician and is adopted by ImageNet [59]. The text embedding models (Word2Vec [42], GloVe [46], and FastText [3]) and pre-trained language models (CLIP [51], BERT [14], GPT-2 [52], and T5 [53]) are learnable models trained in language datasets at word and sentence level respectively. Based on the quantitative analysis and user study (Section IV), we use the Path method of WordNet as it can: 1) generate moderate similarity values. 2) distinguish polysemy. 3) operate unbiasedly to training data. 4) match human perception. However, we note that the Path method is not the only option for this task. Different open-world situations may choose different similarity measurements according to their specific requirements.

Built on the similarity measurements, we introduce Open mIoU, Open AP, and Open PQ to evaluate open-vocabulary semantic, instance, and panoptic segmentation tasks, respectively. Although these three segmentation tasks have distinct

Fig. 1. Segmentation results with error labels evaluated by vanilla and open metrics. Our open metric assigns soft true positive (TP) for entities with error labels and thus provides a more reasonable evaluation for open-vocabulary segmentation.

objectives and evaluation pipelines, our open metrics share two fundamental operations: a class-agnostic matching criterion and a semantic similarity-based scoring computation.

We first use a class-agnostic matching approach, focusing on identifying overlapping entities or pixels. This step differs from conventional evaluation metrics, where matching is confined to the same category's predicted and ground truth regions, regardless of the extent of mask overlap. Subsequently, we utilize the similarity measurement explored above to compute floating-point matching scores instead of the binary true positive (TP), false positive (FP), and false negative (FN) values employed in closed-set segmentation metrics.

Finally, we use the proposed evaluation metrics to benchmark twelve OVS methods of three segmentation tasks in Section V. By comparing these new metrics to vanilla closed-set metrics in Fig. 1, we demonstrate that our novel metrics align more closely with human judgment, providing a more accurate and reasonable assessment of the model's open capability.

The main contributions of this work are:

- We highlight a significant limitation of existing evaluation metrics applied to open-world segmentation. The error label predictions should be assigned soft match scores based on their similarity distance to the ground truth.
- We compare eleven similarity measurements using WordNet linguistics statistics, text embedding, or language models in comprehensive quantitative analysis and user studies.
- We propose innovative open evaluation metrics, including Open mIoU, AP, and PQ. Three mainstream open-vocabulary segmentation (semantic, instance, and panoptic segmentation) and object detection tasks and models are benchmarked with these measures.
- Although the open evaluation metrics are still subjective, they can inspire the community to rethink an effective approach to analyze the model capability for open-vocabulary segmentation.

II. Related Work

A. Closed-Set Segmentation

Closed-set segmentation aims at segmenting images into a predefined set of object classes. The Fully Convolutional Networks (FCNs) [41] is one of the early closed-set segmentation methods based on deep learning. FCNs perform well on closed-set segmentation tasks by leveraging end-to-end trainable deep neural networks. These models learn to predict pixel-wise class labels using convolutional layers and upsampling operations. In

subsequent works, FCNs have been further extended and refined, such as U-Net [58] and DeepLab [6]. Several convolution- [33] or transformer-based [40] works further improve the model's ability in semantic [6], [58], [60], [70], instance [5], [22], [37], [39], [50], and panoptic segmentation [9], [27], [28], [79]. Transformer-based methods primarily focus on enhancing pixel feature extraction for pixel classification.

Although closed-set segmentation methods continue to advance, the predefined label set is ill-posed for addressing the demands of real-world vision applications, such as scene understanding, self-driving, and robotics, where the number of object categories is extensive and subject to change. Our work focuses on enabling the closed-set evaluation metrics to have open-world ability.

B. Open-Vocabulary Segmentation

In [80], Zhao et al. propose an open-vocabulary parsing network to encode the word concept hierarchy based on dictionaries and learn a joint pixel and word feature space. Recently, numerous segmentation methods have been developed based on large vision and language models. Based on CLIP [51] or ALIGN [24] that trained on class-agnostic entity proposals [9], [48], [49], RegionCLIP [81], LSeg [32], MaskCLIP [84], OpenSeg [20], OVSeg [35] and subsequent works [8], [55], [64], [71], [75], [78] utilize text embeddings from CLIP to associate the class-agnostic entity embeddings. Specifically, RegionCLIP [81] matches image areas with corresponding texts by finetuning the CLIP model and LSeg [32] utilizes text embeddings from CLIP as per-pixel classifier. On the other hand, MaskCLIP [84] retains the pre-trained weights frozen and makes minimal adaptations to preserve the visual-language association. OpenSeg [20] employs the ALIGN model and pools global image features with generated class-agnostic masks to get dense/local features. In addition, OVSeg [35] introduces a fine-tuning approach that leverages masked image regions and their corresponding text descriptions to improve the performance of CLIP on masked region classification.

Although these methods consider cross-modal feature representation and alignment, they still rely on closed-set metrics for evaluation. However, existing evaluation metrics still penalize the inconsistency between predicted and ground truth categories even though they are synonyms. Instead, our work aims to design a more reasonable evaluation pipeline considering the similarity between two labels.

C. Open-Vocabulary Object Detection (OVOD)

Considering the close relationship between instance segmentation and object detection, several works [7], [15], [15], [16], [19], [21], [23], [25], [26], [29], [44], [62], [63], [65], [66], [67], [74], [76], [77], [85] explore object detection in the open-vocabulary settings and extend detectors to recognize objects not encountered during the training. Similar to prior segmentation tasks, the existing OVOD approaches still use closed-set metrics (box mAP). Our open metric can also be adapted to OVOD settings.

III. Evaluation Metrics

In this section, we first review three widely-used closed-set evaluation metrics, i.e., mean Intersection over Union (mIoU), Average Precision (AP), and Panoptic Quality (PQ). These metrics have been used in semantic, instance, and panoptic segmentation tasks, with the AP metric also capable of evaluating object detection by replacing mask IoU with box IoU. Next, we introduce effective evaluation metrics for open-vocabulary tasks within these domains.

A. Preliminaries

Despite the various segmentation tasks targeting different levels, such as pixel, object, and entity, we observe that the core of their evaluation metrics is based on IoU at these levels, which measures the precision between predictions and ground truth. We can obtain a confusion matrix by dividing the predicted results into four sets, including True Positives (TP), True Negatives (TN), False Positives (FN), and False Negatives (FN), based on the matching criteria to the ground truth at each category. Here, TP and TN represent the count of accurate positive and negative predictions, respectively. Conversely, FP and FN quantify incorrect positive and negative predictions, respectively. With these definitions in place, Intersection over Union (IoU) can be mathematically formulated by $\mathrm{IoU} = \frac{|TP_p|}{|FP_p|+|TP_p|+|FN_p|}$, where $|TP_p|$, $|FP_p|$, and $|FP_p|$ respectively stand for the number of pixel-level TP, FP, and FN. We calculate the corresponding evaluation metrics given the different segmentation tasks.

For the mIoU metric used in semantic segmentation, the matching criterion is determined by the accurate pixel-level label prediction:

$$\mathrm{mIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\mathrm{IoU}_i, \tag{1}$$

where IoU_i represents the IoU of class i, and $k+1$ is the total number of classes in the evaluated dataset. The AP and PQ metrics used in instance and panoptic segmentation rely on entity-level matches determined by IoU. These matches are evaluated based on whether the IoU between predicted and ground truth entities surpasses a specified threshold δ. Subsequently, AP computes the precision-recall curve using entity-level TP, FP, and FN:

$$\mathrm{AP} = \int_0^1 p(r)\mathrm{d}r, \tag{2}$$

where $p = \frac{|TP_e|}{|TP_e|+|FP_e|}$ and $r = \frac{|TP_e|}{|TP_e|+|FN_e|}$ denotes precision and recall, while $|TP_e|$, $|FP_e|$ and $|FN_e|$ represent the numbers of entity-level TP, FP, and FN respectively. Meanwhile, PQ evaluates the segmentation quality of the model as follows:

$$\mathrm{PQ} = \underbrace{\frac{\sum_{(g_i,d_j)\in TP}\mathrm{IoU}(g_i,d_j)}{|TP_e|}}_{\mathrm{SQ}}$$

$$\times \underbrace{\frac{|TP_e|}{|TP_e| + \frac{1}{2}|FP_e| + \frac{1}{2}|FN_e|}}_{\text{RQ}}, \tag{3}$$

where $\text{IoU}(d_i, g_j)$ represents the IoU between entities d_i and g_j. PQ can be interpreted as the product of Segmentation Quality (SQ) and Recognition Quality (RQ).

B. Open-Vocabulary Evaluation

Existing mIoU, AP, and PQ are ill-equipped to assess OVS tasks as they require strict class matching. We introduce open metrics that integrate semantic similarity between labels to mitigate this limitation. This involves two key steps: first, establishing a method to assess semantic similarity between ground truth and predicted labels; second, using these similarities to calculate softened versions of TP, FP, and FN for metrics suitable in OVS tasks.

1) Similarity Matrix: Two main types of models are used to quantify similarity between words: distributional and knowledge-based models. The latter primarily uses structured knowledge sources, notably WordNet [43], while distributional models derive word embeddings from text embedding or pre-trained language models to compute cosine similarity. Recognizing the advanced capacities of pre-trained language models for sentence or document comprehension, we classify distributional models into two subgroups, resulting in three critical methodologies: 1) WordNet methods, 2) Text embedding models, and 3) Pre-trained language models.

WordNet Methods: WordNet is a lexical database that organizes words into a hierarchical structure using an "IS-A" (hypernym/hyponym) taxonomy, echoing how the human cognitive system stores visual knowledge, proving to be a valuable resource for evaluating semantic relationships between words. Thus, numerous methods [30], [31], [36], [36], [56], [69] have been developed to measure similarity using WordNet. Among these methods, Path [31], Wu-Palmer [69], Lin [36], and Lesk [31] Similarities yield values within the closed interval $[0, 1]$, conducive for computing soft TP, FP, and FN, making them suitable choices.

Text Embedding Models: Text embedding models trained on large text corpora learn to represent words in vector spaces where similar words are closer in the vector space, indicating their semantic similarity. Text embedding models, such as Word2Vec [42], GloVe [46], and FastText [3], are widely used for calculating word similarities using cosine similarity. Note that cosine similarity scores range from -1 to 1, where the interval $[-1, 0)$ signifies negative correlation is redundant for calculating soft TP (FP, FN) scores. Hence, we set cosine similarity values less than 0 to 0.

Pre-trained Language Models: Aside from the word-level embedding ability, pre-trained language models can also comprehend entire sentences or documents that text embedding models cannot, thus extracting better embeddings for compound words. Hence, models such as CLIP [51], BERT [14], GPT-2 [52], and T5 [53] are also well-suited for calculating word similarities.

