Semantic Alignment in Hyperbolic Space for Open-Vocabulary Semantic Segmentation

Hoang M. Truong Hai Nguyen-Truong Dang Huynh*
Fulbright University Vietnam

https://tmhoanggg.github.io/HyRo/

Abstract

Open-vocabulary semantic segmentation requires adapting image-level vision-language models such as CLIP to dense pixel-level prediction, which is challenging due to the mismatch between hierarchical structure and semantic alignment in the embedding space. While recent works leverage hyperbolic geometry to model hierarchical relationships, they align embeddings across hierarchical levels but overlook semantic misalignment among embeddings within the same level. In this work, we propose HyRo, a hyperbolic fine-tuning framework that decouples hierarchical and semantic alignment in the Poincaré ball model. HyRo aligns hierarchical levels by adjusting the hyperbolic radius and refines semantic relationships through angular alignment using an orthogonal transformation that theoretically preserves the hyperbolic radius. Experiments on standard open-vocabulary semantic segmentation benchmarks demonstrate that HyRo achieves state-of-the-art performance over prior methods.

1. Introduction

Open-vocabulary semantic segmentation aims to assign each pixel in an image to a semantic category specified by textual descriptions, including categories unseen during training. To tackle this challenging task, vision-language foundation models such as CLIP [32] have emerged as powerful tools, as they are trained on large-scale image–text pairs and capture rich semantic alignments between visual and linguistic modalities. Consequently, CLIP is widely adopted for open-vocabulary learning by representing class names as text embeddings, effectively enabling language-driven classification for downstream tasks. However, CLIP is originally pre-trained for image-level representation learning, and thus requires additional adaptation to produce dense, pixel-level predictions for semantic segmentation.

*Corresponding author: dang.huynh@fulbright.edu.vn

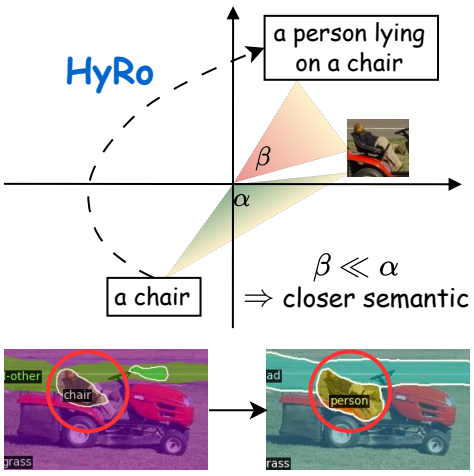

Figure 1. HyRo rotates the text embeddings to achieve a smaller angle (β) relative to the target image embeddings compared to the initial angle (α). This geometric adjustment enables the model to resolve semantic ambiguities and produce more accurate, fine-grained segmentations.

To achieve this adaptation, early approaches [23, 38, 40, 41, 45] decoupled the problem into two stages, utilizing mask proposal generators followed by a pre-trained CLIP model to classify the resulting regions. While such methods benefit from explicit object localization and clear boundaries, the reliance on pre-defined mask proposals introduces a closed-set bias that can limit scalability and generalization to unseen categories. Recent methods [5, 37, 43] instead directly fine-tune CLIP within a shared representation space, enabling dense pixel-level predictions while preserving the semantic richness of the original CLIP representation.

Notably, hyperbolic space has recently gained attention for its ability to capture intrinsic hierarchical structures. Several works [6, 29] incorporate hierarchical representations into CLIP-like vision–language models by embedding

visual and textual entities in hyperbolic space, enabling the model to better capture hierarchical relationships between modalities. In contrast, a recent advancement in open-vocabulary semantic segmentation, HyperCLIP [31], observes that the hierarchy of image embeddings shifts from an image-level to a pixel-level during fine-tuning. To explain this, they integrate the Poincaré ball model to adjust the hyperbolic radius of text embeddings to match this pixel-level granularity. However, a critical limitation of this approach is the absence of explicit constraints on the semantic alignment between embeddings, which is essential for effective open-vocabulary learning. Semantic relationships can be represented in hyperbolic space through geometric structures such as imaginary cones, as explored in prior work [6, 29, 36], or equivalently through angular relationships at the origin.

To address this key limitation, we introduce HyRo (Hyperbolic Rotation), a novel fine-tuning strategy that explicitly disentangles hierarchical and semantic information in hyperbolic space. While prior work primarily focuses on the *hierarchical geometry* of the embedding space, our approach explicitly optimizes the *semantic* aspect. HyRo is motivated by the observation that hierarchy and semantics correspond to distinct geometric properties, as hierarchy is encoded by radial distance, while semantic similarity is reflected in angular orientation. Figure 1 illustrates a failure case in prior work and demonstrates how HyRo resolves it. Specifically, prior work [31] fine-tunes solely via radius scaling and overlooks semantic orientation. This potentially leads to semantically dissimilar concepts being embedded at similar radii but incorrect angles, degrading the discrimination between categories. For instance, in an image depicting a person lying on a chair, HyperCLIP loses semantic understanding and misclassifies both entities as a single "chair" object. HyRo addresses this by rotating the initial embedding to achieve improved angular alignment with related embeddings. After this rotation, the model better segments the scene, correctly recognizing the person as the distinct foreground subject.

