DINOde: Continuous Vision-Text Alignment for Open-Vocabulary Semantic Segmentation

Sung-Hoon Yoon[1]*, Hoyong Kwon[2]*, Changgyoon Oh[2]*, and Kuk-Jin Yoon[2]

[1] Multimodal Intelligence and Perception Lab., DGIST, Republic of Korea
shyoon@dgist.ac.kr
[2] Visual Intelligence Lab., KAIST, Republic of Korea
{kwonhoyong3, changgyoon, kjyoon}@kaist.ac.kr

Abstract. Open-vocabulary semantic segmentation (OVSS) leverages textual semantics to segment objects beyond predefined categories. While the self-supervised model DINOv3 provides strong structured visual representations, its lack of native textual alignment hinders direct application to OVSS. To bridge this gap, we propose DINOde, an ODE-based framework that continuously aligns CLIP text embeddings to the DINO visual manifold. Our approach employs two complementary components: (i) *Semantic Text Flow* (STF), which evolves text embeddings toward the DINO manifold through a continuous ODE trajectory, and (ii) *Global Context Flow* (GCF), which progressively refines the holistic image representation carried by DINO's `CLS` token. To preserve the hyperspherical geometry of the feature space during this evolution, we further introduce *Velocity Tangent Projection*, which constrains the learned velocity field to the tangent space through projection. By modeling alignment as a continuous trajectory, DINOde avoids the manifold entanglement inherent in discrete MLP projections and yields more robust cross-modal alignment. Extensive experiments demonstrate that DINOde consistently outperforms existing methods and achieves state-of-the-art performance across multiple OVSS benchmarks. The code is available at `https://github.com/yoon307/DINOde`.

Keywords: Open-Vocabulary · Semantic Segmentation · ODE

1 Introduction

Semantic segmentation [12, 15, 26, 47, 52, 53, 60, 73] has become a cornerstone task in computer vision, ranging from autonomous systems to medical image analysis. However, real-world deployment remains challenging due to the inherent open nature of the environment, where unseen categories frequently emerge. Consequently, recent research has pivoted toward open-vocabulary semantic segmentation (OVSS), which aims to segment unseen classes.

* Equal contribution.

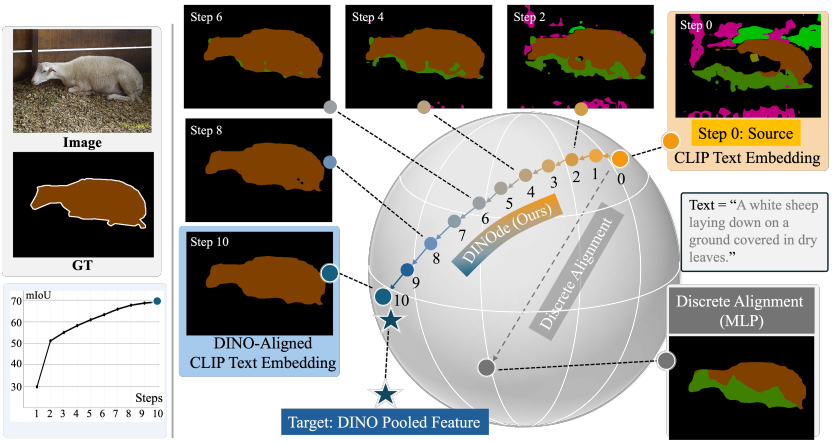

Fig. 1: Overview of the DINOde framework for OVSS. We propose a continuous alignment strategy that bridges the gap between text embeddings and DINO visual features. The illustration depicts how a text embedding is progressively transformed into a DINO-aligned text embedding through an ODE-based trajectory on the unit sphere, resulting in step-by-step refinement of the semantic segmentation.

The emergence of large-scale vision-language pretraining such as the Contrastive Language-Image Pre-training (CLIP [45]) model has significantly advanced this direction by providing a robust alignment between visual features and textual semantics. Early works such as LSeg [33] and TCL [9] employ static contrastive losses to establish direct pixel-text correspondences, while others like MaskCLIP [75] extract dense pseudo-masks directly from frozen encoders for training-free inference. More recently, frameworks have introduced hierarchical region-level grouping as seen in GroupViT [62], or integrated specialized mask-adapted segmentation modules like OVSeg [34] and FC-CLIP [72] to enhance spatial precision. However, CLIP's visual encoder is primarily optimized for global image-text alignment, often yielding coarse and spatially entangled representations that are suboptimal for dense prediction tasks.

Meanwhile, self-supervised Vision Transformers (ViT) trained purely on visual data, such as DINO [8] and DINOv2 [44], exhibit superior localization capabilities and have thus gained increasing attention in dense prediction tasks. Nevertheless, as these self-supervised models operate exclusively within the visual modality, their representation spaces remain unaligned with textual semantics, rendering them incapable of direct open-vocabulary reasoning.

In line with this, the recent work *dino.txt* [24] aligns a frozen DINOv2 backbone with a newly trained text encoder for open-vocabulary segmentation via locked image-text alignment. However, it requires training a text encoder from scratch on large-scale paired data and demands extensive computational resources for vision-text alignment. Another concurrent work, Talk2DINO [3], leverages DINOv2's self-attention heads to selectively align the most relevant

visual regions with textual concepts. While this approach aligns DINOv2 spatial tokens with CLIP text embeddings without fine-tuning the backbones, it simply bridges the two feature spaces using a non-linear mapping function.

Recently, DINOv3 [51] has enabled the preservation of high-quality spatially detailed features through its Gram anchoring regularizer. Notably, simply aligning DINOv3 with the CLIP text embedding space using an MLP provides a surprisingly strong baseline with competitive performance. While aligning two frozen modalities is efficient and straightforward to optimize, MLP-based schemes that attempt to approximate this complex cross-modal transformation through a single-step instantaneous mapping often fail to preserve the topological relationships (e.g., semantic proximity between 'cat' and 'dog') between the representation manifolds. We experimentally observe that MLP-based alignment leads to suboptimal performance, primarily because it adopts a manifold-agnostic Euclidean shortcut that disregards the intrinsic curvature of the data. This structural limitation induces semantic entanglement, wherein the neighborhood relationships from the original space are distorted within target space.

To address these limitations, we introduce DINOde, a continuous ordinary differential equation (ODE)-based vision-text alignment framework. Instead of enforcing a rigid one-step projection, our model learns a smooth trajectory that gradually transforms the textual manifold into DINOv3's representations, as shown in Fig. 1. Furthermore, by incorporating tangent projection on the hypersphere, we ensure the learned velocity field strictly adheres to geometric constraints, enabling a smooth and manifold-preserving flow. Here, we align both text embeddings and DINOv3's `CLS` token through separate ODE trajectories, maximizing global-local semantic coherence. As a result, DINOde demonstrates state-of-the-art OVSS performance with only image-text pairs.