Solutions to Unknown Words: Pre-trained language models, adept at contextualizing words within sentences, can generate embeddings for unknown words using known words or subwords. Text embedding models, however, provide embeddings only for individual words. For unknown words, we initialize them with random vectors. WordNet contains 82,115 noun synsets and is subject to ongoing updates. When words are missing in WordNet, their hypernyms are used as substitutes, as they largely share the hyponyms' characteristics. This results in minor path length variations (± 1) that negligibly affect evaluation results.

Our Preference: In Section IV, we conduct a comprehensive analysis of the eleven similarity measurements mentioned above and prefer using Path Similarity to calculate the similarity matrix $S \in \mathcal{R}^{(k+1)\times(k+1)}$.

2) Open Evaluation Metrics: The proposed open evaluation metrics convert binary TP, FP, and FN into floating-point scores using two steps: first, shifting from class-specific to class-agnostic matching, and second, using a similarity matrix to calculate soft TP, FP, and FN based on the matching conditions. In the following, we introduce the detailed evaluation process for each segmentation task.

Open mIoU: Semantic segmentation is a pixel-level classification task primarily focusing on category prediction at the pixel level. Conventional TP_{hard}, FP_{hard}, and TN_{hard} for class i are calculated as $TP_{\text{hard}} = n_{ii}$, $FP_{\text{hard}} = \sum_{j=0}^{k} n_{ji} - n_{ii}$, and $FN_{\text{hard}} = \sum_{j=0}^{k} n_{ij} - n_{ii}$, where n_{ij} symbolizes pixels with the ground truth label i classified as label j. However, for open-vocabulary evaluation, misclassified pixels should not be strictly counted as binary TP, FP, or FN. Instead, we suggest the calculation of soft TP, FP, and FN for error label prediction situations by utilizing the semantic similarity matrix S, as shown in Table I. Thus, the soft TP, FP, and FN for class i can be defined as:

$$\begin{aligned}
TP_{\text{soft}} &= n_{ii} + \sum_{j=0}^{k} S_{ij} \cdot n_{ij} - S_{ii} \cdot n_{ii} = \sum_{j=0}^{k} S_{ij} \cdot n_{ij}, \\
FP_{\text{soft}} &= \sum_{j=0}^{k} (1 - S_{ji}) \cdot n_{ji} - (1 - S_{ii}) \\
&\quad \cdot n_{ii} = \sum_{j=0}^{k} (1 - S_{ji}) \cdot n_{ji}, \\
FN_{\text{soft}} &= \sum_{j=0}^{k} (1 - S_{ij}) \cdot n_{ij} - (1 - S_{ii}) \\
&\quad \cdot n_{ii} = \sum_{j=0}^{k} (1 - S_{ij}) \cdot n_{ij},
\end{aligned} \tag{4}$$

where S_{ij} represents the similarity between labels i and j in similarity matrix S. The notations $*_{\text{soft}}$ and $*_{\text{hard}}$ represent the soft and binary versions of $*$, where $*$ can be TP, FP, or FN. Next, we calculate the Open mIoU for open-vocabulary semantic segmentation using the IoU formulation.

TABLE I
HARD AND SOFT TP/FP/FN THAT USED IN VANILLA AND OPEN METRICS DEFINED WITH THE UNIFIED NOTATION FROM TABLE II

Metric	Type	Situation	Hard TP/FP/FN (vanilla)	Soft TP/FP/FN (open)
mIoU	pixel-based	$g_i(x,y) = d_j(x,y)$ and $c_i \neq c_j$	$TP_{c_i} = 0$, $FP_{c_j} = 1, FN_{c_i} = 1$	$TP_{c_i} = S_{c_i c_j}, FP_{c_j} = 1 - S_{c_i c_j}$, $FN_{c_i} = 1 - S_{c_i c_j}$
AP	entity-based	$\text{IoU}(g_i, d_j) \geq \delta$ and $c_i \neq c_j$	$TP_{c_i} = 0, FP_{c_j} = 1$	$TP_{c_i} = S_{c_i c_j}, FP_{c_j} = 1 - S_{c_i c_j}$
PQ	entity-based	$\text{IoU}(g_i, d_j) \geq \delta$ and $c_i \neq c_j$ and $(c_i, c_j \in stuff$ or $c_i, c_j \in thing)$	$TP_{c_i} = 0$, $FP_{c_j} = 1, FN_{c_i} = 1$	$TP_{c_i} = S_{c_i c_j}, FP_{c_j} = 1 - S_{c_i c_j}$, $FN_{c_i} = 1 - S_{c_i c_j}, \text{IoU}(g_i, d_j)* = S_{c_i c_j}$

TABLE II
NOTATION USED IN TABLE I

Notation	Definition
g_i	ground truth entity or pixel
d_j	prediction entity or pixel
$g_i(x,y)$	ground truth pixel's coordinates
$d_j(x,y)$	prediction pixel's coordinates
c_i	label of ground truth g_i
c_j	label of prediction d_j
$\text{IoU}(g_i, d_j)$	IoU between entities g_i and d_j
δ	matching threshold for mask IoU

We use g_i and d_j to simultaneously represent ground truth and prediction of entities or pixels for simplicity.

Open AP: For instance segmentation, the matching is performed between predicted and ground truth entities with the same category label. To modify AP for open-vocabulary evaluation, we first convert the class-aware matching strategy to a class-agnostic manner. Specifically, we convert the labels of all ground truth and predicted entities to the "object" category. Then, we sort the predicted entities in descending order of their prediction scores, creating a class-agnostic global ranking. Subsequently, we match the predicted entities using this global descending order. For each predicted entity, we match it with a ground truth entity that has the largest IoU and has not been assigned to any predictions. Let g_i be a ground truth entity and d_j be a predicted entity that satisfies the matching criteria but with different labels. According to the labels of g_i and d_j, we calculate soft TP and FP using the similarity matrix S, as illustrated in Table I. Given the soft TP and FP, the Open AP can be obtained following (2).

Open PQ: The Typica PQ matches predicted and ground-truth entities using two criteria: an IoU threshold above 0.5 and belonging to the same category. For open-vocabulary evaluation, a modification in PQ's category-matching criteria is necessary. Instead of insisting on identical category labels for two overlapping entities, g_i and d_j, it is sufficient that their labels fall into "stuff" or "thing" classes. Then, we calculate the soft TP, FP, and FN using the semantic similarity between the labels of g_i and d_j, as illustrated in Table I. Finally, the Open PQ value can be obtained using (3). Note that the SQ term in PQ calculates the average IoU over matched segments. However, the match criteria for open-vocabulary panoptic segmentation has been updated to a softer manner considering the semantic similarity between labels. Thus, the $\text{IoU}(g_i, d_j)$ in (3) should be multiplied by the same semantic similarity so that the definition of SQ is unchanged.

IV. ANALYSIS OF SIMILARITY MEASURES

In this section, we analyze eleven similarity measurements based on WordNet (Path [31], Wu-Palmer [69], Lin [36], and Lesk [31]), text embedding (Word2Vec [42], GloVe [46], and FastText [3]), and pre-trained language models (CLIP [51], BERT [14], GPT-2 [52], and T5 [53]). The detailed descriptions of investigated similarity measurements are presented in Section IV-H. To analyze these methods, we use them to generate different similarity matrices for ImageNet-1 k's labels. Although ImageNet-1 k is built upon WordNet, we show that its use does not introduce bias when analyzing WordNet-based similarity methods. Because WordNet serves as a framework for organizing words into a lexical database, it can accommodate any word from the lexicon. In addition, we visualize the fitted probability density distribution curves and heatmaps of these similarity matrices in Figs. 2 and 3. We report the metric mean and standard deviation values in Table III. A recent work [13] shows that the average visual similarity between labels decreases as their hierarchical distance in WordNet increases. Specifically, their results on average visual similarities over hierarchical distances show a maximum value of 0.27. Thus, the range $(0, 0.27]$ can be considered a subjectively appropriate range for the mean of a similarity matrix. Additionally, we present the mean and standard deviation of a human-annotated similarity matrix in Table III for reference. This similarity matrix is derived from a user study on a subset of the ImageNet-1k label set. For further details, please refer to Section IV-F.

A. WordNet Methods

Table III shows that the methods by Lin and Lesk have mean values of 0.015 and 0.048. While these values fall inside the proper mean range $(0, 0.27]$, they are much lower than the mean of the human-annotated similarity matrix, indicating an underestimation of similarity. Fig. 2(b) shows most similarity values of these two methods are distributed from 0 to 0.1, indicating that a significant portion of the similarity values is close or equal to zero. The heatmaps in Fig. 3 show these two methods perform similarly. Conversely, Wu-Palmer exhibits a mean of 0.445, which falls outside the proper range and significantly exceeds the mean of the human-annotated similarity matrix, thereby overestimating the semantic relationship. The heatmap of Wu-Palmer in Fig. 3 shows that a substantial proportion of the

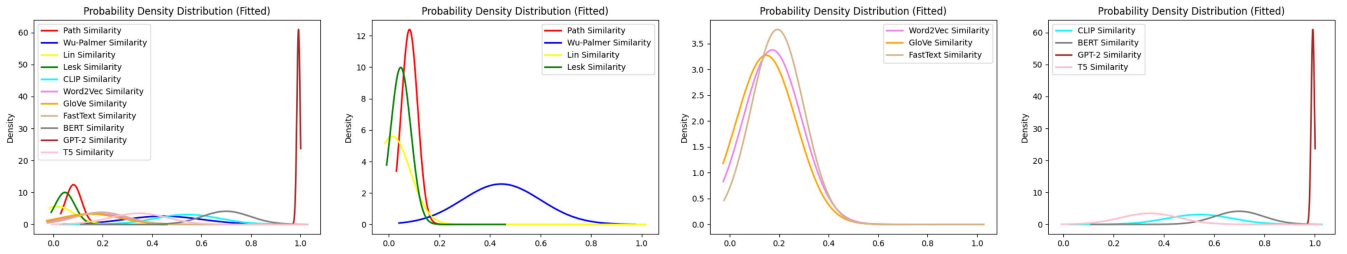


Fig. 2. The fitted probability density distribution curves of similarity matrices for ImageNet-1 k's labels, calculated using various similarity measurements, are presented as follows: (a) displays all the methods under investigation, (b) showcases methods derived from WordNet, (c) presents methods originating from text embedding models, and (d) illustrates methods derived from pre-trained language models.