We theoretically demonstrate that orthogonal transformations in the Poincaré ball model act as ideal rotation operations (see Sec. 3.2). These transformations enable angular adjustment of embeddings, refining semantic alignment without altering their radius, thereby preserving the hierarchical structure established by prior radius tuning. By decoupling these geometric properties, HyRo enables CLIP to simultaneously maintain the pixel-level granularity required for segmentation and enhance the semantic discrimination necessary for open-vocabulary generalization. Overall, our approach highlights the importance of explicitly modeling both hierarchical and angular relationships when adapting vision–language representations for dense open-vocabulary prediction.

Our contributions are summarized as follows:

- We propose HyRo, an orthogonal transformation strategy that adjusts angular relationships in hyperbolic space while preserving hierarchical structure, with theoretical justification provided.
- We introduce a hyperbolic fine-tuning framework that decouples hierarchical alignment (radius) and semantic refinement (angle) for open-vocabulary semantic segmentation.
- We demonstrate that HyRo achieves state-of-the-art performance on standard benchmarks, validating the effectiveness of our geometric approach.

2. Related Works

2.1. Open-Vocabulary Semantic Segmentation.

Previous works [5, 37, 41, 43, 45] on open-vocabulary semantic segmentation aim to adapt foundation vision–language models pre-trained on large-scale datasets, such as CLIP [32], for pixel-level prediction. This adaptation requires transforming high-level vision–language representations into dense predictions, which poses a key challenge for the task. Early approaches directly fine-tune the encoders of CLIP; however, prior methods [41, 42, 45] observe overfitting to seen classes, as such fine-tuning can degrade the generalization ability of CLIP. As a result, many works [12, 17, 23, 40] freeze the CLIP encoders to preserve their generalization.

More recent studies [5, 37, 43] show that fine-tuning CLIP within cost aggregation–based frameworks can alleviate this issue. Beyond Euclidean representations, several works [30, 31] further explore alternative geometric spaces for open-vocabulary semantic segmentation, providing promising insights into modeling complex semantic structures across vision and language. While existing hyperbolic methods primarily optimize the radius of embeddings to capture hierarchical relationships, the angle governing semantic similarity remains largely unexplored. We address this gap by explicitly modeling angular relationships while preserving the hierarchical structure.

2.2. Hyperbolic Deep Learning.

Hyperbolic geometry provides a natural framework for representing hierarchical structures due to its exponential volume growth, which mirrors the branching nature of tree-like hierarchies [18]. In contrast, Euclidean space exhibits polynomial volume growth, making it less suitable for embedding hierarchical data with low distortion [26]. This property has led to widespread adoption of hyperbolic spaces in natural language processing [7, 20, 28, 35], where concepts are naturally organized in taxonomies and can be embedded as tree graphs with minimal distortion [33, 34].

Visual data also exhibits inherent hierarchical struc-

tures [1], spanning multiple levels from pixels and patches to objects and scenes. This observation has motivated recent works [13, 19, 31] to leverage hyperbolic geometry for visual understanding. In the context of vision-language learning, hyperbolic spaces offer a unified framework to model the hierarchical relationships between visual and textual modalities. Recent approaches [6, 29] hypothesize that textual concepts are generally more abstract than visual features and employ hyperbolic embeddings to capture this semantic hierarchy. In this work, we adopt the Poincaré ball model to bridge the hierarchical gap and semantic relationships between vision and language embeddings, enabling more effective alignment for open-vocabulary semantic segmentation.

3. Methodology

We introduce HyRo, a method for open-vocabulary semantic segmentation that improves semantic alignment. HyRo refines semantic relationships through controlled angular transformations in hyperbolic space while preserving hierarchical information encoded in feature radii. Specifically, we first review essential background on the Poincaré ball model in Sec. 3.1, then present our proposed hyperbolic rotation module in Sec. 3.2. Finally, we describe the overall architecture in Sec. 3.3, and our decoder in Sec. 3.4.

3.1. Background: The Poincaré Ball Model

We briefly review the hyperbolic geometry concepts required for our method, focusing on the Poincaré ball model. The m-dimensional Poincaré ball with curvature $-c$ ($c > 0$) is defined as $\mathbb{D}_c^m := \{\mathbf{x} \in \mathbb{R}^m \mid c\|\mathbf{x}\|^2 < 1\}$, which has radius $1/\sqrt{c}$. Its exponential volume growth makes it naturally suited for representing hierarchical relationships [15, 28].

Angle at the origin. A key property of the Poincaré ball model is its conformality: angles measured at the origin are preserved from Euclidean space. Therefore, the angle α between two points $\mathbf{x}, \mathbf{y} \in \mathbb{D}_c^n$ at the origin is given by:

$$\cos(\alpha) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|}, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product.