The contributions of our method can be summarized as follows:

- We introduce **DINOde**, a novel ODE-based alignment framework that continuously transforms CLIP text embeddings toward DINOv3 features, ensuring a smooth and geometry-preserving vision-text mapping.
- By progressively aligning both text embeddings and DINOv3 `CLS` tokens through an ODE-based framework, the proposed DINOde effectively integrates both local and global information.
- We introduce Velocity Tangent Projections to enable geometric-constrained and manifold-preserving ODE learning.
- Extensive experiments across multiple benchmarks and unseen categories demonstrate the effectiveness and generalizability of our approach for open-vocabulary semantic segmentation.

2 Related Work

2.1 VLM-Based Open-Vocabulary Segmentation

Foundational works in Open-Vocabulary Semantic Segmentation (OVSS) predominantly rely on VLMs, such as CLIP [45], for both visual and textual fea-

ture extraction. These approaches generally fall into two categories: decoder-free alignments and conditional decoder frameworks.

Conditional Decoder Frameworks To address the severe localization deficiency of VLM-only features, fully-supervised conditional frameworks [22, 33, 34, 63, 66, 67, 72, 74, 79] integrate the VLM with heavy segmentation decoders trained on base-class pixel annotations. For instance, LSeg [33] appends a dense prediction transformer to the visual encoder to establish direct per-pixel correspondence with textual embeddings, while OVSeg [34] fine-tunes a complex Mask2Former-based decoder to generate class-agnostic mask proposals that are subsequently classified by a mask-adapted CLIP [45]. This highlights an ongoing trade-off between fine-grained localization quality and annotation cost, motivating complementary research into more lightweight paradigms.

Decoder-Free VLM Alignment To avoid heavy, task-specific decoders, researchers often extract dense predictions directly from the VLMs. The training-free scheme [5,25,31,32,40,49,54,55,58,75] leverages off-the-shelf VLMs without any additional training phase or parameter updates. For instance, MaskCLIP [75] alters the self-attention topology of the frozen CLIP [45] visual encoder to extract dense pixel-level features directly, while FreeDA [2] enhances training-free matching by generating diffusion-augmented visual prototypes to enrich semantic representations. Alternatively, similar to weakly-supervised semantic segmentation studies that avoid dense pixel annotations [16, 27–29, 61, 65, 70, 71], the weakly-supervised scheme in OVSS [7, 9, 34, 39, 43, 62, 68] relies solely on image-text pairs to align features without pixel-level annotations. Notably, GroupViT [62] learns hierarchical patch groupings by clustering visual tokens to form semantic regions guided by text, whereas TCL [9] leverages text supervision to generate dense pseudo-masks for cross-modal alignment. While these approaches offer an efficient alternative by alleviating reliance on dense masks, they remain fundamentally bottlenecked by the VLM's visual encoder.

2.2 Self-Supervised Visual Backbones

To compensate for the spatial limitations of VLM encoders, recent works have started utilizing Self-Supervised Vision Models (SSVMs). In particular, the DINO series [8,18,44,51] has shown that it can extract object-centric patch embeddings and maintain fine-grained spatial structures without relying on text supervision. Because of this, a common approach in OVSS is to directly align the spatial features of DINO with the semantic space of CLIP [45].

The effectiveness of SSVMs in dense prediction mainly comes from their emergent object-centric properties. As observed in previous studies [50, 57], the self-attention maps in these models naturally capture semantic boundaries and foreground objects without explicit labels. This localization capability allows SSVMs to act as a spatial anchor, which helps resolve the coarse resolution and grid artifacts commonly found in VLM outputs. Therefore, recent OVSS methods

often adopt a dual-backbone strategy, using DINO for boundary refinement and CLIP [45] for zero-shot classification [30, 59]. By combining global text-image alignment with local geometric priors, these frameworks produce much sharper mask boundaries than VLM-only baselines.

2.3 Continuous and Flow-Based Alignment

Neural ODEs [13] introduced the idea of framing deep networks as continuous dynamic systems. Instead of a discrete sequence of layers, the network learns an ODE to transform the hidden features. This continuous-time formulation is now heavily used in generative modeling. Flow Matching [36, 37] and Normalizing Flows [46], for instance, rely on learning a vector field to map between data distributions. In representation learning, ODE networks also guarantee invertibility and stability for robust feature extraction [4, 19]. Furthermore, because many feature spaces are non-Euclidean, researchers proposed Riemannian and manifold ODEs [38, 41]. By forcing the features to move strictly along geometric manifolds like a hypersphere, these models keep the original topological structure intact. However, cross-modal alignment in OVSS still relies on discrete or static methods. Standard approaches simply use MLP layers or Optimal Transport [21]. These sudden mapping steps easily break the geometric curvature of the embeddings. While continuous flows naturally preserve this geometry, they are rarely used to align different modalities. Therefore, we introduce a manifold-constrained ODE flow that continuously maps the CLIP [45] text space to the DINOv3 [51] visual space, achieving a smooth and geometry-aware alignment.

3 Method

In this work, we introduce an ODE-based alignment framework that continuously transforms text embeddings into the visual space. Section 3.1 presents the fundamental formulation of the ODE used in our method and briefly describes the formulation of the DINOv3 vision encoder and the CLIP text encoder. Then, Section 3.2 details our proposed **DINOde** framework and its key components.

3.1 Preliminaries

Ordinary Differential Equations. An ordinary differential equation (ODE) defines how a continuous variable evolves over time, governed by a vector field. Given a state $z_t \in \mathbb{R}^d$ and a time variable $t \in [0, 1]$, a neural ODE parameterizes the rate of change using a learnable neural network $v_\theta(z_t, t)$. The dynamics of z_t are described by the initial value problem:

$$\frac{dz_t}{dt} = v_\theta(z_t, t), \quad z_{t=0} = z_0. \tag{1}$$

Integrating this ODE from time $t = 0$ to $t = 1$ yields a continuous transformation trajectory from the initial state to the final state:

$$z_1 = z_0 + \int_0^1 v_\theta(z_t, t)\, dt. \tag{2}$$

In practice, this continuous integral is numerically approximated using discrete solvers such as the Euler method. For a step size $\Delta t = 1/N_{step}$, the state is iteratively updated as:

$$z_{t+\Delta t} = z_t + \Delta t \cdot v_\theta(z_t, t). \tag{3}$$

This iterative update can be viewed as the *continuous-depth limit of a residual network*, where each block performs an infinitesimal update along the time-dependent velocity field v_θ. Unlike discrete, single-step projections (e.g., MLPs), the ODE formulation provides a smooth, invertible, and stable trajectory for modeling complex feature transformations.

