Reconstructing CLIP for Open-Vocabulary Dense Perception

Yajie Liu[1,2], Jinjin Zhang[1,2], Qingjie Liu[2,3], Di Huang[1,2,3*]
[1]State Key Laboratory of Complex and Critical Software Environment,
Beihang University, Beijing 100191, China
[2]School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[3]Zhejiang Industrial Big Data and Robot Intelligent System Key Laboratory,
Hangzhou Innovation Institute, Beihang University, China

Abstract

*Large-scale vision–language models (VLMs) such as CLIP have excelled in zero-shot image classification, yet they struggle to achieve the dense cross-modal alignment required by open-vocabulary dense perception (OVDP). While recent self-distillation methods address this by aligning dense features with the generalizable global semantics, a key question remains: how should such dense features be constructed to achieve optimal alignment? To address this, we propose DenseRC, a principled **Dense R**epresentations **C**onstruction framework that reconstructs CLIP for OVDP based on two key insights. First, by analyzing the internal semantics encoded in the global cls token, we identify that multi-layer value embeddings serve as an informative basis for dense features. Second, we reveal that spatial aggregation tends to amplify semantic misalignment. Motivated by this, we design a lightweight Head-Selective Gating (HSG) module that adaptively reweights feature heads according to their intrinsic heterogeneity, enabling discriminative and alignment-friendly dense representations construction. Extensive experiments demonstrate that DenseRC delivers consistent and substantial gains across OVDP tasks including object detection and semantic segmentation, setting new state-of-the-art performance on multiple benchmarks.*

1. Introduction

Large-scale vision–language models (VLMs) such as CLIP [28] have demonstrated strong transferable representations for zero-shot image classification. This success has motivated their adaptation to open-vocabulary dense perception (OVDP), which aims to recognize a broad spectrum of visual concepts beyond a fixed set of categories.

Most OVDP methods build on CLIP to leverage its generalization capability for dense visual recognition at the region or pixel level. However, CLIP is optimized for global image–text alignment, resulting in underwhelming dense representations, as noted in recent works [18, 35]. Specifically, CLIP's dense features suffer from weak cross-modal alignment with text and poor feature coherence [23].

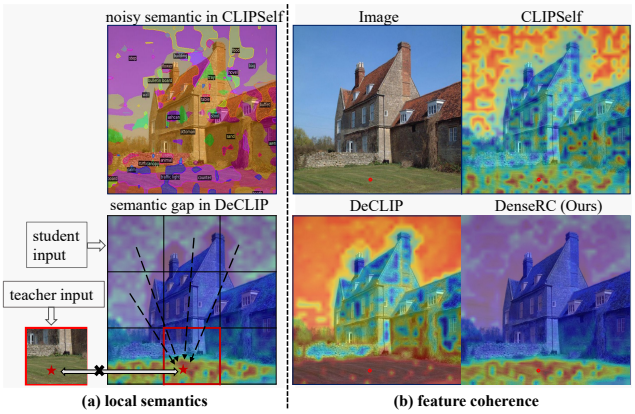

Figure 1. Limitations of prior methods in local semantics and feature coherence. **Left:** CLIPSelf [37] suffers from semantic noise in dense embeddings due to residual connections, while DeCLIP [35] aggregates irrelevant patches (e.g., sky regions) when forming dense student features, leading to a semantic gap with the cropped teacher input. **Right:** The dense features from prior methods show weak feature coherence, exhibiting large intra-class variance and low inter-class separation, as reflected by the irrelevant semantics activated by the given red point.

To equip CLIP with dense alignment capability that benefits various dense downstream tasks, existing works primarily follow two directions. The first line [14, 41] retrains CLIP with fine-grained region–text supervision. However, curating high-quality dense annotations is far more costly than collecting image–text pairs, limiting scalability. Alternatively, another line [35, 37] employs self-distillation to transfer CLIP's global alignment capability towards dense level without supervision.

These self-distillation methods use the global image rep-

*Corresponding author

resentation, the *cls* token, as a bridge to guide dense alignment with text. By aligning pooling student dense features with the frozen *cls* token from corresponding image crops, they aim to instill generalizable semantics into the dense representations. Thus, constructing dense features that are consistent with the *cls* token is essential. Nevertheless, current implementations fall short. As shown in Fig. 1, CLIPSelf [37] directly uses CLIP's final spatial output, but its residual connections inevitably inject noise [18]. DeCLIP [35] enhances patch semantics through self–self attention aggregation, but this mechanism inevitably introduces contamination from irrelevant patches, thereby leading to a semantic gap. Meanwhile, the issue of feature coherence remains insufficiently resolved. Although DeCLIP attempts to enforce coherence on their attention blocks, the final dense representations still suffer from limited category discriminability, particularly within stuff regions. Overall, how to effectively construct discriminative dense representations well-aligned with the global semantics remains an open problem.

To tackle these issues, we propose DenseRC, a novel framework for constructing *cls*-compatible dense representations for OVDP. We decompose the construction into two key aspects: (1) *what* features to extract and (2) *how* to construct them. For *what*, we reveal that the generalizable semantics of *cls* token are encoded within its multi-layer value embeddings, which therefore are employed as the informative basis to construct dense features in DenseRC. For *how*, we re-examine the self-attention mechanism for OVDP by decoupling it along the spatial and head dimensions. Our analysis shows that apply spatial aggregation in dense feature construction would increases the alignment error with the *cls* token. Further, we observe strong heterogeneity across heads in terms of both cross-modal semantic contribution and locality preference, suggesting that uniform aggregation is suboptimal. These findings motivate the Head-Selective Gating (HSG) module, which employs learnable gating to adaptively reweight head-wise features based on their semantic and spatial characteristics, yielding more discriminative and alignment-friendly dense representations.