Möbius matrix-vector multiplication. For $\mathbf{M} \in \mathbb{R}^{m \times m}$ and $\mathbf{x} \in \mathbb{D}_c^m$ such that $\mathbf{Mx} \neq \mathbf{0}$, denoting $\widetilde{\mathbf{x}} = \mathbf{Mx}$, the Möbius matrix-vector multiplication is defined as:

$$\mathbf{M} \otimes_c \mathbf{x} = \frac{1}{\sqrt{c}} \tanh\left(\frac{\|\widetilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \tanh^{-1}\left(\sqrt{c}\|\mathbf{x}\|\right) \right) \frac{\widetilde{\mathbf{x}}}{\|\widetilde{\mathbf{x}}\|}. \tag{2}$$

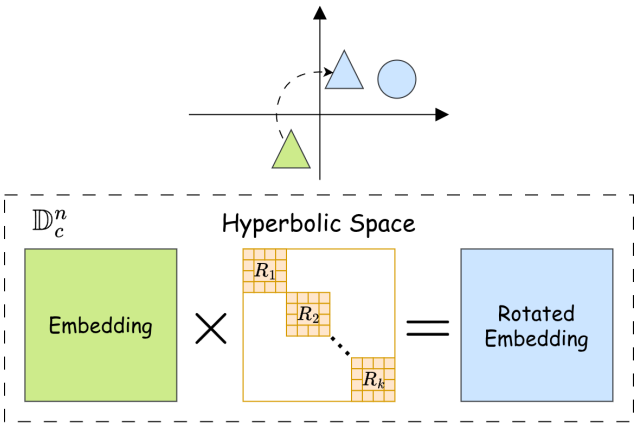

Figure 2. Overview of HyRo. Embeddings are rotated around the origin in hyperbolic space using an orthogonal block matrix to minimize the angle between visual and textual features while preserving their hyperbolic radii, thereby enhancing cross-modal semantic alignment.

Exponential and logarithmic maps. Hyperbolic space is a curved manifold where standard Euclidean operations (e.g., addition, linear transformations) are not directly applicable. To bridge this gap, we use exponential and logarithmic maps that convert between Euclidean tangent space and the hyperbolic manifold.

The exponential map $\exp_{\mathbf{x}}^{\mathbb{D},c}$ takes a Euclidean vector $\mathbf{v}$ from the tangent space at a point $\mathbf{x} \in \mathbb{D}_c^m$ and projects it onto the Poincaré ball along a geodesic (the hyperbolic equivalent of a straight line). Conversely, the logarithmic map $\log_{\mathbf{x}}^{\mathbb{D},c}$ maps a point $\mathbf{y} \in \mathbb{D}_c^m$ on the manifold back to the tangent space at $\mathbf{x}$. At the origin $\mathbf{0}$, these maps have particularly simple forms:

$$\exp_{\mathbf{0}}^{\mathbb{D},c}(\mathbf{v}) = \frac{1}{\sqrt{c}} \tanh\left(\sqrt{c}\|\mathbf{v}\|\right) \frac{\mathbf{v}}{\|\mathbf{v}\|}, \tag{3}$$

$$\log_{\mathbf{0}}^{\mathbb{D},c}(\mathbf{y}) = \frac{1}{\sqrt{c}} \tanh^{-1}\left(\sqrt{c}\|\mathbf{y}\|\right) \frac{\mathbf{y}}{\|\mathbf{y}\|}. \tag{4}$$

These maps act as nonlinear scaling operations that preserve direction while mapping between spaces. This enables us to perform operations in the Euclidean tangent space and then map the results back to the hyperbolic manifold.

3.2. Semantic Refinement via Rotation Matrix

Intuitively, rotating embeddings around the origin modifies their angular relationships (semantic similarity) while preserving their radial distance (hierarchical level). We therefore introduce an orthogonal rotation matrix to refine the angular alignment between feature embeddings while preserving their hyperbolic radii, thereby improving semantic alignment across modalities. This transformation is depicted in Fig. 2. Specifically, given a hyperbolic embedding

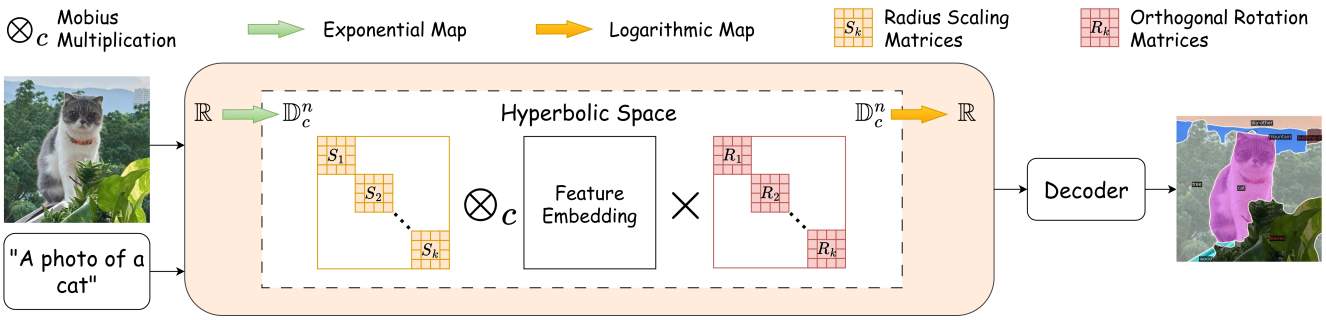

Figure 3. **Overall architecture of HyRo.** Given image and text inputs, Euclidean embeddings are mapped to the Poincaré ball via the exponential map. HyRo then decouples alignment into two stages: (1) *Hierarchical Adjustment* using block-diagonal radius scaling matrices to align granularity, and (2) *Semantic Refinement* using orthogonal rotation matrices to adjust angular relationships without altering the radius. The refined hyperbolic embeddings are mapped back to the tangent space for decoding.

$\mathbf{q} \in \mathbb{D}_c^d$ and an orthogonal matrix $\mathbf{R}$, the resulting embedding $\mathbf{v} \in \mathbb{D}_c^d$ after adjusting the angle is

$$\mathbf{v} = \mathbf{R}\mathbf{q}. \tag{5}$$

Since the Poincaré ball model is conformal, angles measured at the origin coincide with their Euclidean counterparts.