Task Formulation. We build upon two complementary pre-trained models: DINOv3 for visual representation and CLIP for text embedding. Our goal is to bridge these two representations by learning a continuous transformation that evolves the CLIP semantic manifold toward DINOv3's visual manifold for OVSS. Given an input image $I \in \mathbb{R}^{3\times H\times W}$, we denote the DINOv3 visual encoder as $f_{vis} : \mathbb{R}^{3\times H\times W} \to \mathbb{R}^{(N+1)\times D}$. The backbone encodes the image into N non-overlapping patch tokens together with a learnable `[CLS]` token:

$$\left(\mathbf{z}_{cls}^{img}, \mathbf{Z}^{img}\right) = f_{vis}(I), \tag{4}$$

where $\mathbf{Z}^{img} \in \mathbb{R}^{N\times D}$ and $\mathbf{z}_{cls}^{img} \in \mathbb{R}^{D}$, with D denoting the embedding dimension of DINOv3. For a textual prompt T, we denote the CLIP text encoder as f_{text}, and the resulting semantic embedding is:

$$\mathbf{z}^{text} = f_{text}(T) \in \mathbb{R}^{D_{text}}, \tag{5}$$

where D_{text} is the CLIP embedding dimension.

3.2 DINOde

In this work, we aim to exploit the potential of a self-supervised vision model for open-vocabulary semantic segmentation. We experimentally observe that the MLP-based method leads to suboptimal alignment, primarily because it employs a Euclidean shortcut that disregards the manifold of the feature space. To this end, we propose **DINOde**, an ODE-based image-text alignment framework that continuously transforms CLIP text space toward the visual space of DINO. DINOde consists of two complementary components: (*i*) **Semantic Text Flow (STF)** that gradually aligns the text manifold into the vision manifold, and (*ii*) **Global Context Flow (GCF)** that also progressively refines the DINO's holistic `[CLS]` token, producing a globally aligned representation.

Semantic Text Flow. As shown in Fig. 2, to continuously align the text embedding with the visual manifold, we model the image-text alignment as an ODE as in Eq. 1: $\frac{d\mathbf{z}_t}{dt} = v_\theta(\mathbf{z}_t, t)$. To specify the initial state and the visual supervision

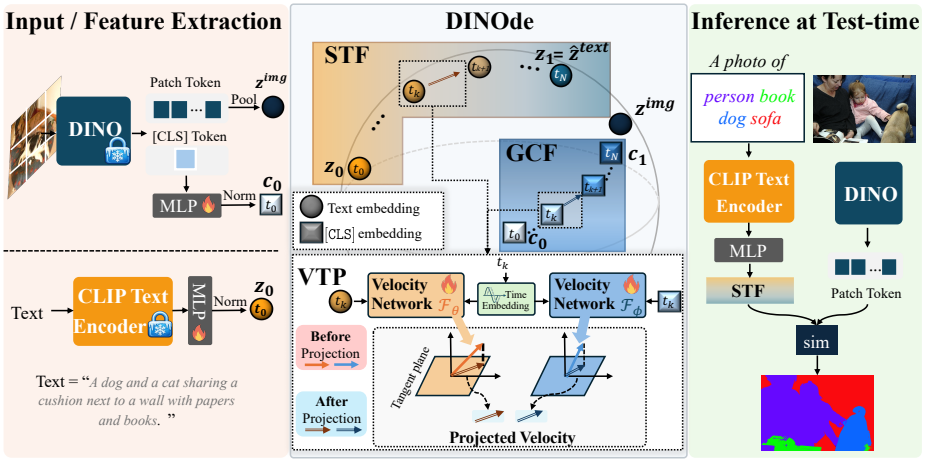

Fig. 2: Overview of the proposed DINOde framework. Given an input image, a frozen DINOv3 encoder extracts patch tokens and a `[CLS]` token, while a CLIP text encoder produces a semantic text embedding. **Semantic Text Flow (STF)** continuously aligns the text embedding to the DINO visual manifold via an ODE-based transformation, while **Global Context Flow (GCF)** refines the global `[CLS]` representation. To preserve hyperspherical geometry, **Velocity Tangent Projection (VTP)** constrains the velocity to the tangent space. The aligned embeddings are used to compute patch–text similarity for open-vocabulary semantic segmentation.

target, we first map the text embedding to the DINO feature dimension via a learnable linear projection: $\mathbf{z}_0 = \mathrm{Norm}(W\mathbf{z}^{text})$, where $W \in \mathbb{R}^{D\times D_{text}}$, and use $\mathbf{z}_0$ as the initial condition. Then, we obtain a pooled visual embedding

$$\mathbf{z}^{img} = \mathrm{Norm}(\mathrm{Pool}(\mathbf{Z}^{img})) \in \mathbb{R}^{D}, \tag{6}$$

which serves as the target visual representation. Here, $\mathrm{Norm}(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_2$ denotes ℓ_2 normalization. For the Pool operation, we employ top-K pooling as in [69]. To effectively inject temporal information into the velocity network $\mathcal{F}_\theta$, we realize time conditioning via a sinusoidal embedding $\gamma(t)$, which is incorporated through feature-wise modulation, i.e., $v_\theta(\mathbf{z}, t) = \mathcal{F}_\theta(\mathbf{z}; \gamma(t))$. The details about the velocity network are provided in the *Supplementary Materials*. To preserve the intrinsic geometry of the hyperspherical feature space, we propose **Velocity Tangent Projection (VTP)** that constrains the velocity to the tangent space at $\mathbf{z}_t$:

$$\tilde{v}_\theta(\mathbf{z}_t, t) = v_\theta(\mathbf{z}_t, t) - \langle v_\theta(\mathbf{z}_t, t), \mathbf{z}_t\rangle \mathbf{z}_t. \tag{7}$$

We then integrate the ODE forward from $t = 0$ to $t = 1$ to obtain the aligned embedding as in Eq. 3:

$$\hat{\mathbf{z}}^{text} = \mathbf{z}_1 = \Phi_\theta^{text}(\mathbf{z}_0), \tag{8}$$

where Φ_θ^{text} denotes the numerical flow map obtained by integrating Eq. 1 with an explicit Euler solver (Eq. 3), i.e., for N_{step} with $\Delta t = 1/N_{step}$,

$$\mathbf{z}_{t_{k+1}} = \mathrm{Norm}\big(\mathbf{z}_{t_k} + \Delta t\, \tilde{v}_\theta(\mathbf{z}_{t_k}, t_k)\big), \quad t_k = k\Delta t. \tag{9}$$