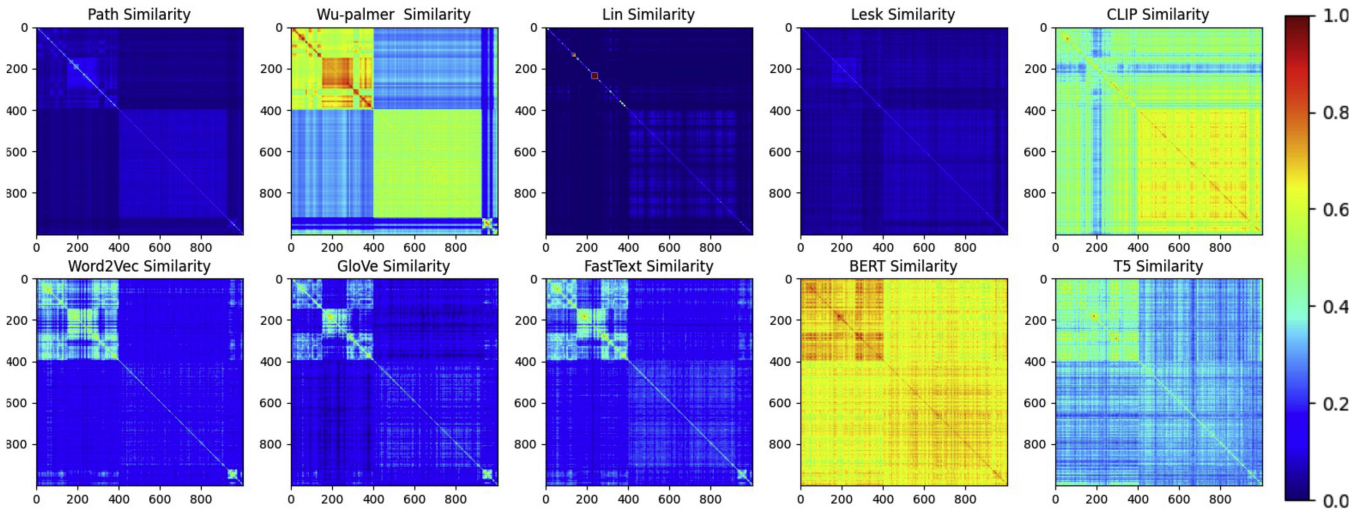


Fig. 3. Heatmaps depicting similarity matrices for ImageNet-1 k's labels, computed using various similarity measurements. Since a significant portion of values in the similarity matrix calculated by GPT-2 is close or equal to one, we do not present its heatmap here.

TABLE III
THE MEAN AND STANDARD DEVIATION OF THE SIMILARITY MATRICES CALCULATED USING VARIOUS SIMILARITY MEASUREMENTS

Method	WordNet method				Text embedding model			Pre-trained language model				Human
	Lin	Lesk	Path	WuP	Glove	WV	FT	T5	CLIP	BERT	GPT-2	
Mean	0.015	0.048	0.083	0.445	0.148	0.173	0.194	0.346	0.543	0.700	0.992	0.109
Std	0.078	0.050	0.043	0.156	0.125	0.121	0.109	0.117	0.130	0.098	0.007	0.101

Herein, 'WV,' 'FT,' and 'WuP' represent abbreviations for the Word2Vec, FastText, and Wu-Palmer methods. 'Human' represents a human-annotated similarity matrix investigated from a user study using a subset of ImageNet-1k's label set.

similarity values exceed 0.5, further supporting this observation. In such cases, our open metrics would yield evaluation results with significant bias compared to these metrics. While the mean of the Path method falls within the proper range $(0, 0.27]$ and closely aligns with the mean of the human-annotated similarity matrix.

B. Text Embedding Models

Figs. 2(c) and 3, along with Table III show that Word2Vec, GloVe, and FastText share similar distributions and mean values. Notably, their mean values fall within the proper range $(0, 0.27]$ and closely align with the mean of the human-annotated similarity matrix, making them suitable choices. However, they have two issues in calculating similarity. First, text embedding models, treating words as context-independent entities, cannot discern different meanings of a word, leading to potentially inaccurate similarity scores. For example, in the label set of the ImageNet-22 k dataset, "crane" denotes both a bird and a machine, "compass" signifies both a navigational instrument and a drafting instrument, and "bank" describes both sloping land and a financial institution, as shown in Fig. 6. However, these models extract the same embedding for polysemy. In contrast, WordNet's base unit is a synset organized by word meanings, which can inherently handle polysemy. Second, the quality of text embeddings heavily relies on training corpus. Inadequate or biased training data can lead to biased embeddings. For example, Fig. 4 shows that text embeddings from GloVe models trained on

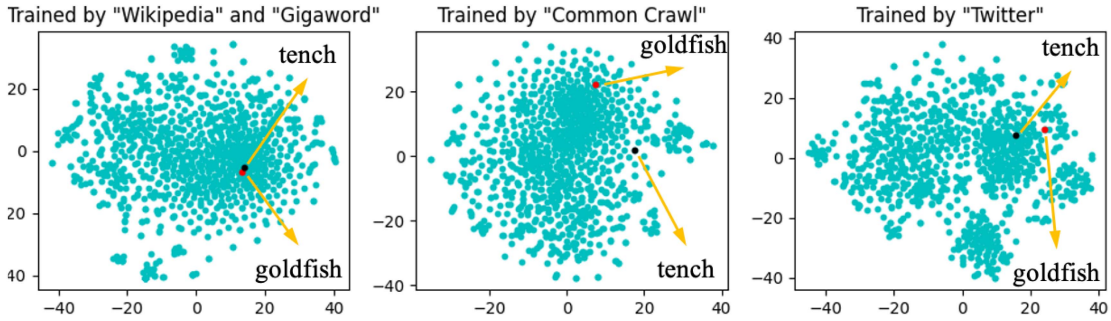


Fig. 4. T-SNE visualization of text embeddings for ImageNet-1 k's labels extracted from GloVe models trained on different datasets. The text embeddings trained by different datasets have varied distributions. The labels "tench" and "goldfish" display varying semantic distances within these distributions.

different datasets exhibit varying distributions. The labels "tench" and "goldfish" exhibit differing semantic distances within these distributions. However, WordNet, developed by linguists and NLP researchers over three decades [18], inherently minimizes biases in word representations due to its structured organization.

C. Pre-Trained Language Models

We use pre-trained language models to extract text embeddings. The two main issues in text embedding models, i.e. cannot handle polysemy and training data dependency, also exist in pre-trained language models. In addition, Figs. 2(d) and 3, and Table III show that the similarity matrices generated by pre-trained language models exhibit mean values that fall outside of the proper range and significantly exceed the mean of the human-annotated similarity matrix, leading to an overestimate of semantic relationship.

D. Case Study of Similarity Measures

In Fig. 5, we visualize three similarity lists of typical labels "sea lion", "ballplayer", and "bee" of ImageNet-1 k calculated by Path Similarity, GloVe model, and T5 model. The similarity values of each word are sorted in descending order. We present three largest, two middle, and three smallest values of each method. In addition, we show pictures of the labels with Top-3 similarity values for each method.

Fig. 5(a) illustrates the phrase "sea lion", composed of the words "sea" and "lion". The GloVe and T5 models assign high similarity values to labels containing "sea" or "lion", despite lacking visual resemblance. In contrast, the Path method prioritizes labels with similar textures and closer biological relationships.

Fig. 5(b) presents the compound word "ballplayer" that can be broken down into tokens "ball", "play", and "er" through morphological tokenization. The GloVe and T5 methods assign higher similarities to words containing these tokens, even with minimal texture similarity. In contrast, the Path method assigns higher scores to labels related to occupations or human statuses. Note that only two human-related labels appear in the Path method's Top-3 similarities because the ImageNet-1 k contains just three human identity-related labels. The third label, "whippet" ranks closely due to its biological proximity to humans.

Fig. 5(c) shows a simple word "bee". The GloVe and T5 methods assign high similarity scores to labels containing tokens such as "bee" or "bea", even with little texture similarity. In contrast, the Path method assigns high scores to labels like "ant", "fly", and "cicada", as they all belong to the class Insecta and share high texture similarities. Among the Top-3 labels, "ant" receives the highest score because both "ant" and "bee" belong to the same order, Hymenoptera.

The analyses above reveal that the embedding models GloVe and T5 compute high similarities for labels containing similar text despite their differing appearances. In contrast, the Path method assigns large values to labels with inherent biological similarity and similar appearances. Furthermore, the GloVe and T5 models generate much higher similarity values than the Path method, which overestimates the semantic relationship between labels.

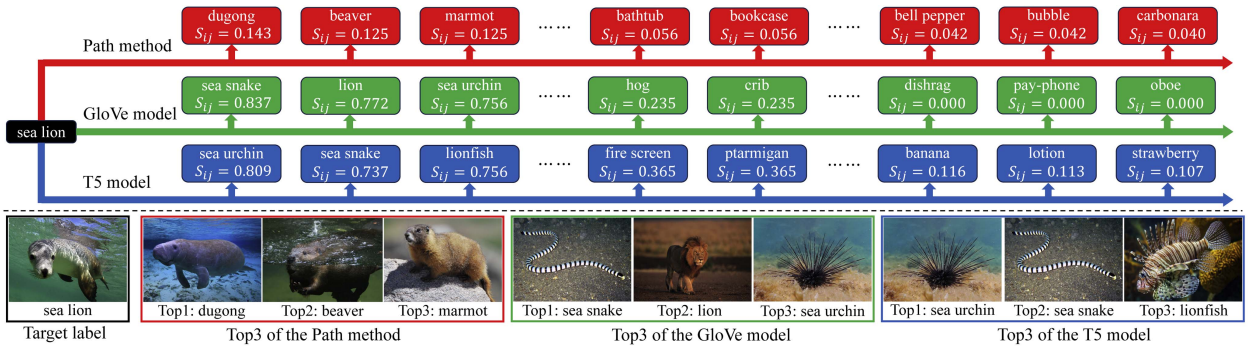
E. User Study on Similarity Measures

A user study involving 1,000 subjects is conducted to investigate the human preference for different similarity measures. In this study, we randomly selected 50 images from the ADE20K [82] dataset and visualized the largest entity in each image. For each visualized entity, we randomly chose one incorrect label and calculated its semantic similarities with the ground truth label using all investigated similarity measurements. Fig. 7 illustrates four samples of this study. We ask each subject to rate their level of satisfaction with the semantic similarities calculated by different similarity measurements for each image. The satisfaction scores ranged from 0 (lowest) to 10 (highest).

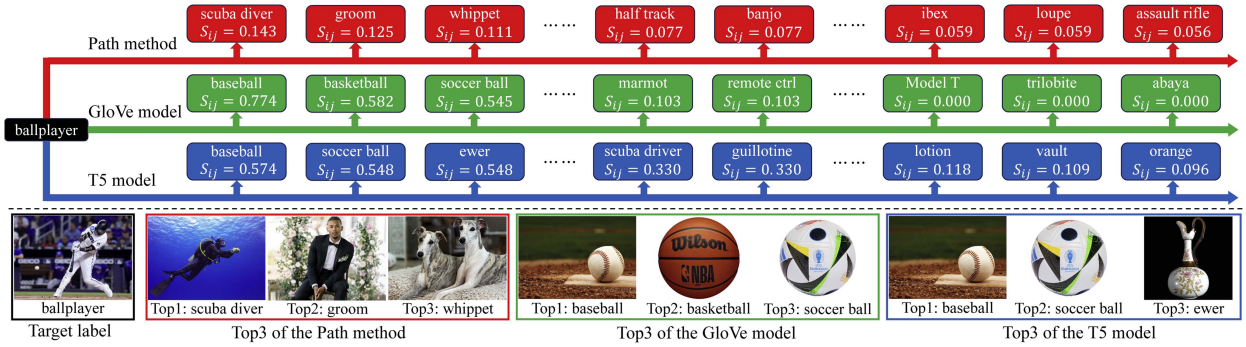
The aggregated satisfaction scores are visualized in Fig. 8. The results indicate that most subjects rate Path Similarity with satisfaction scores higher than 6. The average scores for each similarity measure further validate the superiority of Path Similarity. This user study demonstrates that Path Similarity is the most preferred method among human participants for calculating the semantic similarity between words.