To ensure strict orthogonality, we employ the Cayley transform [4] to parameterize a learnable unconstrained matrix $\mathbf{\Theta} \in \mathbb{R}^{n\times n}$:

$$\mathbf{A} = \mathbf{\Theta} - \mathbf{\Theta}^\top, \tag{6}$$

$$\mathbf{R} = (\mathbf{I} + \mathbf{A})\,(\mathbf{I} - \mathbf{A})^{-1}, \tag{7}$$

where $\mathbf{A}$ is a skew-symmetric matrix and $\mathbf{I}$ denotes the identity matrix. This parameterization guarantees that $\mathbf{R}$ satisfies the orthogonality constraint $\mathbf{R}^\top\mathbf{R} = \mathbf{I}$.

For computational efficiency, we adopt a block-diagonal structure inspired by [30]. A naive application of the Cayley transform to a full $d\times d$ matrix incurs an $\mathcal{O}(d^3)$ matrix inversion cost, which becomes prohibitive for high-dimensional CLIP embeddings. Furthermore, CLIP encoders produce embeddings of different dimensionalities across modalities (*e.g.*, 768 for vision and 512 for text in ViT-B/16), making a single shared full matrix infeasible. By decomposing $\mathbf{R}$ into $K_\mathbf{R} = d/n$ independent blocks:

$$\mathbf{R} = \mathrm{diag}\left(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_{K_\mathbf{R}}\right), \tag{8}$$

where each block $\mathbf{R}_i \in \mathbb{R}^{n\times n}$ is constructed via the Cayley transform, the inversion cost reduces from $\mathcal{O}(d^3)$ to $K_\mathbf{R}\cdot\mathcal{O}(n^3) = \mathcal{O}(d^3/n^2)$, while also enabling parallelization across blocks. Note that setting $n = d$ (*i.e.*, $K_\mathbf{R} = 1$) recovers a full unconstrained rotation matrix, while smaller n imposes stronger structural constraints with fewer learnable parameters. We ablate this choice in Sec. 4.3.

Theoretical Justification. We formally justify that applying an orthogonal transformation to a point in the Poincaré ball induces a rotation of the angle at the origin without altering the hyperbolic radius.

Let $\mathbf{x} \in \mathbb{D}_c^d$. Its representation in the tangent space at the origin is $\mathbf{v}_\mathbf{x} = \log_\mathbf{0}^c(\mathbf{x})$. Applying an orthogonal matrix $\mathbf{R} \in \mathrm{O}(d)$ to this tangent vector yields $\mathbf{v}'_\mathbf{x} = \mathbf{R}\mathbf{v}_\mathbf{x}$. Since $\mathbf{R}$ preserves the Euclidean norm, we have $\|\mathbf{v}'_\mathbf{x}\| = \|\mathbf{v}_\mathbf{x}\|$.

Mapping $\mathbf{v}'_\mathbf{x}$ back to the manifold via the exponential map:

$$\begin{aligned}
\mathbf{x}' &= \exp_\mathbf{0}^c(\mathbf{v}'_\mathbf{x}) \\
&= \frac{1}{\sqrt{c}}\tanh\left(\sqrt{c}\|\mathbf{v}'_\mathbf{x}\|\right)\frac{\mathbf{v}'_\mathbf{x}}{\|\mathbf{v}'_\mathbf{x}\|} \\
&= \mathbf{R}\left[\frac{1}{\sqrt{c}}\tanh\left(\sqrt{c}\|\mathbf{v}_\mathbf{x}\|\right)\frac{\mathbf{v}_\mathbf{x}}{\|\mathbf{v}_\mathbf{x}\|}\right] \\
&= \mathbf{R}\exp_\mathbf{0}^c(\mathbf{v}_\mathbf{x}) \\
&= \mathbf{R}\mathbf{x}.
\end{aligned} \tag{9}$$

This confirms that the transformation simplifies to a direct matrix multiplication on the coordinates. Furthermore, the angle α' between the rotated point $\mathbf{x}'$ and a target $\mathbf{y}$ becomes:

$$\cos(\alpha') = \frac{\langle \mathbf{x}', \mathbf{y}\rangle}{\|\mathbf{x}'\|\|\mathbf{y}\|} = \frac{\langle \mathbf{R}\mathbf{x}, \mathbf{y}\rangle}{\|\mathbf{x}\|\|\mathbf{y}\|}. \tag{10}$$

Crucially, since $\|\mathbf{x}'\| = \|\mathbf{R}\mathbf{x}\| = \|\mathbf{x}\|$ (due to orthogonality), the hyperbolic radius $\mathrm{Rad}_{\mathbf{x}'} = \mathrm{Rad}_\mathbf{x}$ remains unchanged. Thus, HyRo enables independent control over semantic alignment (angle) and hierarchical depth (radius).

3.3. Overall Architecture

Figure 3 illustrates our overall architecture. We first map the input Euclidean feature embedding $\mathbf{z} \in \mathbb{R}^d$ into the Poincaré ball model. We utilize the origin ($\mathbf{0}$) as the reference point, as it represents the root of the hierarchy [15]. The mapping is performed via the exponential map:

$$\mathbf{h} = \exp_\mathbf{0}^{\mathbb{D},c}(\mathbf{z}), \tag{11}$$

where d is the feature dimension, and $\mathbf{h} \in \mathbb{D}_c^d$ is the resulting hyperbolic embedding.