Global Context Flow. In addition to text alignment, we introduce **Global Context Flow (GCF)** to refine the holistic image representation carried by the `[CLS]` token. DINOv3 produces a global representation $\mathbf{z}_{cls}^{img} \in \mathbb{R}^D$ that summarizes the overall image context by aggregating information from all patch tokens. While the patch tokens $\mathbf{Z}^{img}$ capture fine-grained, part-level details, the `[CLS]` token provides holistic information of the given image. We leverage this property to align global semantics between the visual and textual spaces. Given the raw `[CLS]` token $\mathbf{z}_{cls}^{img}$ from the frozen DINOv3 backbone, we first apply a lightweight projection head and ℓ_2 normalization to obtain an initial state on the hypersphere:

$$\mathbf{c}_0 = \mathrm{Norm}\big(W_{cls}\mathbf{z}_{cls}^{img}\big) \in \mathbb{R}^D, \tag{10}$$

where $W_{cls} \in \mathbb{R}^{D\times D}$. We then evolve $\mathbf{c}_0$ via an ODE as in STF:

$$\frac{d\mathbf{c}_t}{dt} = u_\phi(\mathbf{c}_t, t), \tag{11}$$

where u_ϕ is a learnable velocity network. Similar to STF, time conditioning is realized using a sinusoidal embedding $\gamma(t)$ injected via feature-wise modulation. Here, we also apply **VTP** to preserve the hyperspherical geometry:

$$\tilde{u}_\phi(\mathbf{c}_t, t) = u_\phi(\mathbf{c}_t, t) - \langle u_\phi(\mathbf{c}_t, t), \mathbf{c}_t\rangle \mathbf{c}_t. \tag{12}$$

Finally, the globally refined representation is obtained by forward integration:

$$\hat{\mathbf{z}}_{cls}^{img} = \mathbf{c}_1 = \Psi_\phi(\mathbf{c}_0), \tag{13}$$

where Ψ_ϕ denotes the numerical flow map induced by Eq. 11. Again, we use Euler solver with N_{step} and step size $\Delta t = 1/N_{step}$ as follows:

$$\mathbf{c}_{t_{k+1}} = \mathrm{Norm}\big(\mathbf{c}_{t_k} + \Delta t\, \tilde{u}_\phi(\mathbf{c}_{t_k}, t_k)\big), \quad t_k = k\Delta t. \tag{14}$$

Loss formulation. In DINOde, we adopt a CLIP-style symmetric contrastive objective to train the velocity networks (v_θ, u_ϕ). Here, motivated by the prior work [24], we construct a global image representation by concatenating the pooled patch descriptor $\mathbf{z}^{img}$ and the refined `CLS` token $\hat{\mathbf{z}}_{cls}^{img}$ obtained via GCF:

$$\mathbf{z}_{||}^{img} = \Big[\mathbf{z}^{img}\,;\, \hat{\mathbf{z}}_{cls}^{img}\Big] \in \mathbb{R}^{2D}. \tag{15}$$

For each image-caption pair, we encode the caption with CLIP and produce a vision-aligned text feature $\hat{\mathbf{z}}^{text} \in \mathbb{R}^D$ obtained with STF (Eq. 8). To match dimensionality with $\mathbf{z}_{||}^{img}$, we repeat the $\hat{\mathbf{z}}^{text}$ to form $\mathbf{z}_{||}^{text} \in \mathbb{R}^{2D}$. Given a minibatch of B image-caption pairs, we compute the similarity matrix

$$\mathbf{S}_{ij} = \frac{1}{\tau}\,(\mathbf{z}_{||,i}^{img})^\top \mathbf{z}_{||,j}^{text}, \tag{16}$$

where τ is a temperature. The symmetric contrastive loss is then defined as

$$\mathcal{L}_{NCE} = \frac{1}{2}\Big(\mathrm{CE}(\mathbf{S}, \mathbf{y}) + \mathrm{CE}(\mathbf{S}^\top, \mathbf{y})\Big), \qquad \mathbf{y} = [1, \ldots, B], \tag{17}$$

where CE and $\mathbf{y}$ are cross-entropy loss and matching indices, respectively.

Inference. As illustrated in Fig. 2, the inference stage of DINOde utilizes the learned STF to generate semantic anchors for arbitrary categories and performs pixel-level segmentation by measuring their similarity with DINOv3 patch tokens. Specifically, for a given set of M candidate categories $\{c_m\}_{m=1}^{M}$, we first extract their textual embeddings using a frozen CLIP text encoder and map them to the source state $\mathbf{z}_0 \in \mathbb{R}^{D\times M}$ via the learnable projection W. We then perform forward integration from $t = 0$ to $t = 1$ by utilizing the learned velocity network v_θ as described in Eq. 9. The resulting target state, $\hat{\mathbf{z}}^{text} = \mathbf{z}_1 \in \mathbb{R}^{D\times M}$, serves as a set of semantic anchors optimally aligned with the visual manifold of DINO. Finally, the dense segmentation map is generated by computing the cosine similarity between the DINO patch tokens and these aligned semantic anchors $\hat{\mathbf{z}}^{text}$.

4 Experiments

4.1 Experimental Settings

Datasets. To evaluate the robustness and generalizability of our proposed method, we conduct extensive experiments on eight common OVSS benchmarks. Following the categorization in prior literature [2, 3, 9, 59], these datasets are divided into two groups based on the inclusion of a background (bg) class. The first group consists of three benchmarks with a background class: **Pascal VOC 2012 (V21)** [20], **Pascal Context (C60)** [42], and **COCO Object (Object)** [6], containing 21, 60, and 81 classes, respectively. The second group includes five benchmarks without a background class. Specifically, we evaluate on **COCO Stuff (Stuff)** [6], **Cityscapes (City)** [17], and **ADE20K (ADE)** [76, 77], which comprise 171, 19, and 150 classes, respectively. Furthermore, we report performance on **Pascal VOC 2012 (V20)** [20] and **Pascal Context (C59)** [42], both of which exclude the background class to provide a comprehensive comparison.

Evaluation Protocol and Metrics. We follow the evaluation protocols established in previous works [3,9,59,62]. During inference, for a fair comparison with the state-of-the-art work [3], input images are resized to a shorter-side length of 448 pixels and processed via sliding window inference with a 448×448 resolution and a 224-pixel stride. Following prior works [3, 54], during inference, the textual embeddings are generated by averaging the encodings of standard prompt templates (*e.g.*, "*a photo of a {class}*"). The primary evaluation metric for every experiment is the mean-Intersection-over-Union (mIoU).

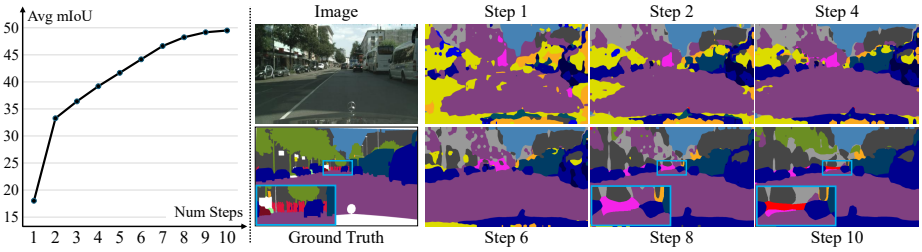

Fig. 3: Analysis of Semantic Text Flow (STF) with varying numbers of ODE steps. Segmentation maps are obtained using the aligned text embedding $\mathbf{z}_{t_k}$ at each step $k \in \{1, 2, \ldots, 10\}$. (Left) Average mIoU across eight benchmarks, showing the progressive performance improvement during the manifold transition. (Right) Qualitative results on Cityscapes across steps.