F. Human-Annotated Similarity Matrix

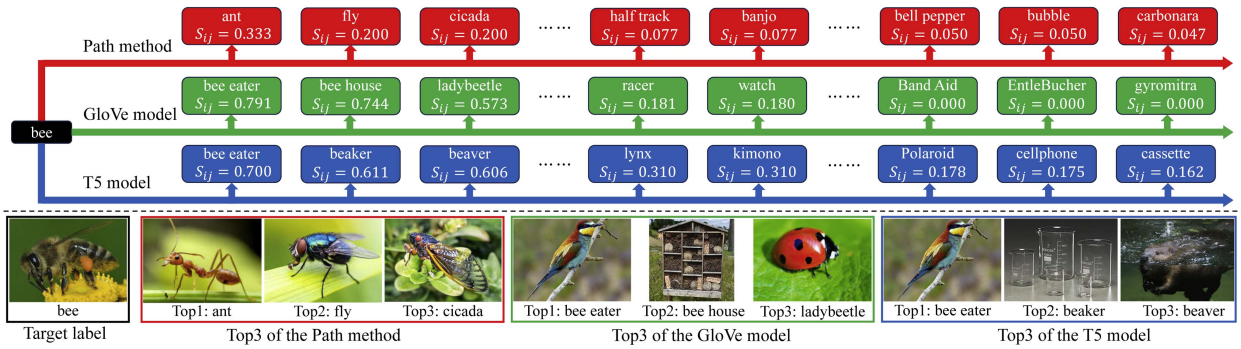
Human-annotated similarities offer a subjective benchmark for analyzing similarity measurements. To this end, we collect a similarity matrix through a user study. Annotating the similarity



(a) Case study of phrase "sea lion".



(b) Case study of compound word "ballplayer".



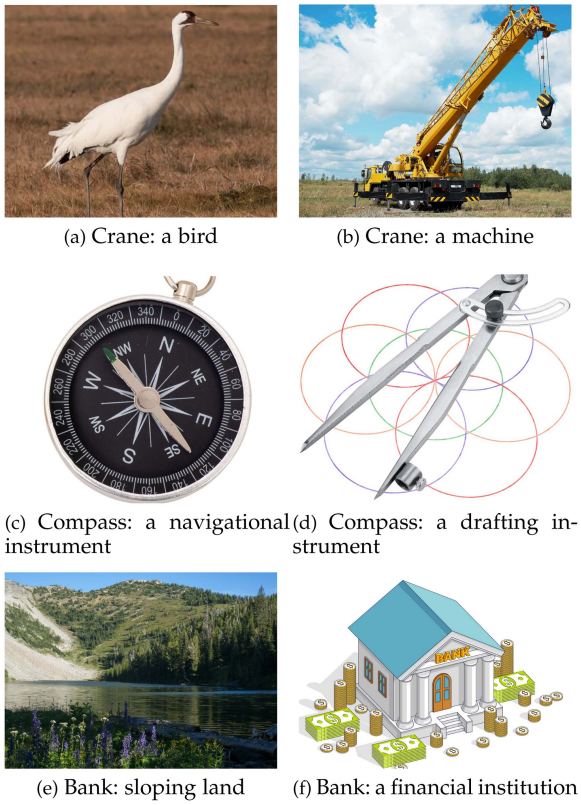
(c) Case study of simple word "bee".

Fig. 5. Similarity lists of typical labels "sea lion", "ballplayer", and "bee", calculated by the Path method, GloVe model, and T5 model. The GloVe and T5 models calculate large similarities to labels containing similar text. While the Path method assigns large values to labels with inherent biological similarity. In addition, the GloVe and T5 models overestimate the semantic relationship between labels compared to the Path method of WordNet.

matrix for ImageNet-1 k involves 499,500 elements (excluding diagonal and symmetric elements), making it impractical to complete it within a short timeframe. Thus, we randomly select 100 labels from ImageNet-1 k, named Mini-ImageNet, and conduct a simplified user study focusing on 4,950 label pairs. For each label pair, we collect 10 similarity scores from subjects (49,500 scores for 4,950 label pairs). The average score for each label pair is used to create an annotated similarity matrix. Each of the 1,000 subjects scores 50 distinct label pairs on a scale from 0 (lowest) to 1 (highest similarity). The mean and standard deviation of this matrix are shown in Table III.

To provide a qualitative analysis, we show the distributions of the human-annotated similarity matrix alongside those generated by other methods for Mini-ImageNet in Fig. 9. The Path method aligns most closely with the human-annotated similarities among the investigated methods. We also visualize the heatmaps of the human-annotated similarity matrix alongside those generated by the Path method, Glove model, and T5 model, which are the most human-aligned methods within the WordNet-based methods, text embedding models, and pre-trained language models, respectively, as shown in Fig. 10. The heatmap of the Path method exhibits the highest similarity to that of humans, further indicating its closest alignment with human-annotated similarities.

We evaluate various similarity measurements by analyzing their difference matrices to provide a quantitative assessment. These matrices are computed through element-wise subtraction between the human-annotated similarity matrix and the similarity matrices generated by other methods for the Mini-ImageNet labels. The absolute values of all elements in each difference matrix are then used to compute their mean and standard deviation, as shown in Table IV. Among all the methods examined, the



(a) Crane: a bird (b) Crane: a machine

(c) Compass: a navigational instrument (d) Compass: a drafting instrument

(e) Bank: sloping land (f) Bank: a financial institution

Fig. 6. Polysemy in category names within the ImageNet-22 k dataset. (a) and (b) illustrate the different meanings of the word "crane". (c) and (d) indicate the different meanings of the word "compass". (e) and (f) show the different meanings of the word "bank". Text embedding and pre-trained language models extract identical embeddings for these terms, thus failing to distinguish between their varied meanings.

TABLE IV
MEAN AND STANDARD DEVIATION OF THE DIFFERENCE MATRICES BETWEEN THE HUMAN-ANNOTATED SIMILARITY MATRIX AND THOSE GENERATED BY OTHER METHODS FOR MINI-IMAGENET

Method	WordNet method				Text embedding model			Pre-trained language model			
	Lin	Lesk	Path	WuP	Glove	WV	FT	T5	CLIP	BERT	GPT-2
Mean	0.101	0.060	0.028	0.346	0.074	0.098	0.108	0.252	0.415	0.596	0.884
Std	0.053	0.044	0.024	0.133	0.063	0.084	0.073	0.119	0.139	0.120	0.100

Path method achieves the minimal mean and standard deviation across all investigated methods, quantitatively demonstrating its closest alignment with the human-annotated similarities.

G. Discussion

The above analyses verify that Path Similarity has five advantages in calculating the similarity matrix for open-vocabulary tasks: 1) generates moderate similarity values. 2) distinguishes polysemy. 3) performs unbiasedly to training data. 4) aligns with human cognition. and 5) agrees with subjects in the user study.

H. Details of Similarity Measures

We analyze and evaluate eleven similarity measurements, including Path [31], Wu-Palmer [69], Lin [36], Lesk [31], Word2Vec [42], GloVe [46], FastText [3], CLIP [51], BERT [14], GPT [52] and T5 [53]). The specific details of these methods are listed as follows.

Path Similarity: It measures the shortest path length between two synsets in the WordNet hierarchy. Path similarity is calculated as the inverse of the shortest path length plus one.

Wu-Palmer Similarity: It measures the depth of the two synsets in the WordNet hierarchy and the depth of their least common subsumed. This similarity is calculated as the ratio of twice the depth of the least common subsumer to the sum of the depths of the two synsets.

Lin Similarity: It measures the similarity between two synsets based on the information content of their least common subsumer and their individual information content. This metric is calculated as twice the information content of the least common subsumer divided by the sum of the information content of the two synsets.

Lesk Similarity: It measures the overlap of words in the glosses of two synsets. Lesk similarity is calculated based on the number of common words and the total number of words in the glosses.

Word2Vec: Word2Vec uses either the Continuous Bag of Words (CBOW) or Skip-gram architecture to capture semantic relationships between words. It represents words in a continuous vector space where similar words are closer, making it suitable for various natural language processing tasks.

GloVe: GloVe focuses on creating word embeddings that consider the global co-occurrence statistics of words. It leverages a matrix factorization approach to generate word vectors that encode semantic relationships between words based on their co-occurrence frequencies.

FastText: FastText is an extension of traditional word embeddings. It breaks words into smaller subword units called character n-grams and learns embeddings for these subword units in addition to full words. This approach allows FastText to represent out-of-vocabulary words effectively and capture morphological information, making it particularly useful for languages with rich morphology.

CLIP: CLIP is designed for multimodal understanding, capable of comprehending both text and images. CLIP learns to map textual descriptions and images into a shared embedding space, enabling it to perform tasks such as image classification and cross-modal retrieval.