Once projected, we adjust the hyperbolic radius of $\mathbf{h}$ using a learnable diagonal matrix $\mathbf{S}$ introduced in [31]. In hyperbolic space, linear transformations are realized via Möbius matrix-vector multiplication. The adjusted embedding $\mathbf{q} \in \mathbb{D}_c^d$ is obtained as:

$$\mathbf{q} = \mathbf{S} \otimes_c \mathbf{h}. \tag{12}$$

To balance expressiveness and computational efficiency, we impose a block-diagonal structure on $\mathbf{S}$. Specifically, $\mathbf{S}$ is composed of $K_\mathbf{S}$ independent sub-matrices:

$$\mathbf{S} = \text{diag}(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{K_\mathbf{S}}), \tag{13}$$

where each block $\mathbf{S}_k \in \mathbb{R}^{b \times b}$ is a learnable parameter, and the number of blocks is given by $K_\mathbf{S} = d/b$. This structure allows the model to learn distinct scaling factors for different feature subspaces.

Then, given the radius-adjusted embedding $\mathbf{q} \in \mathbb{D}_c^d$ obtained from Eq. (12), we apply an orthogonal matrix transformation as in Eq. (5) to obtain the semantically refined embedding $\mathbf{v} \in \mathbb{D}_c^d$.

After refinement, the embeddings are mapped back to the Euclidean tangent space using the logarithmic map to facilitate compatibility with the decoder:

$$\mathbf{e} = \log_\mathbf{0}^{\mathbb{D},c}(\mathbf{v}). \tag{14}$$

In general, given a Euclidean feature embedding $\mathbf{x} \in \mathbb{R}^d$ from the visual encoder, we apply semantic refinement in hyperbolic space through the following procedure:

$$\mathbf{x}' = \log_\mathbf{0}^{\mathbb{D},c}\left(\mathbf{R} \cdot \left(\mathbf{S} \otimes_c \exp_\mathbf{0}^{\mathbb{D},c}(\mathbf{x})\right)\right), \tag{15}$$

where $\exp_\mathbf{0}^{\mathbb{D},c}$ maps the feature to the Poincaré ball, the scaling matrix $\mathbf{S}$ adjusts the hyperbolic radius, the rotation $\mathbf{R}$ refines angular relationships while preserving hyperbolic radii (and thus hierarchical information), and $\log_\mathbf{0}^{\mathbb{D},c}$ projects back to Euclidean space for subsequent processing.

After getting the refined embeddings, we use the decoder (see Sec. 3.4) to generate dense pixel-level predictions.

3.4. Cost Aggregation Decoder

We adopt the cost aggregation decoder introduced by CAT-Seg [5] to adapt CLIP [32] for dense open-vocabulary segmentation. Instead of directly predicting pixel labels, the decoder aggregates similarity scores between visual and textual embeddings to produce pixel-level predictions.

Dense visual and textual embeddings are first extracted using the CLIP encoders [32]. These embeddings are refined through the hyperbolic refinement stage described in Sec. 3.3. The resulting refined embeddings are denoted as $D^V \in \mathbb{R}^{(H\times W)\times d}$ and $D^L \in \mathbb{R}^{N_C \times d}$, where $H \times W$ is the spatial resolution and $N_\mathcal{C}$ is the number of candidate classes. Then, a cost volume $C \in \mathbb{R}^{(H\times W)\times N_C}$ is constructed by computing cosine similarity between pixel and text embeddings:

$$C(i,n) = \frac{D^V(i) \cdot D^L(n)}{\|D^V(i)\| \, \|D^L(n)\|}. \tag{16}$$

The cost volume is then projected into a higher-dimensional embedding space using a convolution layer, producing $F \in \mathbb{R}^{(H\times W)\times N_C \times d_F}$.

To exploit both spatial and semantic relationships, cost aggregation is decomposed into spatial and class aggregation modules, which enforce spatial consistency and suppress background noise. For each class n, spatial aggregation refines the cost map using Swin Transformer blocks [24] with window and shifted-window attention:

$$F'(:,n) = \mathcal{T}^{sa}(F(:,n)). \tag{17}$$

Class aggregation models relationships among category tokens. A transformer layer without positional encoding is applied across class tokens:

$$F''(i,:) = \mathcal{T}^{ca}(F'(i,:)). \tag{18}$$

A linear transformer is used for efficiency when handling a large number of classes.

To further improve aggregation, the original visual and textual embeddings (D^V and D^L) are used as guidance. Projected embeddings are concatenated with cost features during attention:

$$F'(:,n) = \mathcal{T}^{sa}([F(:,n); \mathcal{P}^V(D^V)]) \tag{19}$$

$$F''(i,:) = \mathcal{T}^{ca}([F'(i,:); \mathcal{P}^L(D^L)]). \tag{20}$$

Finally, a lightweight upsampling decoder produces high-resolution predictions. The aggregated cost volume is progressively upsampled and fused with intermediate features from the CLIP image encoder (e.g., layers 4 and 8 in ViT-B/16). Each stage performs bilinear upsampling, concatenation with upsampled CLIP features, and a 3×3 convolution. Starting from 24×24 features, the decoder progressively upsamples to 96×96 before the prediction head outputs the final segmentation map.