Implementation Details. Throughout the experiments, we employ DINOv3 ViT-L/16 as our visual backbone and CLIP ViT-L/14 as the textual encoder. In contrast to conventional weakly-supervised OVSS methods [9, 11, 39, 62, 64] that rely on large-scale image-caption datasets such as CC3M [48] (∼3.3M) or CC12M [10] (∼12.4M), our DINO and ODE-based approach achieves efficient alignment with significantly less data, specifically training on the COCO 2017 Caption [14, 35] *train* set (∼118k images). We use the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 0.01. The model is trained for 20 epochs with a batch size of 256 on a single NVIDIA RTX 3090 GPU. Since we utilize frozen DINOv3 and CLIP encoders, we can leverage pre-cached embeddings, and the aligned text anchors are cached once per class set, allowing further computational efficiency. The entire training process, including the initial caching phase, is completed within 4 hours of wall-clock time. Consistent with [3, 9], we apply Pixel-Adaptive Mask Refinement (PAMR) [1] as a post-processing step to enhance the boundary adherence of the final semantic maps. The hyperparameters for our method are as follows: the number of ODE steps $N_{step} = 10$, $K = 20$ for top-K pooling, the temperature $\tau = 0.07$ for NCE loss.

4.2 Discussions

Analysis of Progressive Manifold Transitions. To verify whether STF effectively learns the progressive manifold transition via ODE as intended, we perform quantitative and qualitative comparisons across varying numbers of ODE integration steps. The average mIoU across eight benchmarks for each step is illustrated in the graph of Fig. 3 (left). Notably, the performance exhibits a rapid increase from steps 1 to 8. This trend demonstrates that the velocity network is effectively trained to transport CLIP text embeddings toward the DINO visual manifold in a progressive and stable manner. Between steps 8 and 10, the mIoU curve shows a more gradual yet steady improvement. This suggests that once the embeddings reach the vicinity of the DINO manifold, the STF module performs fine-grained adjustments to converge toward the optimal semantic coordinates

Table 1: Component analysis of the proposed method.

	STF	VTP	GCF	V20	C59	Stuff	City	ADE	V21	C60	Object	Avg
(a)				88.4	43.2	29.7	42.5	23.2	68.0	39.6	46.7	47.7
(b)	✓			88.4	43.0	31.2	44.7	25.2	66.6	39.5	47.1	48.2
(c)	✓	✓		88.5	43.4	31.4	45.4	25.5	67.3	39.6	47.2	48.5
(d)	✓		✓	90.2	44.3	31.3	45.6	24.8	69.5	40.3	47.2	49.2
(e)	✓	✓	✓	91.1	44.5	31.6	45.4	25.4	69.2	40.7	48.0	**49.5**

within the visual space. The qualitative results on the Cityscapes dataset, shown in Fig. 3, consistently align with these quantitative observations. In the first 8 steps, we observe rapid improvements in global structure and semantic consistency. In the subsequent steps, the results exhibit further improvements in fine details. In conclusion, these results demonstrate that our STF successfully learns an optimal manifold transition trajectory, effectively bridging the modality gap between CLIP and DINO.

Component Analysis. Table 1 presents an ablation study evaluating the individual contributions of our proposed Semantic Text Flow (STF) and Global Context Flow (GCF), as well as the impact of Velocity Tangent Projection (VTP). In this table, result (a) is the baseline where text-to-visual alignment is performed using a standard MLP with a parameter capacity comparable to that of the STF Velocity Network. A comparison between (a) and (b)–(e) demonstrates that the ODE-based progressive transition via STF achieves superior performance compared to simple static mapping. Notably, the steady improvement in average mIoU from (a) to (c) and (e) suggests that STF successfully learns an optimal transition trajectory for transferring CLIP text embeddings to the DINO visual manifold, surpassing the capabilities of mere feature alignment. To verify the role of VTP in preserving the intrinsic geometry of the hyper-spherical feature space, we compare (b) with (c) and (d) with (e); in both cases, VTP leads to meaningful performance improvements. This demonstrates that projecting velocity vectors onto the tangent space, thereby maintaining the feature norm during transition, plays an important role in ensuring the stability and accuracy of the progressive manifold transfer. Finally, we analyze the effect of GCF, which leverages the DINO `[CLS]` token to incorporate global contextual information. GCF ensures that the transition of text embeddings (STF) is not biased toward local visual features but instead accounts for the broader global context. Comparisons between (b) and (d), as well as (c) and (e), demonstrate that GCF yields consistent performance improvements. These results indicate that STF and GCF complementarily optimize the continuous alignment between manifolds.

Ablation on Number of Steps. Table 2 analyzes the impact of the number of ODE steps (N_{step}) used during both the training and inference of the STF and GCF modules. We observe that even with $N_{step} = 5$, our method outperforms

Table 2: Ablation study on the number of ODE steps (N_{step}).

Align Method	N_{step}	V20	C59	Stuff	City	ADE	V21	C60	Object	Avg
MLP-based	-	88.4	43.2	29.7	42.5	23.2	68.0	39.6	46.7	47.7
ODE-based	5	89.7	44.4	31.7	43.5	25.2	66.7	40.4	48.0	48.7
	10	91.1	44.5	31.6	45.4	25.4	69.2	40.7	48.0	49.5
	50	91.0	44.1	31.9	46.1	25.4	69.1	40.2	48.5	49.6

Table 3: Generalization capability across various visual and textual backbones. We compare our proposed alignment (Ours) against the discrete projection baseline (MLP).

DINOv3(Visual)	CLIP(Textual)	Alignment	V20	C59	Stuff	City	ADE	V21	C60	Object	Avg
ViT-L	ViT-L	MLP	88.4	43.2	29.7	42.5	23.2	68.0	39.6	46.7	47.7
		Ours	91.1	44.5	31.6	45.4	25.4	69.2	40.7	48.0	**49.5**
ViT-B	ViT-L	MLP	87.3	39.2	27.7	38.1	20.7	62.4	36.1	41.8	44.2
		Ours	89.2	41.6	29.3	37.7	22.4	66.2	38.4	43.8	**46.1**
ViT-S	ViT-L	MLP	80.3	34.8	19.9	25.8	15.7	54.2	32.7	31.6	36.9
		Ours	84.2	36.1	22.6	29.4	16.3	55.6	33.6	32.9	**38.8**
ViT-L	ViT-B	MLP	87.7	43.1	30.6	43.8	24.3	67.3	39.5	46.3	47.8
		Ours	90.8	42.8	30.5	42.8	25.0	69.8	39.7	48.1	**48.7**

the MLP-based discrete mapping, demonstrating the effectiveness of ODE-based transitions for cross-modal alignment. The improvement attained at $N_{step} = 10$ further suggests that a sufficient level of granularity enables the model to more successfully navigate the complex manifold gap between the disparate domains. For $N_{step} \geq 10$, the average mIoU exhibits a subtle upward trend as the number of steps increases. This suggests that a higher number of steps facilitates a smoother transition trajectory, assisting the model in converging more closely to the optimal position on the target manifold. However, we also observe marginal performance drops in certain benchmarks. This can be attributed to the accumulation of minute errors during the numerical integration process over an excessive number of steps, which potentially causes the final embedding to 'over-shoot' its optimal target within the visual manifold. Considering the trade-off between performance gains and computational overhead, we select $N_{step} = 10$ as the optimal balance for our framework.