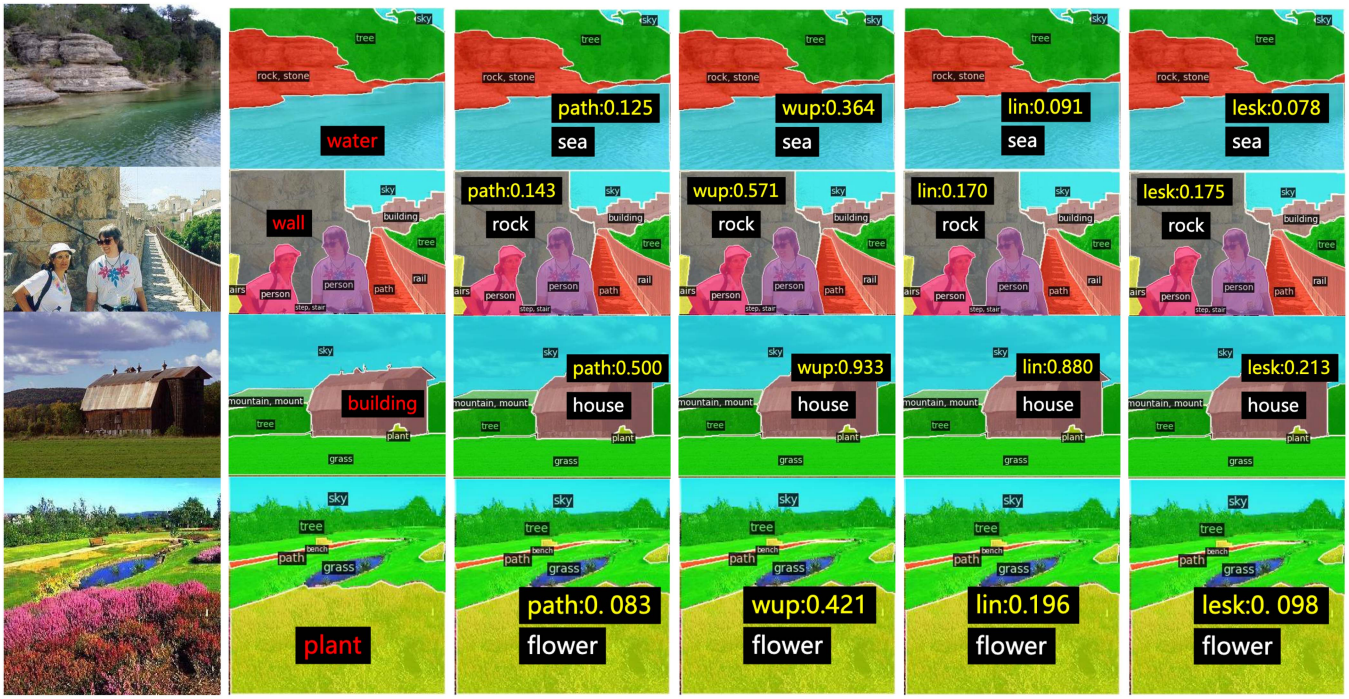
BERT: BERT is a language model that introduces bidirectional pre-training of contextual word representations. It learns to predict missing words in sentences by considering the entire context from both directions.

GPT-2: GPT-2 is a large-scale transformer-based language model that demonstrates the capabilities of generative text models. It can generate coherent and contextually relevant text based on a given prompt.

T5: T5 is a large language model that follows a text-to-text framework, where all NLP tasks are cast into a consistent format. It pre-trains on a wide range of text-to-text tasks, making it adaptable to various natural language understanding and generation tasks.

V. BENCHMARKS

In this section, we employ Open mIoU, PQ, and AP metrics to assess open-vocabulary semantic (Section V-A), panoptic (Section V-B), and instance segmentation (Section V-C),



(a) Image (b) mask (c) Path method (d) Wup method (e) Lin method (f) Lesk method

Fig. 7. Visualization of four test samples for user study on similarity measurements. (c) to (f) respectively present the semantic similarities calculated by the Path Similarity, Wu-Palmer Similarity, Lin Similarity, and Lesk Similarity in WordNet.

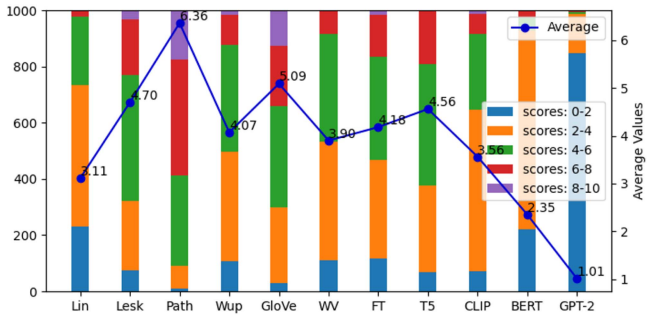


Fig. 8. Visualization of distributions and mean scores from a study with 1,000 subjects, assessing 50 label pairs' similarities calculated by various methods. Scores ranged from 0 (lowest) to 10 (highest satisfaction).

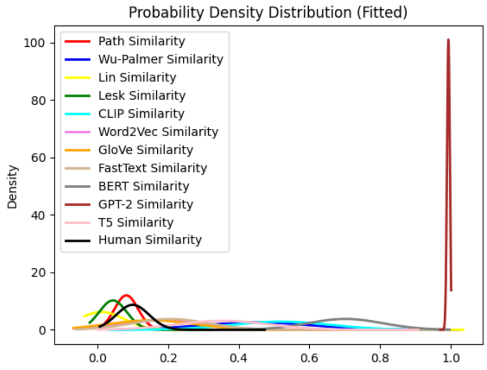


Fig. 9. Probability density distribution curves of the similarity matrices annotated by humans and those computed by other methods for Mini-ImageNet's labels.

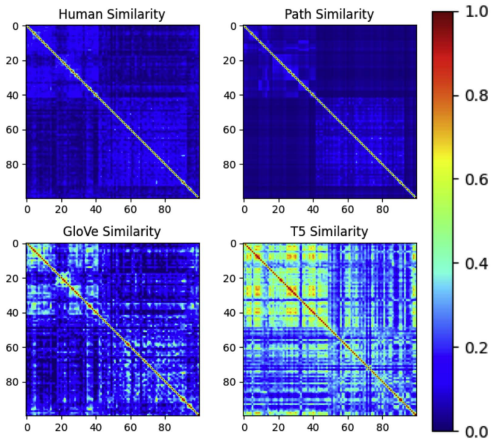


Fig. 10. Heatmaps of the similarity matrices annotated by humans and those computed by the Path method, GloVe model, and T5 model for Mini-ImageNet's labels.

respectively. All results presented in this section are evaluated using publicly released models from various methods. We utilize widely used benchmark datasets for performance evaluation, including ADE20K [83], Pascal VOC [45], Pascal Context [17], COCO [37], and COCO-Stuff [4].

ADE20 K (A-150/847): ADE20 K contains over 22 K annotated images, split into 20 K training and 2 K validation images. It provides two labeling schemes: A-150, with 150 categories, and A-847, with 847 categories. The A-150 supports benchmarking

TABLE V
RESULTS ON OPEN-VOCABULARY SEMANTIC SEGMENTATION EVALUATED BY VANILLA AND OPEN MIOU IN A CROSS-DATASET SETTING

Method	Metric	A-847	PC-459	A-150	PC-59	PAS-20
SimSeg [73]	Vanilla mIoU	6.7	8.7	20.4	47.3	88.3
	Open mIoU	12.9$^{(+6.2)}$	14.5$^{(+5.8)}$	29.0$^{(+8.6)}$	50.8$^{(+3.5)}$	90.2$^{(+1.9)}$
POMP [55]	Vanilla mIoU	-	-	20.7	51.0	-
	Open mIoU	-	-	29.6$^{(+8.9)}$	54.5$^{(+3.5)}$	-
OVSeg [35]	Vanilla mIoU	8.9	12.3	29.6	55.8	94.7
	Open mIoU	15.7$^{(+6.8)}$	17.4$^{(+5.1)}$	36.9$^{(+7.3)}$	59.6$^{(+3.8)}$	95.6$^{(+1.1)}$
SAN [72]	Vanilla mIoU	12.8	16.6	31.9	57.5	95.2
	Open mIoU	19.2$^{(+6.4)}$	19.9$^{(+3.3)}$	39.0$^{(+7.1)}$	60.4$^{(+2.9)}$	96.0$^{(+0.8)}$
CAT-Seg [11]	Vanilla mIoU	11.4	16.2	31.5	61.6	96.5
	Open mIoU	18.4$^{(+7.0)}$	20.3$^{(+4.1)}$	39.9$^{(+8.4)}$	65.4$^{(+3.8)}$	97.2$^{(+0.7)}$
FC-CLIP [75]	Vanilla mIoU	14.8	13.8	34.1	58.4	95.4
	Open mIoU	20.6$^{(+5.8)}$	16.3$^{(+2.5)}$	40.4$^{(+6.3)}$	61.3$^{(+2.9)}$	96.2$^{(+0.8)}$

Yellow cells show a performance reversal between vanilla and Open mIoU for SAN and CAT-Seg methods.

TABLE VI
RESULTS ON OPEN-VOCABULARY PANOPTIC SEGMENTATION EVALUATED BY VANILLA AND OPEN PQ IN A CROSS-DATASET SETTING

Method	Metric	Source Dataset: COCO PQ	SQ	RQ	Target Dataset: ADE20K (A-150) PQ	SQ	RQ
OPSNet [8]	Vanilla PQ	44.0	83.5	62.1	17.8	54.6	21.6
	Open PQ	45.9$^{(+1.9)}$	83.4$^{(-0.1)}$	63.4$^{(+1.3)}$	21.9$^{(+4.1)}$	76.0$^{(+21.4)}$	26.7$^{(+6.1)}$
ODISE [71]	Vanilla PQ	34.9	84.3	53.5	23.2	74.4	27.9
	Open PQ	37.1$^{(+2.2)}$	84.1$^{(-0.2)}$	55.4$^{(+1.9)}$	25.6$^{(+2.4)}$	79.4$^{(+5.0)}$	30.9$^{(+3.0)}$
FC-CLIP [75]	Vanilla PQ	54.4	83.0	64.8	26.8	71.6	32.3
	Open PQ	55.3$^{(+0.9)}$	83.0	65.9$^{(+1.1)}$	28.8$^{(+2.0)}$	77.9$^{(+6.3)}$	34.8$^{(+2.5)}$

TABLE VII
RESULTS ON OPEN-VOCABULARY PANOPTIC SEGMENTATION IN THE ZERO-SHOT SETTING

Method	Metric	Known PQ^{Th}	SQ^{Th}	RQ^{Th}	Unknow PQ^{Th}	SQ^{Th}	RQ^{Th}
CGG [64]	Vanilla PQ	48.1	82.4	57.8	34.6	77.6	40.9
	Open PQ	48.8$^{(+0.7)}$	82.3$^{(-0.1)}$	58.6$^{(+0.8)}$	35.3$^{(+0.7)}$	81.9$^{(+4.3)}$	41.8$^{(+0.9)}$

Since only the "thing" classes are classified into "known" and "unknown" classes, we only present the results for the "thing" classes here.

for semantic, instance, and panoptic segmentation, while the A-847 is solely for semantic segmentation.

Psacal Context (PC-59/459): Pascal Context is a semantic segmentation dataset consisting of 5 K training and 5 K validation images. It offers two labeling schemes: PC-459, which includes 459 labels, and PC-59, which features 59 labels.

Pascal VOC (PAS-20): Pascal VOC is a semantic segmentation dataset comprising 1,464 training and 1,449 validation images, categorized into 20 classes (PAS-20).

COCO: COCO is a widely used dataset for instance segmentation and object detection, consisting of over 200 K annotated images spanning 80 category labels.

COCO-Stuff: COCO-Stuff is an extension of the COCO dataset, comprising over 164 K images annotated with 80 "thing" and 91 "stuff" categories. It is used to train open-vocabulary semantic segmentation models.