Training Objectives. Following standard practice in open-vocabulary semantic segmentation methods [31, 37, 41], we train our model using pixel-wise cross-entropy loss. Given the predicted segmentation logits $\hat{Y} \in \mathbb{R}^{H\times W\times N_C}$ and ground truth labels $Y \in \{1, \ldots, N_C\}^{H\times W}$, the loss is defined as:

$$\mathcal{L} = -\frac{1}{H \times W} \sum_{i=1}^{H\times W} \log \frac{\exp(\hat{Y}_{i,y_i})}{\sum_{n=1}^{N_C} \exp(\hat{Y}_{i,n})}, \tag{21}$$

Model	VLM	Additional Backbone	Fine-tuning Space	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
ZS3Net [2]	-	ResNet-101	E	-	-	-	19.4	38.3	-
LSeg [21]	CLIP ViT-B/32	ResNet-101	E	-	-	-	-	47.4	-
ZegFormer [12]	CLIP ViT-B/16	ResNet-101	E	4.9	9.1	16.9	42.8	86.2	62.7
ZSseg [39]	CLIP ViT-B/16	ResNet-101	E	7.0	-	20.5	47.7	88.4	-
OpenSeg [16]	ALIGN	ResNet-101	E	4.4	7.9	17.5	40.1	-	63.8
OVSeg [22]	CLIP ViT-B/16	ResNet-101c	E	7.1	11.0	24.8	53.3	92.6	-
ZegCLIP [46]	CLIP ViT-B/16	-	E	-	-	-	41.2	93.6	-
SED [37]	CLIP ConvNeXt-B	-	E	11.4	<u>18.6</u>	<u>31.6</u>	**57.3**	94.4	-
SAN [41]	CLIP ViT-B/16	Side Adapter	E	10.1	12.6	27.5	53.8	94.0	-
HyperCLIP* [31]	CLIP ViT-B/16	-	H	<u>11.9</u>	18.2	**31.7**	<u>57.1</u>	94.9	**77.1**
HyRo (Ours)	CLIP ViT-B/16	-	H	**12.0**	**18.9**	31.2	**57.3**	**95.0**	<u>76.7</u>

Table 1. **Comparison with state-of-the-art methods on standard benchmarks.** The best-performing results are presented in bold, while the second-best results are underlined. "E": Euclidean Space. "H": Hyperbolic Space.

where y_i denotes the ground truth class label at spatial position i.

During training, we only fine-tune the hyperbolic transformation parameters (radius scaling matrices $\mathbf{S}$ and rotation matrices $\mathbf{R}$), while keeping the CLIP encoders frozen to preserve their generalization ability.

4. Experiments

4.1. Experimental Setup

Dataset. We evaluate our approach under the standard open-vocabulary semantic segmentation protocol following prior work [5]. The model is trained on COCO-Stuff [3] and evaluated on multiple benchmark datasets, including ADE20K [44], PASCAL VOC [14], and PASCAL-Context [27]. ADE20K contains 20K training images and 2K validation images and is evaluated using two label sets: A-150, which contains the 150 most frequent classes, and A-847, which covers all 847 categories [12]. PASCAL-Context includes 5K images for training and validation, with results reported on both the full 459-class setting (PC-459) and the 59 most frequent classes (PC-59). PASCAL VOC contains 20 foreground object classes and one background class; we report results on PAS-20 following standard practice. Additionally, PAS-20^b is reported, where background labels are defined as categories present in PC-59 but absent from PAS-20, following [17].

Metrics. We use mean Intersection-over-Union (mIoU) for evaluation, consistent with prior open-vocabulary semantic segmentation works [16, 22, 37, 41].

Implementation Details. We fine-tune the CLIP [32] ViT-B/16 model following the training protocol of [31]. AdamW [25] is used as the optimizer, with a learning rate of 2×10^{-4} for our hyperbolic transformation parameters and 1×10^{-6} for the CLIP encoder. We set the block size to 256 for both the diagonal radius-scaling matrix and the orthogonal rotation matrix. Following common practice in open-vocabulary segmentation [5, 31], we use a small batch size of 8 (distributed across 8 NVIDIA A100 GPUs, 1 sample per GPU) to preserve CLIP's generalization capability. Training is performed for 40,000 iterations, taking approximately 8 hours.

4.2. Main Results

Quantitative Results. Table 1 compares our method with state-of-the-art open-vocabulary semantic segmentation approaches across multiple benchmarks. HyRo achieves the best performance on four out of six benchmarks, demonstrating the effectiveness of hyperbolic rotation for improving semantic alignment.

The improvements on large-vocabulary benchmarks such as A-847 and PC-459 highlight HyRo's ability to capture fine-grained semantic relationships among many visually similar categories. In particular, HyRo achieves 12.0 mIoU on A-847 and 18.9 mIoU on PC-459. On commonly reported settings, our method obtains 31.2 mIoU on A-150, 57.3 mIoU on PC-59, 95.0 mIoU on PAS-20, and 76.7 mIoU on PAS-20^b matching or surpassing strong baselines [31, 37].