Generalization to Diverse Backbones. To demonstrate the generalization capability of the DINOde, we conduct experiments with various visual and textual backbones, as shown in Table 3. As the DINOv3 visual backbone scales down to ViT-B and ViT-S, our continuous flow-based alignment method consistently outperforms the baseline, MLP-based alignment. This suggests that, regardless of the granularity of visual representation, ODE-based trajectory modeling is inherently more effective at bridging the manifold gap between modalities than discrete projection methods. Furthermore, DINOde surpasses the MLP-based baseline even when the CLIP textual backbone is transitioned to ViT-B. This

Table 4: Geometric diagnostics of manifold preservation. STF consistently outperforms the MLP baseline across all criteria.

Geometric Criterion	MLP	STF(Ours)
(i) *Neighborhood preservation*	0.575	**0.601**
(ii) *Angular / geodesic consistency*	0.693	**0.712**
(iii) *Class-structure preservation*	0.843	**0.871**
(iv) *Alignment-space compactness*	0.412	**0.423**

demonstrates that our method effectively guides the transition trajectory toward the target DINO visual manifold, even when the underlying composition of the text embedding space is altered. These results demonstrate that the DINOde possesses high generalization capability, flexibly accommodating diverse visual and textual manifolds without relying on a specific backbone.

Geometric Diagnostics of Manifold Preservation. To directly verify that STF preserves the intrinsic geometry of the feature space, we report four geometric diagnostics in Table 4. We measure (i) the *top-10 nearest-neighbor overlap* between the original CLIP class space and the aligned space, (ii) the *local Pearson correlation of geodesic-distance vectors* between neighboring classes, (iii) linear *CKA between class-level Gram matrices* to assess whether the overall semantic structure is preserved, and (iv) the *mean cosine similarity* between aligned text features and mask-aware DINO patch prototypes of the corresponding classes. Across all four criteria, STF consistently surpasses the MLP baseline, indicating that the continuous ODE-based transition better retains neighborhood structure, geodesic consistency, and class-level semantic organization during cross-modal alignment.

Comparison to state-of-the-art. In Table 5, we present a quantitative comparison between our proposed DINOde and state-of-the-art OVSS methods across eight benchmarks. Baseline methods are categorized into training-free OVSS approaches [2, 5, 23, 31, 32, 40, 49, 54–56, 78] and weakly-supervised OVSS methods [3, 7, 9, 24, 34, 39, 43, 62, 68] that leverage image-caption pairs. We conduct comprehensive comparisons with representative baselines and recent state-of-the-art approaches, particularly dino.txt [24] and Talk2DINO [3], which also utilize image-caption pairs for training. To ensure a fair comparison, the results for Talk2DINO* in Table 5 and Fig. 4 were obtained using officially released checkpoints and source code integrated with a DINOv3 backbone. Specifically for benchmarks with a background, we conducted an exhaustive search for the optimal background threshold to ensure peak performance of Talk2DINO*.

Our method outperforms existing state-of-the-art approaches on most benchmarks (6 out of 8), both with and without mask refinement. Notably, compared with Talk2DINO*, which employs the same DINOv3 visual encoder backbone, our approach achieves an average mIoU of 49.5% before mask refinement, yielding a 1.9%p improvement. Upon incorporating the final mask refinement step,

Table 5: Quantitative comparison with state-of-the-art methods. **Bold** and Underline denote the best and second-best results, respectively.

Model	Visual Encoder	V20	C59	Stuff	City	ADE	V21	C60	Object	Avg
without Mask Refinement										
MaskCLIP [78]$_{ECCV22}$	CLIP	29.4	12.4	8.8	11.5	7.2	23.3	11.7	7.2	13.9
SCLIP [56]$_{ECCV24}$	CLIP	70.6	25.2	17.6	21.3	10.9	44.0	22.3	26.9	29.9
ClearCLIP [31]$_{ECCV24}$	CLIP	80.0	29.6	19.9	27.9	15.0	-	-	-	-
NACLIP [23]$_{WACV25}$	CLIP	78.7	32.1	21.4	31.4	17.3	52.2	28.7	29.9	36.5
dino.txt [24]$_{CVPR25}$	DINOv2(reg)	62.1	30.9	20.9	32.1	20.6	-	-	-	-
FreeDA [2]$_{CVPR24}$	DINOv2	71.8	35.4	24.2	32.3	19.4	44.9	31.1	24.6	35.5
FreeDA [2]$_{CVPR24}$	CLIP+DINOv2	85.7	39.7	26.3	33.6	21.4	44.1	34.8	33.9	39.9
ProxyCLIP [32]$_{ECCV24}$	CLIP+DINOv2(reg)	85.2	36.2	24.6	35.2	21.6	56.6	33.0	36.7	41.1
ProxyCLIP [32]$_{ECCV24}$	CLIP+DINO	83.2	37.7	25.6	40.1	22.6	60.6	34.5	39.2	43.0
CLIPer [54]$_{ICCV25}$	CLIP	88.2	39.8	25.8	-	21.8	61.2	34.3	39.6	-
Talk2DINO [3]$_{ICCV25}$	DINOv2(reg)	87.1	39.1	27.0	35.8	21.1	60.1	34.2	37.6	42.8
Talk2DINO* [3]$_{ICCV25}$	DINOv3	87.7	43.0	**32.6**	40.8	24.2	65.9	37.9	**48.6**	47.6
Ours	DINOv3	**91.1**	**44.5**	31.6	**45.4**	**25.4**	**69.2**	**40.7**	48.0	**49.5**
with Mask Refinement										
SCLIP [56]$_{ECCV24}$	CLIP	76.3	27.4	18.7	23.9	11.8	47.8	23.8	26.9	32.1
NACLIP [23]$_{WACV25}$	CLIP	84.5	36.4	24.6	37.1	19.6	57.9	36.4	34.6	41.4
FreeDA [2]$_{CVPR24}$	DINOv2	75.2	39.0	27.0	33.1	21.3	45.3	34.3	26.7	37.7
FreeDA [2]$_{CVPR24}$	CLIP+DINOv2	87.1	42.4	28.1	33.8	22.6	55.4	37.1	36.1	42.8
ProxyCLIP [32]$_{ECCV24}$	CLIP+DINOv2(reg)	85.8	37.6	25.6	37.5	22.5	59.4	34.6	39.0	42.8
ProxyCLIP [32]$_{ECCV24}$	CLIP+DINO	83.2	38.0	26.2	41.0	22.6	60.7	34.7	39.4	43.2
CLIPer [54]$_{ICCV25}$	CLIP	90.0	43.6	28.7	-	24.4	**69.8**	38.0	43.3	-
Talk2DINO [3]$_{ICCV25}$	DINOv2(reg)	89.8	42.7	29.6	38.4	22.9	66.1	37.3	42.3	46.1
Talk2DINO* [3]$_{ICCV25}$	DINOv3	88.2	44.3	**33.8**	40.5	24.8	67.0	39.2	**50.4**	48.5
Ours	DINOv3	**91.6**	**45.3**	32.1	**46.2**	**25.9**	**69.8**	**41.1**	48.4	**50.1**