A. Open-Vocabulary Semantic Segmentation

Table V shows the results on open-vocabulary semantic segmentation methods which trained on the COCO-Stuff [4] dataset and evaluated on datasets such as Pascal VOC (PAS-20) [45], Pascal Context (PC-59/459 for 59/459 categories) [17], and ADE20 K (A-150/847 for 150/847 categories) [83]. Open mIoU consistently outperforms vanilla mIoU across all five test datasets, as it measures wrong predictions by considering the similarity between predicted and ground-truth labels. Additionally, the improvement of Open mIoU becomes less distinct with higher vanilla mIoU scores due to fewer label misclassifications in these cases. Furthermore, we observe a performance reversal between vanilla and Open mIoU for SAN and CAT-Seg. Specifically, CAT-Seg exceeds SAN in Vanilla mIoU on PC-459 and A-150 datasets, but SAN outperforms CAT-Seg in Open mIoU on these datasets. This reversal indicates that, compared to CAT-Seg, SAN struggles more with classifying pixels into specific ground truth categories but performs better at assigning pixels to semantically similar labels when direct classification is difficult, leading to superior Open mIoU performance. These results further demonstrate that Open mIoU provides a more comprehensive evaluation of models' segmentation ability by considering label similarity.

The mIoU evaluates pixel-wise classification. Table V indicates that methods generally achieve lower Open mIoU in datasets with a larger number of categories, while higher Open mIoU on datasets with fewer categories. This discrepancy arises because accurately classifying pixels into correct or related categories becomes more challenging with more classes. To address this, we suggest grouping labels based on biological relationships, such as WordNet hierarchies, and adopting a multilevel classification strategy. For example, pixels can first be classified into broader groups and then into the specific labels within each group. This strategy improves group classification accuracy due to the small number of groups and reduces the difficulty of specific label classification within each group. Most importantly, misclassifications within the same group can still yield relatively high scores, as labels within a group share higher similarities.

The methods evaluated in this experiment include SimSeg [73], POMP [55], OVSeg [35], SAN [72], and FC-CLIP [75]. SimSeg adopts a two-stage framework, using MaskFormer [10] with a ResNet-101 backbone for mask proposal extraction and CLIP with a ViT-B/16 backbone for open-vocabulary classification. POMP, a pre-trained vision-language model, utilizes CLIP with a ViT-B/16 backbone to generate text embeddings for mask classification within SimSeg. OVSeg fine-tunes CLIP on masked regions paired with text descriptions, using MaskFormer with a Swin-B backbone to generate masks and fine-tuned CLIP with a ViT-L/14 backbone for classification. SAN integrates a side module after frozen CLIP, with two branches: one predicting mask proposals and the other attention bias. The frozen CLIP uses a ViT-L/14 backbone. FC-CLIP introduces a shared Frozen Convolutional CLIP with a ConvNeXt-Large backbone, transitioning methods to a one-stage framework and improving the accuracy-cost trade-off.

B. Open-Vocabulary Panoptic Segmentation

Tables VI and VII present the results of open-vocabulary panoptic segmentation methods in the cross-dataset and

zero-shot settings. In Table VI, all methods are trained on the dataset COCO [37] and evaluated on the target dataset ADE20 K. Similarly, methods in Table VII are trained and tested on the COCO dataset, where 64 thing classes are categorized as known classes, while the remaining 16 thing classes are designated as unknown classes.

Segmentation Quality: The SQ in vanilla PQ considers only matched entities, possibly skewing true segmentation quality by including classification accuracy. In contrast, the SQ in Open PQ evaluates overlapping entities (IoU $> \delta$) regardless of their labels. In Table VI, Open SQ shows results akin to vanilla SQ on the source dataset, indicating that entities with incorrect labels possess segmentation quality comparable to those with correct labels. Conversely, on the target dataset, Open SQ substantially increases to vanilla SQ, indicating that overlapping entities with incorrect labels exhibit significantly better segmentation quality than those with correct labels. Similar results can be seen in Table VII when comparing known and unknown classes. Apparently, the vanilla SQ severely underestimates the methods' segmentation quality on the target/unknown dataset. Consequently, the SQ term in Open PQ proves sensitive to segmentation quality and can evaluate the segmentation quality of entity pairs with incorrect label predictions.

Recognition Quality: Table VI shows a disparity in the improvement of Open RQ between the source and target datasets. This discrepancy arises because the model trained on the source dataset tends to classify entities into the most accurate category rather than semantic similar categories when inferring on the same source dataset. However, the target dataset contains numerous unseen classes, and in such cases, the model classifies these entities into semantically similar categories. Consequently, more substantial improvements are observed for Open RQ on the target dataset. This observation validates the sensitivity of the proposed Open RQ to the semantic distances between labels, enabling the evaluation of recognition quality that considers semantic similarity.

The PQ metric evaluates both the segmentation quality (SQ) and recognition quality (RQ) of masks. The methods in Tables VI and VII exhibit high and comparable Open SQ values, leaving limited room for improvement. However, their low Open RQ values on the target dataset significantly influence the final Open PQ scores, highlighting the importance of Open RQ. Since Open RQ primarily focuses on the classification accuracy of masks, as illustrated in (3) and Table I, future works are recommended to adopt the multilevel classification strategy proposed in Section V-A to improve mask classification and, consequently, Open RQ.

The methods benchmarked in this experiment include FC-CLIP, OPSNet [8], ODISE [71], and CGG [64]. FC-CLIP follows the settings described in Section V-A. OPSNet [8] introduces an Embedding Modulation module that enhances embedding and facilitates information exchange between the segmentation model Mask2Former [9] and the CLIP encoder, which uses an RN50 backbone. ODISE [71] develops a diffusion-based approach, combining stable diffusion [57] with the Mask2Former segmentation model. CGG [64] introduces a joint grounding and generation framework, reducing the reliance on caption datasets and complex pipelines, using ResNet-50 as the backbone and BERT [14] embeddings for classification and encoding.

TABLE VIII
RESULTS ON OPEN-VOCABULARY INSTANCE SEGMENTATION EVALUATED BY VANILLA AND OPEN AP IN A CROSS-DATASET SETTING

Method	Metric	Target Dataset: ADE20K (A-150) AP	AP50	AP75	AP_s	AP_m	AP_l
ODISE [71]	Vanilla AP	14.0	23.4	13.9	3.3	16.2	27.6
	Vanilla AP†	12.7	20.9	12.7	2.8	14.5	25.3
	Open AP†	14.3$^{(+1.6)}$	23.7$^{(+2.8)}$	14.2$^{(+1.5)}$	3.8$^{(+1.0)}$	16.4$^{(+2.0)}$	27.3$^{(+2.0)}$
FC-CLIP [75]	Vanilla AP	16.8	27.4	17.2	5.7	19.9	29.7
	Vanilla AP†	15.8	25.5	16.3	5.4	18.7	27.7
	Open AP†	17.1$^{(+1.3)}$	27.8$^{(+2.3)}$	17.4$^{(+1.1)}$	6.1$^{(+0.7)}$	20.0$^{(+1.3)}$	29.2$^{(+1.5)}$

*† signifies the use of a class-agnostic global ranking during the matching process.

TABLE IX
RESULTS ON OPEN-VOCABULARY INSTANCE SEGMENTATION IN THE ZERO-SHOT SETTING

Method	Metric	Constrained (AP50) Base	Novel	Generalized (AP50) Base	Novel	All
CGG [64]	Vanilla AP	46.2	29.5	46.0	28.1	41.4
	Vanilla AP†	44.2	28.9	43.6	26.4	39.1
	Open AP†	45.2$^{(+1.0)}$	30.0$^{(+1.1)}$	44.8$^{(+1.2)}$	28.7$^{(+2.3)}$	40.6$^{(+1.5)}$

This table tests AP under IoU of 0.5.*† signifies the use of a class-agnostic global ranking during the matching process.

C. Open-Vocabulary Instance Segmentation

In Tables VIII and IX, we present the results of open-vocabulary instance segmentation in the cross-dataset and zero-shot settings. In the cross-dataset setting, the methods are trained on the COCO dataset and tested on the ADE20 K dataset. In the zero-shot setting, the methods undergo both training and evaluation exclusively on the COCO dataset, where 48 classes are categorized as base classes and 17 as novel classes. Table VIII highlights the superiority of Open AP over vanilla AP in all aspects, showing that the proposed Open AP is robust to variations in entity size and matching thresholds when assessing the semantic relationship between different object classes. Table IX exhibits results similar to those in Table VI, where there is a discrepancy in the increase of Open AP between base and novel classes. These results also underscore the sensitivity of Open AP to the semantic relatedness between different labels. Consequently, Open AP proves to be a viable metric for evaluating open-vocabulary instance segmentation tasks.

Open-Vocabulary Object Detection: The Open AP metric can also evaluate object detection by replacing the mask IoU with box IoU. Table X presents the results for open-vocabulary object detection in the same zero-shot setting as used in Table IX, revealing analogous observations to those in Table IX. In addition, the stability of Open AP improvements for CGG [64] in Tables IX and X suggests that Open AP can be applied across different IoU types. These results verify that Open AP is a reasonable metric for open-vocabulary object detection. Furthermore, we witness an increment reversal between CGG and RegionCLIP in Table X. Specifically, for novel classes under the generalized setting, CGG's Open AP exceeds its vanilla AP from 26.9 to 29.1 by 2.2, while RegionCLIP's Open AP

TABLE X
RESULTS ON OPEN-VOCABULARY OBJECT DETECTION IN A ZERO-SHOT SETTING

Method	Metric	Constrained (AP50) Base	Constrained (AP50) Novel	Generalized (AP50) Base	Generalized (AP50) Novel	Generalized (AP50) All
CGG [64]	Vanilla AP	46.0	30.4	45.9	28.9	41.4
	Vanilla AP†	44.3	29.4	43.6	26.9	39.2
	Open AP†	45.3$^{(+1.0)}$	30.6$^{(+1.2)}$	44.9$^{(+1.3)}$	29.1$^{(+2.2)}$	40.8$^{(+1.6)}$
RegionCLIP [81]	Vanilla AP	62.1	42.9	61.7	39.3	55.9
	Vanilla AP†	61.0	39.8	60.3	30.1	52.4
	Open AP†	61.6$^{(+0.6)}$	40.6$^{(+0.8)}$	61.6$^{(+1.3)}$	33.6$^{(+3.5)}$	53.9$^{(+1.5)}$
Object-Centric-OVD [1]	Vanilla AP	-	-	56.7	40.5	52.5
	Vanilla AP†	-	-	54.8	36.9	50.1
	Open AP†	-	-	55.8$^{(+1.0)}$	38.7$^{(+1.8)}$	51.3$^{(+1.2)}$
CORA [68]	Vanilla AP	-	-	44.7	41.5	43.9
	Vanilla AP†	-	-	42.2	37.9	41.1
	Open AP†	-	-	43.6$^{(+1.4)}$	39.8$^{(+1.9)}$	42.6$^{(+1.5)}$

This table tests AP under IoU of 0.5.*† signifies the use of a class-agnostic global ranking during the matching process. Yellow cells show an increment reversal between CGG and RegionCLIP.

outperforms its vanilla AP from 30.1 to 33.6 by 3.5. Notably, CGG, with a lower vanilla AP, shows a smaller increase than RegionCLIP, which has a higher vanilla AP. Generally, methods with lower vanilla AP tend to have higher increments in Open AP, as a lower vanilla AP means more misclassified entities that can be rewarded with soft scores. This reversal suggests that RegionCLIP obtains more scores with fewer misclassified entities compared to CGG, highlighting its ability to assign semantically closer labels when direct classification is challenging. These results demonstrate that Open AP offers a valuable framework for evaluating misclassified entities in open-vocabulary instance segmentation or detection.