Compared with the prior hyperbolic method HyperCLIP [31], our approach improves performance on several benchmarks, including A-847 (+0.1 mIoU), PC-459 (+0.7 mIoU), PC-59 (+0.2 mIoU), and PAS-20 (+0.1 mIoU). These improvements validate that explicitly learning angular transformations, rather than relying solely on hyperbolic radius scaling, better capture the semantic relationships and the complex hierarchical structure of visual concepts. Overall, these results indicate that refining angular relationships through hyperbolic rotations improves semantic alignment and enables more effective adaptation of vision–language representations for dense open-vocabulary prediction.

Figure 4. Qualitative comparison between HyperCLIP [31] and our method on the A-847 setting of ADE20K. Our approach mitigates several semantic misalignment failures observed in HyperCLIP.

Qualitative Results. Figure 4 presents qualitative comparisons on the ADE20K [44] dataset, which contains 847 semantic categories (A-847) and serves as the most challenging benchmark in our evaluation due to its highly diverse vocabulary.

Compared with the state-of-the-art HyperCLIP [31], our proposed HyRo effectively mitigates several common semantic misalignment issues, yielding more cohesive and accurate segmentation masks. In the first row, the baseline fails to distinguish between the person and the chair, incorrectly labeling both as "chair" and producing noisy, artifact-heavy predictions along the background tree line. HyRo, however, cleanly disentangles these adjacent objects. In the second row, although HyperCLIP produces a cleaner floor segmentation, it exhibits semantic inconsistency by fragmenting visually uniform regions into multiple contradictory labels, whereas our approach maintains spatial coherence. Finally, in the last row, the baseline succumbs to background dominance, over-predicting the "dirt" category across most of the image and completely failing to detect multiple people in the scene. Taken together, these examples highlight HyRo's superior ability to overcome the common failure modes of baseline models, delivering precise and spatially coherent segmentation in highly challenging, open-vocabulary environments.

Attention Visualization. Figure 5 visualizes attention maps of visual embeddings with and without HyRo across three target classes: "person", "building", and "window". Without semantic refinement, the model struggles to ground visual features to the correct class regions, resulting in diffuse attention that bleeds into semantically unrelated background areas. After applying HyRo, the angular refinement in hyperbolic space explicitly improves the semantic correspondence between visual and textual embeddings. This causes attention to concentrate sharply on the precise target regions, effectively suppressing background noise. These visualizations confirm that modeling angular relationships in hyperbolic space is an effective mechanism for enhancing cross-modal semantic alignment.

(a) Image (b) Without HyRo (c) With HyRo (Ours)

Figure 5. Attention map visualization for target classes "person", "building", and "window" (top to bottom). Without HyRo, attention is diffusely spread across semantically irrelevant regions. With HyRo, attention is more concentrated on the target class regions, indicating improved semantic alignment.

4.3. Ablation Study

Component Importance Analysis. We ablate the contributions of radius scaling and rotation in Tab. 2. While independently applying either hierarchical scaling (radius) or semantic refinement (rotation) improves the baseline, their combination in HyRo yields the best performance across nearly all benchmarks. The rotation module provides substantial gains on large-scale sets like A-847, indicating that angular alignment is the primary driver for open-vocabulary generalization. Meanwhile, radius scaling further enhances performance, confirming that positioning embeddings at the correct hierarchical level is essential for fine-grained separability. These results confirm that hierarchical positioning (radius) and angular alignment (rotation) play complementary roles in improving semantic consistency for open-vocabulary segmentation. Specifically, while radius captures abstraction, rotation prevents semantic collapse among closely related categories.

Choice of Curvature. We evaluate the impact of curvature c on performance in Tab. 3. Results show that a "gentler" curvature ($c = 0.01$) consistently outperforms higher values across most benchmarks. While a higher curvature ($c = 1.0$) improves results on specific sets like PAS-20^b, it appears to distort the pre-trained Euclidean CLIP space too aggressively for the diverse label space of A-847. Consequently, $c = 0.01$ is chosen as the default to preserve CLIP's zero-shot generalization while enabling hierarchical adjustment. We do not explore values smaller than 0.01, as extremely small curvature would make the space nearly Euclidean and diminish the benefits of hyperbolic geometry.

Radius	Rotation	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
✗	✗	11.4	17.6	29.8	56.2	94.8	75.9
✓	✗	11.9	18.2	**31.7**	57.1	94.9	76.4
✗	✓	11.6	18.3	30.6	56.5	**95.4**	**76.7**
✓	✓	**12.0**	**18.9**	31.2	**57.3**	95.0	**76.7**

Table 2. Ablation results on the contribution of Radius Scaling and Rotation modules.

c	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
0.01	**12.0**	18.9	**31.2**	**57.3**	95.0	76.7
0.05	11.4	**19.0**	30.1	56.5	**95.1**	76.3
0.1	11.4	17.6	30.4	55.9	**95.1**	75.9
1.0	11.2	17.6	30.1	57.2	94.8	**77.8**

Table 3. Ablation results on the choice of curvature c.

n	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
32	11.4	17.6	29.8	56.2	94.8	75.9
128	11.6	18.3	30.6	56.5	**95.4**	**76.7**
256	**12.0**	**18.9**	**31.2**	**57.3**	95.0	**76.7**

Table 4. Ablation results on the size of diagonal rotation blocks n.

Size of Diagonal Rotation Blocks. We evaluate the impact of block size n on capturing semantic transformations in Tab. 4. Performance scales with n, with $n = 256$ reaching the optimal balance of representational capacity and generalization across nearly all benchmarks. While smaller datasets like PAS-20 are less sensitive to block size, the significant gains on A-847 suggest that high-capacity rotation is critical for disentangling fine-grained semantic relationships in large-vocabulary settings. Consequently, $n = 256$ is selected as the default configuration.