our framework reaches a peak average mIoU of 50.1%, further consolidating its superior performance across most of the benchmarks. This result suggests that the gains are not merely due to the superior visual representations of the backbone. The superiority of our approach is further supported by the qualitative comparisons presented in Fig. 4, which illustrates the results across three distinct benchmarks: Pascal VOC (V21), Cityscapes (City), and Pascal Context (C59). Notably, in the Cityscapes examples (Fig. 4 middle row), our method successfully captures small *Bus* and *vegetation* classes that the compared baselines fail to identify. Additional qualitative comparison results are provided in the *Supplementary Materials*. Our results exhibit noticeable improvements in fine details compared to other SOTA methods, demonstrating that the ODE-based transition effectively achieves optimal textual-visual manifold alignment.

5 Conclusion

In this work, we propose DINOde, an ODE-based alignment framework that bridges the semantic gap between self-supervised visual representations and text representations for open-vocabulary semantic segmentation. Existing MLP-based projection methods approximate the complex relationship between het-

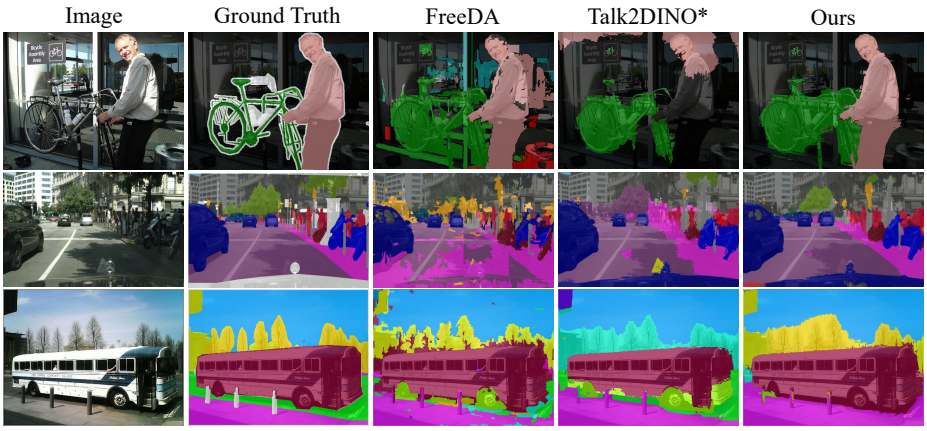

Fig. 4: Qualitative comparison with the state-of-the-art methods. The result of Talk2DINO* is obtained using the DINOv3 backbone for a fair comparison.

erogeneous representation manifolds through a single-step mapping, which often fails to preserve the intrinsic geometry of the feature space. To address this limitation, in DINOde, we propose Semantic Text Flow (STF) that gradually evolves text embeddings toward the DINO feature space along an ODE trajectory, enabling smooth cross-modal alignment. In addition, Global Context Flow (GCF) progressively refines the global image representation carried by the `[CLS]` token to better capture holistic visual semantics. To further preserve the hyperspherical geometry of the feature space, we introduce Velocity Tangent Projection (VTP), which constrains the velocity field to the tangent space and enables geometry-preserving alignment. Extensive experiments demonstrate that the proposed DINOde effectively improves cross-modal representation learning and consistently enhances performance on OVSS benchmarks. These results highlight the potential of flow-based alignment for bridging vision and language representations. As future work, since modern multimodal large language models map frozen visual features into the language space via a single MLP projector, which is analogous to the discrete baseline we revisit, we expect continuous ODE-based alignment to be a promising direction for improving fine-grained visual grounding in MLLMs.

Acknowledgements

This research was supported by the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2026-25561904) and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2026-25473963).

References

1. Araslanov, N., Roth, S.: Single-stage semantic segmentation from image labels. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4253–4262 (2020)
2. Barsellotti, L., Amoroso, R., Cornia, M., Baraldi, L., Cucchiara, R.: Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3689–3698 (2024)
3. Barsellotti, L., Bianchi, L., Messina, N., Carrara, F., Cornia, M., Baraldi, L., Falchi, F., Cucchiara, R.: Talking to dino: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22025–22035 (2025)
4. Behrmann, J., Grathwohl, W., Chen, R.T., Duvenaud, D., Jacobsen, J.H.: Invertible residual networks. In: International conference on machine learning. pp. 573–582. PMLR (2019)
5. Bousselham, W., Petersen, F., Ferrari, V., Kuehne, H.: Grounding everything: Emerging localization properties in vision-language transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3828–3837 (2024)
6. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018)
7. Cai, K., Ren, P., Zhu, Y., Xu, H., Liu, J., Li, C., Wang, G., Liang, X.: Mixreorg: Cross-modal mixed patch reorganization is a good mask learner for open-world semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1196–1205 (2023)
8. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
9. Cha, J., Mun, J., Roh, B.: Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11165–11174 (2023)
10. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3558–3568 (2021)
11. Chen, J., Zhu, D., Qian, G., Ghanem, B., Yan, Z., Zhu, C., Xiao, F., Culatana, S.C., Elhoseiny, M.: Exploring open-vocabulary semantic segmentation from clip vision encoder distillation only. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 699–710 (2023)
12. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
13. Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. Advances in neural information processing systems **31** (2018)
14. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)

15. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in neural information processing systems **34**, 17864–17875 (2021)
16. Cho, H., Yoon, S.H., Kweon, H., Yoon, K.J.: Finding meaning in points: Weakly supervised semantic segmentation for event cameras. In: European Conference on Computer Vision. pp. 266–286. Springer (2024)
17. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
18. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023)
19. Dupont, E., Doucet, A., Teh, Y.W.: Augmented neural odes. Advances in neural information processing systems **32** (2019)
20. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
21. Gandhamal, A., Sikdar, A., Sundaram, S.: Ov-coast: Cost aggregation with optimal transport for open-vocabulary semantic segmentation. arXiv preprint arXiv:2506.03706 (2025)
22. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: European conference on computer vision. pp. 540–557. Springer (2022)
23. Hajimiri, S., Ayed, I.B., Dolz, J.: Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5061–5071. IEEE (2025)
24. Jose, C., Moutakanni, T., Kang, D., Baldassarre, F., Darcet, T., Xu, H., Li, D., Szafraniec, M., Ramamonjisoa, M., Oquab, M., et al.: Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24905–24916 (2025)
25. Kang, D., Cho, M.: In defense of lazy visual grounding for open-vocabulary semantic segmentation. In: European Conference on Computer Vision. pp. 143–164. Springer (2024)
26. Kim, J., Kwon, H., Kweon, H., Yoon, K.J.: Bootstrapping video semantic segmentation model via distillation-assisted test-time adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10766–10777 (2026)
27. Kweon, H., Yoon, S.H., Kim, H., Park, D., Yoon, K.J.: Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6994–7003 (2021)
28. Kweon, H., Yoon, S.H., Yoon, K.J.: Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11329–11339 (2023)
29. Kwon, H., Jeong, J., Yoon, S.H., Yoon, K.J.: Phase concentration and shortcut suppression for weakly supervised semantic segmentation. In: European Conference on Computer Vision. pp. 293–312. Springer (2024)
30. Lai, Z.: Exploring simple open-vocabulary semantic segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30221–30230 (2025)

31. Lan, M., Chen, C., Ke, Y., Wang, X., Feng, L., Zhang, W.: Clearclip: Decomposing clip representations for dense vision-language inference. In: European Conference on Computer Vision. pp. 143–160. Springer (2024)
32. Lan, M., Chen, C., Ke, Y., Wang, X., Feng, L., Zhang, W.: Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In: European Conference on Computer Vision. pp. 70–88. Springer (2024)
33. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=RriDjddCLN
34. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7061–7070 (2023)
35. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
36. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
37. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
38. Lou, A., Lim, D., Katsman, I., Huang, L., Jiang, Q., Lim, S.N., De Sa, C.M.: Neural manifold ordinary differential equations. Advances in Neural Information Processing Systems **33**, 17548–17558 (2020)
39. Luo, H., Bao, J., Wu, Y., He, X., Li, T.: Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In: International Conference on Machine Learning. pp. 23033–23044. PMLR (2023)
40. Luo, J., Khandelwal, S., Sigal, L., Li, B.: Emergent open-vocabulary semantic segmentation from off-the-shelf vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4029–4040 (2024)
41. Mathieu, E., Nickel, M.: Riemannian continuous normalizing flows. Advances in neural information processing systems **33**, 2503–2515 (2020)
42. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 891–898 (2014)
43. Mukhoti, J., Lin, T.Y., Poursaeed, O., Wang, R., Shah, A., Torr, P.H., Lim, S.N.: Open vocabulary semantic segmentation with patch aligned contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19413–19423 (2023)
44. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
46. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International conference on machine learning. pp. 1530–1538. PMLR (2015)

47. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
48. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018)
49. Shin, G., Xie, W., Albanie, S.: Reco: Retrieve and co-segment for zero-shot transfer. Advances in Neural Information Processing Systems **35**, 33754–33767 (2022)
50. Siméoni, O., Puy, G., Vo, H.V., Roburin, S., Gidaris, S., Bursuc, A., Pérez, P., Marlet, R., Ponce, J.: Localizing objects with self-supervised transformers and no labels. arXiv preprint arXiv:2109.14279 (2021)
51. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
52. Sun, G., Liu, Y., Ding, H., Probst, T., Van Gool, L.: Coarse-to-fine feature mining for video semantic segmentation. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3126–3137 (2022)
53. Sun, G., Liu, Y., Ding, H., Wu, M., Van Gool, L.: Learning local and global temporal contexts for video semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
54. Sun, L., Cao, J., Xie, J., Jiang, X., Pang, Y.: Cliper: Hierarchically improving spatial representation of clip for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23199–23209 (2025)
55. Sun, S., Li, R., Torr, P., Gu, X., Li, S.: Clip as rnn: Segment countless visual concepts without training endeavor. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13171–13182 (2024)
56. Wang, F., Mei, J., Yuille, A.: Sclip: Rethinking self-attention for dense vision-language inference. In: European conference on computer vision. pp. 315–332. Springer (2024)
57. Wang, Y., Shen, X., Yuan, Y., Du, Y., Li, M., Hu, S.X., Crowley, J.L., Vaufreydaz, D.: Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. IEEE transactions on pattern analysis and machine intelligence **45**(12), 15790–15801 (2023)
58. Wysoczańska, M., Ramamonjisoa, M., Trzciński, T., Siméoni, O.: Clip-diy: Clip dense inference yields open-vocabulary semantic segmentation for-free. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1403–1413 (2024)
59. Wysoczańska, M., Siméoni, O., Ramamonjisoa, M., Bursuc, A., Trzciński, T., Pérez, P.: Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In: European Conference on Computer Vision. pp. 320–337. Springer (2024)
60. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021)
61. Xie, J., Hou, X., Ye, K., Shen, L.: Clims: Cross language image matching for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4483–4492 (2022)

62. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18134–18144 (2022)
63. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2955–2966 (2023)
64. Xu, J., Hou, J., Zhang, Y., Feng, R., Wang, Y., Qiao, Y., Xie, W.: Learning open-vocabulary semantic segmentation models from natural language supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2935–2944 (2023)
65. Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Xu, D.: Multi-class token transformer for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4310–4319 (2022)
66. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2945–2954 (2023)
67. Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X.: A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In: European conference on computer vision. pp. 736–753. Springer (2022)
68. Yi, M., Cui, Q., Wu, H., Yang, C., Yoshie, O., Lu, H.: A simple framework for text-supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7071–7080 (2023)
69. Yoon, S.H., Kweon, H., Cho, J., Kim, S., Yoon, K.J.: Adversarial erasing framework via triplet with gated pyramid pooling layer for weakly supervised semantic segmentation. In: European conference on computer vision. pp. 326–344. Springer (2022)
70. Yoon, S.H., Kwon, H., Jeong, J., Park, D., Yoon, K.J.: Diffusion-guided weakly supervised semantic segmentation. In: European Conference on Computer Vision. pp. 393–411. Springer (2024)
71. Yoon, S.H., Kwon, H., Kim, H., Yoon, K.J.: Class tokens infusion for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3595–3605 (2024)
72. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. Advances in Neural Information Processing Systems **36**, 32215–32234 (2023)
73. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. pp. 173–190. Springer (2020)
74. Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Yang, J., Zhang, L.: A simple framework for open-vocabulary segmentation and detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1020–1031 (2023)
75. Zheng Ding, Jieke Wang, Z.T.: Open-vocabulary universal image segmentation with maskclip. In: International Conference on Machine Learning (2023)
76. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)

77. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International journal of computer vision **127**(3), 302–321 (2019)
78. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European conference on computer vision. pp. 696–712. Springer (2022)
79. Zou, X., Dou, Z.Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., et al.: Generalized decoding for pixel, image, and language. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15116–15127 (2023)