Vanilla AP adopts a class-aware matching strategy, which conflicts with the intention of the new metric. We design Open AP with a class-agnostic matching strategy. Tables VIII to X show that Open AP achieves higher values than the class-agnostic vanilla AP, reflecting its strength in considering label similarities. However, ranking proposals using classification scores of different categories introduces inconsistencies, causing Open AP to be lower than the class-aware vanilla AP, as shown in Tables VIII to X. To improve Open AP, we suggest generating two scores per proposal: one indicating object presence and another indicating category confidence. The first score ensures unbiased ranking during matching proposals, while the second refines category-specific AP calculations. In addition, the multilevel classification strategy in Section V-A can be used to generate the second score, further enhancing Open AP.

The methods evaluated in this experiment include ODISE, FC-CLIP, CGG, RegionCLIP [81], Object-Centric-OVD [1], and CORA [68]. FC-CLIP [75], ODISE [71], and CGG [64] follow the evaluation settings outlined in Sections V-A and V-B. RegionCLIP [81] extends CLIP with an RN50x5 backbone for object detection tasks by adapting it to recognize region-level visual representations detected by Faster R-CNN [54] using an RN50x4-C4 backbone. Object-Centric-OVD [1] improves open-vocabulary detection through knowledge distillation and object-level pseudo-labeling, using Faster R-CNN with an RN50x5 backbone as the detector. CORA [68] introduces a DETR-style framework with region prompting and anchor pre-matching to improve object localization without extra training datasets. This method uses DAB-DETR [38] as the detector and CLIP with an RN50x4 backbone for classification.

TABLE XI
RESULTS ON OPEN-VOCABULARY PANOPTIC SEGMENTATION EVALUATED BY OPEN PQ USING THREE DIFFERENT SIMILARITY MATRICES CALCULATED BY THE PATH SIMILARITY, GLOVE MODEL, AND T5 MODEL

Method	Metric	Similarity Measures	Target Dataset: ADE20K (A-150) PQ	SQ	RQ
FC-CLIP [75]	Vanilla PQ	-	26.8	71.6	32.3
	Open PQ	Path Similarity	28.8$^{(+2.0)}$	77.9$^{(+6.3)}$	34.8$^{(+2.5)}$
		GloVe model	31.5$^{(+4.7)}$	77.7$^{(+6.1)}$	38.2$^{(+5.9)}$
		T5 model	30.7$^{(+3.9)}$	77.8$^{(+6.2)}$	37.1$^{(+4.8)}$

D. Ablation Study

Table XI presents the results for open-vocabulary panoptic segmentation evaluated by Open PQ with similarity matrices calculated by the Path Similarity, GloVe model, and T5 model. The FC-CLIP method evaluated in this study follows the evaluation settings described in Section V-A. The increments in Open SQ exhibit similar values when using different similarity matrices. As Open SQ focuses solely on segmentation quality, disregarding classification accuracy in performance evaluation is preferable. When using the GloVe or T5 model, the increases in Open RQ exhibit significantly higher values than the Path method, indicating an overestimated evaluation of classification accuracy for entities. The evaluation of open metrics yields moderate results when using the Path method, making it the preferred choice for open-vocabulary tasks.

E. Real-World Open-Vocabulary Segmentation

In the current evaluation settings, the evaluated categories of each test set are limited by a fixed label set. However, the images in test set are collected from real-world scenes that contain categories far more than a fixed label set. Evaluating methods' performance solely on these fixed labels can not fully explore their actual open ability. To address this, we introduce another evaluation setting.

In this setting, we utilize labels outside A-150 to fully explore the models' open ability across a broader range of categories. First, we apply label sets of A-847, PC-459, PC-59, and PAS-20 on FC-CLIP to generate semantic segmentation results for the test images of A-150. Since A-150 lacks annotations for these external label sets, we then leverage a cross-dataset similarity matrix to map the predicted categories to the A-150 categories. Specifically, for each predicted category from the external label set, we check its similarities with the labels of A-150 and choose the highest as its mapped category. Finally, we evaluate the mapped predictions using Open mIoU. Note that the FC-CLIP method evaluated in this study adheres to the evaluation settings described in Section V-A.

Experimental results in Table XII reveal that using label sets out of the current dataset decreases Open mIoU. This is expected because transferring the predicted label sets to the A-150 label set will enlarge their semantic distance from the ground truth labels. Additionally, model performance does not directly correlate with the number of predicted labels. For instance, PC-459 outperforms A-847, and PAS-20 outperforms PC-59. This occurs because there is no correlation between the overlaps

TABLE XII
RESULTS ON OPEN-VOCABULARY SEMANTIC SEGMENTATION ADOPT A NOVEL EVALUATION SETTING IN WHICH THE LABEL SETS OF A-847, PC-459, A-150, PC-59, AND PAS-20 ARE USED TO GENERATE DIFFERENT LABEL PREDICTIONS FOR THE A-150 DATASET

Method	Metric	Test Dataset and Annotation: A-150				
		A-847	PC-459	A-150	PC-59	PAS-20
FC-CLIP [75]	Open mIoU	19.2	25.2	40.4	11.2	11.6

TABLE XIII
MEAN AND STANDARD DEVIATION OF THE SATISFACTION SCORES OBTAINED FROM A USER STUDY INVOLVING 1000 SUBJECTS WITH 50 EVALUATED IMAGES ARE PRESENTED

Similarity methods	Path method	GloVe model	T5 model
Mean	6.12	3.05	4.29
Std	1.71	0.99	1.23

Scores ranged from0 (lowest) to 10 (highest satisfaction).

of novel label sets and current dataset categories and the number of categories within the novel label sets.

F. User Study on Open Metrics

We conduct a user study to assess human preferences in evaluating segmentation results with different similarity matrices. This study involves surveying 1,000 subjects using 50 images randomly selected from the ADE20 K dataset. Predictions are generated using FC-CLIP [75], and evident error label predictions are evaluated using Open PQ with three different similarity matrices derived from Path Similarity, GloVe model, and T5 model. The FC-CLIP method evaluated in this study adheres to the evaluation settings outlined in Section V-A. The fourth row of Fig. 1 shows some samples of this study. Subjects rate their satisfaction with the results on a scale from 0 (lowest) to 10 (highest). Table XIII shows subjects perfer evaluation results based on Path Similarity.

We conduct a user study to assess human preferences regarding segmentation results generated by different methods. The study involves 300 subjects, each evaluating 20 images randomly selected from the ADE20 K dataset. We first apply six open-vocabulary segmentation methods to segment these images and present their results alongside ground truth annotations. Participants are then asked to rank the six predicted results for each image using ranking IDs from #1 to #6 based on their satisfaction with the segmentation performance. The rankings from #1 (best) to #6 correspond to satisfaction levels. Finally, we calculate the mean of all chosen ranking IDs as the ranking scores for each method, where lower scores indicate higher satisfaction. The ranking scores and the open mIoU values of the different methods are presented in Table XIV. The open mIoU values correlate negatively with the human ranking scores, demonstrating that the Open metrics accurately reflect the true performance of the evaluated methods.

G. Open Metrics With Path Similarity for Polysemy

Open-vocabulary segmentation methods often struggle to classify polysemous terms correctly, as the same word embedding is extracted for different synsets within a polysemous word. The Path method, which calculates similarities based on synsets, effectively mitigates this challenge. To demonstrate the advantage of Open metrics enhanced with Path similarity, we annotate a dataset derived from the ImageNet-21 k. Specifically, we select 10 polysemous words, each with two distinct synsets, and annotate 10 images for each synset, resulting in a dataset of 200 images across 20 categories, named PolyImage. The polysemous words include “flip”, “cannikin”, “cricket”, “bed”, “garden”, “jumper”, “pump”, “rotisserie”, “screwdriver”, and “weir”.

We use these 10 words to extract CLIP embeddings for mask classification. The state-of-the-art method FC-CLIP [71] is employed to evaluate the PolyImage. Table XV shows the results using vanilla and Open metrics. To ensure a fair comparison with vanilla metrics, we adjust the Open metrics to only reward misclassifications within the same polysemous word. The experimental results indicate that Open metrics outperform vanilla metrics, demonstrating their ability to assign reasonable scores to misclassifications arising from polysemy.

TABLE XIV
COMPARISONS BETWEEN OPEN MIOU VALUES AND RANKING SCORES OBTAINED THROUGH A USER STUDY INVOLVING 300 SUBJECTS WITH 20 EVALUATED IMAGES

Method	Number of ranking ID						Ranking score	Open mIoU
	#1	#2	#3	#4	#5	#6		
SimSeg [73]	353	327	338	620	1412	2950	4.88	30.4
POMP [55]	333	563	825	1161	1391	1727	4.32	30.9
OVSeg [35]	317	333	1144	1974	1898	334	3.97	35.3
SAN [72]	335	1956	1955	1068	349	337	3.03	41.6
CAT-Seg [11]	1613	1410	1159	854	620	344	2.75	42.2
FC-CLIP [75]	3049	1411	579	323	330	308	2.07	43.5

The rankings from #1 (best) to #6, indicate satisfaction levels. Ranking scores are the mean values of all chosen ranking IDs, where lower scores correspond to higher satisfaction levels.

TABLE XV
VANILLA AND OPEN METRICS RESULTS ON THE POLYIMAGE DATASET EVALUATED BY FC-CLIP

Method	Semantic segmentation		Panoptic segmentation		Instance segmentation	
	Vanilla mIoU	Open mIoU	Vanilla PQ	Open PQ	Vanilla AP†	Open AP†
FC-CLIP	80.9	82.3$^{(+1.4)}$	64.7	66.8$^{(+2.1)}$	39.6	40.1$^{(+0.9)}$

*†signifies the use of a class-agnostic global ranking during the matching process.

VI. CONCLUSION

In this work, we address the limitations of existing evaluation metrics in open-vocabulary image segmentation by proposing new open evaluation metrics that incorporate semantic similarity between predicted and ground truth labels. We emphasize the superiority of the Path method in calculating semantic similarity for open-vocabulary tasks, as it captures meaningful relationships between category labels. Through extensive evaluations, we demonstrate the effectiveness of our proposed metrics in assessing model performance on open-vocabulary image segmentation tasks. Our contributions enhance the evaluation framework for open-vocabulary segmentation, enabling better model development and assessment in real-world applications.