5. Conclusion

In this work, we introduce HyRo, a fine-tuning strategy for open-vocabulary semantic segmentation in hyperbolic space that explicitly decouples hierarchical alignment and semantic refinement. By operating in the Poincaré ball model, HyRo first aligns embeddings to appropriate hierarchical levels via radius adjustment and then refines semantic relationships through angular optimization using orthogonal transformations that preserve the radius. This strategy enables more precise alignment between vision and language representations without affecting generalization to unseen categories. Experiments on multiple benchmarks demonstrate that HyRo consistently improves performance over strong baselines. This work highlights the importance of jointly modeling hierarchy and semantics in non-Euclidean spaces and opens new directions for geometric representation learning in dense vision–language tasks.

Future Work. While HyRo demonstrates strong open-vocabulary segmentation performance on static images, extending this hyperbolic geometric approach to more complex open-vocabulary video segmentation settings, such as MOSE [9], MOSEv2 [11], MeViS [8], and MeViSv2 [10], is a promising direction for future work.

References

[1] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987. 3

[2] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. 6

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[4] A. Cayley. Sur quelques propriétés des déterminants gauches. *Journal für die reine und angewandte Mathematik*, 1846(32):119–123, 1846. 4

[5] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4113–4123, 2024. 1, 2, 5, 6

[6] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Ramakrishna Vedantam. Hyperbolic Image-Text Representations. In *International Conference on Machine Learning (ICML)*, 2023. 1, 2, 3

[7] Bhuwan Dhingra, Christopher Shallue, Mohammad Norouzi, Andrew Dai, and George Dahl. Embedding text in hyperbolic spaces. In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, pages 59–69, 2018. 2

[8] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 8

[9] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 8

[10] Henghui Ding, Chang Liu, Shuting He, Kaining Ying, Xudong Jiang, Chen Change Loy, and Yu-Gang Jiang. MeViS: A multi-modal dataset for referring motion expression video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 8

[11] Henghui Ding, Kaining Ying, Chang Liu, Shuting He, Xudong Jiang, Yu-Gang Jiang, Philip HS Torr, and Song Bai. MOSEv2: A more challenging dataset for video object segmentation in complex scenes. *arXiv preprint arXiv:2508.05630*, 2025. 8

[12] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11583–11592, 2022. 2, 6

[13] Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7409–7419, 2022. 3

[14] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–308, 2009. 6

[15] Octavian Ganea, Gary Becigneul, and Thomas Hofmann. Hyperbolic neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 3, 4

[16] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision (ECCV)*, pages 540–557, 2022. 6

[17] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision (ECCV)*, pages 540–557, 2022. 2, 6

[18] Mikhael Gromov. Hyperbolic groups. In *Essays in group theory*, pages 75–263, 1987. 2

[19] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[20] Matthew Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. Inferring concept hierarchies from text corpora via hyperbolic embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3231–3241, 2019. 2

[21] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations (ICLR)*, 2022. 6

[22] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7061–7070, 2023. 6

[23] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7061–7070, 2023. 1, 2

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 5

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Machine Learning (ICML)*, 2019. 6

[26] Jiří Matoušek. On embedding trees into uniformly convex banach spaces. *Israel Journal of Mathematics*, 114(1):221–237, 1999. 2

[27] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and

Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 891–898, 2014. 6

[28] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 3

[29] Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *International Conference on Learning Representations (ICLR)*, 2025. 1, 2, 3

[30] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Yu Huang, Yaoming Wang, and Wei Shen. Parameter-efficient fine-tuning in hyperspherical space for open-vocabulary semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15009–15020, 2025. 2, 4

[31] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Changsong Wen, Yu Huang, Menglin Yang, Feilong Tang, and Wei Shen. Understanding fine-tuning clip for open-vocabulary semantic segmentation in hyperbolic space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4562–4572, 2025. 2, 3, 5, 6, 7

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 1, 2, 5, 6

[33] Frederic Sala, Chris De Sa, Albert Gu, and Christopher Re. Representation tradeoffs for hyperbolic embeddings. In *International Conference on Machine Learning (ICML)*, pages 4460–4469, 2018. 2

[34] Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *Graph Drawing*, pages 355–366, 2012. 2

[35] Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincaré glove: Hyperbolic word embeddings. In *International Conference on Machine Learning (ICML)*, 2019. 2

[36] Ziwei Wang, Sameera Ramasinghe, Chenchen Xu, Julien Monteil, Loris Bazzani, and Thalaiyasingam Ajanthan. Learning visual hierarchies in hyperbolic space for image retrieval. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9924–9934, 2025. 2

[37] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3426–3436, 2024. 1, 2, 5, 6

[38] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2955–2966, 2023. 1

[39] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision (ECCV)*, pages 736–753, 2022. 6

[40] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision (ECCV)*, pages 736–753, 2022. 1, 2

[41] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2945–2954, 2023. 1, 2, 5, 6

[42] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[43] Ziyu Zhao, Xiaoguang Li, Lingjia Shi, Nasrin Imanpour, and Song Wang. Dpseg: Dual-prompt cost volume learning for open-vocabulary semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25346–25356, 2025. 1, 2

[44] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 6, 7

[45] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2

[46] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11175–11185, 2023. 6