REFERENCES

[1] H. Bangalath, M. Maaz, M. U. Khattak, S. H. Khan, and F. Shahbaz Khan, "Bridging the gap between object and image-level representations for open-vocabulary detection," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 33781–33794.

[2] D. S. Bassett and M. S. Gazzaniga, "Understanding complexity in the human brain," *Trends Cogn. Sci.*, vol. 15, no. 5, pp. 200–209, 2011.

[3] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, 2017.

[4] H. Caesar, J. Uijlings, and V. Ferrari, "COCO-stuff: Thing and stuff classes in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1209–1218.

[5] K. Chen et al., "Hybrid task cascade for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4974–4983.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[7] P. Chen, K. Sheng, M. Zhang, Y. Shen, K. Li, and C. Shen, "Open vocabulary object detection with proposal mining and prediction equalization," 2022, *arXiv: 2206.11134*.

[8] X. Chen, S. Li, S.-N. Lim, A. Torralba, and H. Zhao, "Open-vocabulary panoptic segmentation with embedding modulation," 2023, *arXiv:2303.11324*.

[9] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," 2021, *arXiv:2112.01527*.

[10] B. Cheng, A. G. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 17864–17875.

[11] S. Cho et al., "CAT-seg: Cost aggregation for open-vocabulary semantic segmentation," 2023, *arXiv:2303.11797*.

[12] R. Chu, Y. Chen, T. Kong, L. Qi, and L. Li, "ICM-3D: Instantiated category modeling for 3D instance segmentation," *IEEE Robot. Automat. Lett.*, vol. 7, no. 1, pp. 57–64, Jan. 2022.

[13] T. Deselaers and V. Ferrari, "Visual and semantic similarity in imagenet," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1777–1784.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv: 1810.04805*.

[15] Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, and G. Li, "Learning to prompt for open-vocabulary object detection with vision-language model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14064–14073.

[16] Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, and G. C. Li, "Learning to prompt for open-vocabulary object detection with vision-language model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14064–14073.

[17] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, pp. 303–338, 2010.

[18] C. Fellbaum, *About Wordnet*. Princeton, NJ, USA: Princeton Univ., 2010.

[19] M. Gao et al., "Open vocabulary object detection with pseudo bounding-box labels," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 266–282.

[20] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Open-vocabulary image segmentation," 2021, *arXiv:2112.12143*.

[21] X. Gu, T. Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," 2021, *arXiv:2104.13921*.

[22] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[23] D. Huynh, J. Kuen, Z. Lin, J. Gu, and E. Elhamifar, "Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7010–7021.

[24] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. IEEE Int. Conf. Mach. Learn. Appl.*, 2021, pp. 7010–7021.

[25] D. Kim, A. Angelova, and W. Kuo, "Contrastive feature masking open-vocabulary vision transformer," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 15556–15566.

[26] D. Kim, A. Angelova, and W. Kuo, "Region-aware pretraining for open-vocabulary object detection with vision transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11144–11154.

[27] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6399–6408.

[28] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9404–9413.

[29] W. Kuo, Y. Cui, X. Gu, A. J. Piergiovanni, and A. Angelova, "F-VLM: Open-vocabulary object detection upon frozen vision and language models," 2022, *arXiv:2209.15639*.

[30] C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification," *WordNet: An Electron. Lexical Database*, vol. 49, no. 2, pp. 265–283, 1998.

[31] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone," in *Proc. 5th Annu. Int. Conf. Syst. Documentation*, 1986, pp. 24–26.

[32] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," 2022, *arXiv:2201.03546*.

[33] X. Li et al., "Transformer-based visual segmentation: A survey," 2023, *arXiv:2304.09854*.

[34] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He, "Scaling language-image pre-training via masking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23390–23400.

[35] F. Liang et al., "Open-vocabulary semantic segmentation with mask-adapted clip," 2022, *arXiv:2210.04150*.

[36] D. Lin et al., "An information-theoretic definition of similarity," in *Proc. IEEE Int. Conf. Mach. Learn. Appl.*, 1998, pp. 296–304.

[37] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[38] S. Liu et al., "DAB-DETR: Dynamic anchor boxes are better queries for DETR," 2022, *arXiv:2201.12329*.

[39] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.

[40] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[41] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[42] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[43] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[44] M. Minderer et al., "Simple open-vocabulary object detection with vision transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 728–755.

[45] R. Mottaghi et al., "The role of context for object detection and semantic segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 891–898.

[46] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[47] L. Qi, L. Jiang, S. Liu, X. Shen, and J. Jia, "Amodal instance segmentation with KINS dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3009–3018.

[48] L. Qi et al., "High quality entity segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 4024–4033.

[49] L. Qi et al., "Open world entity segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8743–8756, Jul. 2023.

[50] L. Qi et al., "PointINS: Point-based instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6377–6392, Oct. 2022.

[51] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. IEEE Int. Conf. Mach. Learn. Appl.*, 2021, pp. 8748–8763.

[52] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[53] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.

[54] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

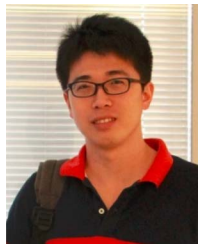
[55] S. Ren et al., "Prompt pre-training with twenty-thousand classes for open-vocabulary visual recognition," 2023, *arXiv:2304.04704*.
[56] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proc. Int. Joint Conf. Artif. Intell.*, 1995, pp. 448–453.
[57] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.
[58] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Peroc. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.
[59] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, 2015.
[60] T. Shen et al., "High quality segmentation for ultra high-resolution images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1300–1309.
[61] L. G. Ungerleider and J. V. Haxby, "'what' and 'where' in the human brain," *Curr. Opin. Neurobiol.*, vol. 4, no. 2, pp. 157–165, 1994.
[62] L. Wang et al., "Object-aware distillation pyramid for open-vocabulary object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11186–11196.
[63] T. Wang and N. Li, "Learning to detect and segment for open vocabulary object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7051–7060.
[64] J. Wu et al., "Betrayed by captions: Joint caption grounding and generation for open vocabulary instance segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 21881–21891.
[65] J. Wu et al., "Towards open vocabulary learning: A survey," 2023, *arXiv:2306.15880*.
[66] S. Wu, W. Zhang, S. Jin, W. Liu, and C. C. Loy, "Aligning bag of regions for open-vocabulary object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 21881–21891.
[67] S. Wu et al., "Clipself: Vision transformer distills itself for open-vocabulary dense prediction," 2023, *arXiv:2310.01403*.
[68] X. Wu, F. Zhu, R. Zhao, and H. Li, "CORA: Adapting CLIP for open-vocabulary detection with region prompting and anchor pre-matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7031–7040.
[69] Z. Wu and M. Palmer, "Verb semantics and lexical selection," 1994, *arXiv:cmp-lg/9406033*.
[70] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.
[71] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," 2023, *arXiv:2303.04803*.
[72] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "Side adapter network for open-vocabulary semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2945–2954.
[73] M. Xu et al., "A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2022, pp. 736–753.
[74] S. Xu et al., "DST-Det: Simple dynamic self-training for open-vocabulary object detection," 2023, *arXiv:2310.01393*.
[75] Q. Yu, J. He, X. Deng, X. Shen, and L.-C. Chen, "Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip," 2023, *arXiv:2308.02487*.
[76] Y. Zang, W. Li, K. Zhou, C. Huang, and C. C. Loy, "Open-vocabulary DETR with conditional matching," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 106–122.
[77] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, "Open-vocabulary object detection using captions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14393–14402.
[78] H. Zhang et al., "A simple framework for open-vocabulary segmentation and detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 1020–1031.
[79] W. Zhang, J. Pang, K. Chen, and C. C. Loy, "K-net: Towards unified image segmentation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 10326–10338.
[80] H. Zhao, X. Puig, B. Zhou, S. Fidler, and A. Torralba, "Open vocabulary scene parsing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2021–2029.
[81] Y. Zhong et al., "Regionclip: Region-based language-image pretraining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16772–16782.
[82] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5122–5130.
[83] B. Zhou et al., "Semantic understanding of scenes through the ADE20K dataset," *Int. J. Comput. Vis.*, vol. 127, pp. 302–321, 2019.
[84] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 696–712.
[85] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 350–368.

Hao Zhou received the BS, MS, and PhD degrees from Harbin Engineering University, Harbin, China, in 2016, 2019, and 2024, respectively. He is currently working as a postdoctoral researcher with the School of Information Science and Technology, Great Bay University, and Tsinghua Shenzhen International Graduate School. His current research interests include trajectory prediction and image segmentation.

Lu Qi received the PhD degree from the Chinese University of Hong Kong, in 2021. He works as a postdoc with Prof. Ming-Hsuan Yang in UC Merced. He obtained the Hong Kong PhD Fellowship in 2017. His research interests include instance-level detection, image generation, and cross-modal pretraining. He was the senior program commitee of AAAI 2023/2024 and area chair of ICLR2024.

Tiancheng Shen received the BS degree from Shandong University, in 2017 and the MS degree from Peking University, in 2020. He is currently working toward the PhD degree with UC Merced. His research interests include computer vision and deep learning, primarily focusing on image segmentation and editing.

Hai Huang received the BS and PhD degrees from the Harbin Institute of Technology, Harbin, China, in 2001 and 2008, respectively. He is a professor with the National Key Laboratory of Science and Technology of Underwater Vehicles, Harbin Engineering University. His current research interests include underwater vehicles and autonomous operation.



Xu Yang (Senior Member, IEEE) received the BE degree in electronic engineering from the Ocean University of China, Qingdao, China, in 2009 and the PhD degree in computer science from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2014. He is an associate professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, CASIA, Beijing. His current research interests include computer vision, graph algorithms, and robotics.

Xiangtai Li received the PhD degree from Peking University, in 2022. He is working as a research fellow with S-Lab, a member of the Multimedia Laboratory of NTU (MMLab@NTU), Nanyang Technological University. His research interests include computer vision and machine learning with a focus on scene understanding, segmentation, video understanding, and multi-modal learning. Several of his works have been published in top-tier conferences and journals. He regularly reviews top-tier conferences and journals, including CVPR, ICCV, ICLR, ICML, NeurIPS, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, and *International Journal of Computer Vision*.

Ming-Hsuan Yang (Fellow, IEEE) is a professor of electrical engineering and computer science with the University of California, Merced. He serves as a program co-chair of the ICCV in 2019. He received the Best Paper Award at ICML 2024, Longuet-Higgins Prize at CVPR 2023, Best Paper Honorable Mention at CVPR 2018, Nvidia Pioneer Research Award in 2018 and 2017, NSF CAREER award in 2012 and Google Faculty Award in 2009. He is a fellow of the ACM.