To enhance feature coherence, we distill feature correlations from visual foundation models (VFM) [26] directly on the dense features. Unlike DeCLIP [35], which applies separate losses to decoupled attention (for feature coherence) and value (for dense alignment), our method enforces a unified constraint on a single feature stream. This integration enforces dense features to simultaneously attain cross-modal alignment and feature coherence.

We evaluate DenseRC on a range of OVDP tasks, including dense zero-shot classification [37], object detection [9, 25], and semantic segmentation [4]. Extensive experiments show that our approach outperforms previous state-of-the-art methods by large margins across multiple OVDP benchmarks, demonstrating its strong capability in fine-grained visual understanding.

In summary, our contributions are threefold:

- We propose DenseRC, a framework for constructing *cls*-compatible dense representations for OVDP. We analyze the internal semantics encoded in *cls* token and reveal that multi-layer value embeddings offer a semantically consistent basis to construct dense features.
- We establish the limitation of spatial aggregation and highlight the importance of head-wise modeling in dense feature construction. Accordingly, we design a lightweight HSG module that adaptively integrates head-wise information to construct dense representations.
- Extensive experiments on multiple open-vocabulary dense prediction benchmarks demonstrate the effectiveness and strong generalization capability of our approach.

2. Related Work

Open-Vocabulary Dense Perception. Large-scale vision–language models such as CLIP [28] have demonstrated strong transferable representations for zero-shot image classification. This capability has motivated growing interest in OVDP [4, 9], which aims to detect or segment arbitrary concepts based on natural language descriptions. Most existing approaches [4, 13, 25] fine-tune CLIP on specific downstream tasks using densely annotated data. While effective in specialized settings, these methods remain task-specific and do not address the fundamental limitations of CLIP in capturing fine-grained visual representations.

Adapting VLMs for Dense Prediction. A separate line of work seeks to enhance CLIP during the **pre-finetuning (upstream)** stage, after VLM pre-training but before downstream adaptation, to improve its ability to generate general-purpose dense representations that benefit a wide range of OVDP tasks. Existing methods in this category fall into two main paradigms: 1) Region-Supervised Retraining. Methods [14, 41] collect large-scale region–text datasets to retrain CLIP with fine-grained alignment. For example, FG-CLIP constructs 40M region–caption pairs to enforce local correspondence. While these approaches improve localization, they require costly data curation and substantial computational resources, hindering their scalability and practical deployment. 2) Self-Distillation without Extra Supervision. As a more scalable alternative, this paradigm transfers global semantics from a frozen CLIP model to a trainable dense encoder via self-distillation. CLIPSelf [37] aligns student region features with CLIP's image-level embeddings extracted from cropped views. R-SC-CLIPSelf [27] further distills spatial structure from CLIP and introduces a refinement stage to enhance the teacher's spatial awareness. DeCLIP [35] improves local discriminability by de-

coupling attention and value streams and enforcing spatial consistency regularization.

While existing methods have made strides in improving either dense semantics or feature coherence, the core question of how to *construct* dense representations that are intrinsically consistent with CLIP's global embedding space remains under-explored. In this work, we propose a principled framework for building *cls*-compatible dense representations that inherently align with global semantics, thereby facilitating effective dense alignment.

3. Method

In this section, we present DenseRC, a framework designed to construct dense representations that are structurally consistent with the global semantics in CLIP.

3.1. Overview

As illustrated in the upper left of Fig. 2, self-distillation framework [37] supervises the pooling region-level features $\mathcal{X}_{dense} \in \mathbb{R}^{N\times D}$ with the frozen *cls* token $\mathbf{c} \in \mathbb{R}^D$ extracted from the corresponding cropped image:

$$\mathcal{L}_{\text{semantics}}\big(\text{RoIAlign}(\mathcal{X}_{dense}, b),\, \mathbf{c}\big), \tag{1}$$

where b denotes a sampled region proposal, N is the number of patches, and D is the feature dimension. Although effective, the paradigm leaves a fundamental question largely unexplored: *what* visual semantics does the *cls* token actually encode? We answer this by analyzing the internal semantics encoded in the *cls* token and derive a principled guideline for constructing dense representations.

3.2. Semantics encoded in the *cls* Token

In CLIP's vision transformer, the *cls* token is updated at each layer l through self-attention and an MLP:

$$\mathbf{c}^{(l)} = \mathbf{c}^{(l-1)} + \text{Proj}\Big(\text{Attn}^{(l)} \cdot \mathbf{v}^{(l)}\Big), \tag{2}$$

$$\mathbf{c}^{(l)} = \mathbf{c}^{(l)} + \text{MLP}\Big(\mathbf{c}^{(l)}\Big), \tag{3}$$

where $\mathbf{v}^{(l)} \in \mathbb{R}^{N\times D}$ are the value embeddings. The MLP blocks mainly act as intra-token transformations, thus, to analyze where visual information originates, we isolate the attention pathway and rewrite it as:

$$\mathbf{c}^{(l)} \leftarrow \mathbf{c}^{(l-1)} + A^{(l)}\mathbf{v}^{(l)}, \tag{4}$$

where $A^{(l)}$ absorbs attention weights and projection parameters, and $\leftarrow$ denotes information flow.

Unrolling over all L layers (excluding the constant initialization $\mathbf{c}^{(0)}$, which is data-agnostic) yields:

$$\mathbf{c}^{(L)} \leftarrow \sum_{l=1}^{L} A^{(l)}\mathbf{v}^{(l)}. \tag{5}$$

This decomposition reveals that the the final *cls* token fundamentally aggregates information from multi-layer value embeddings. Motivated by this, we design the dense representations of student model to align with this aggregation form:

$$\mathcal{X}_{\text{dense}} = \sum_{l=1}^{L} A_p^{(l)} \cdot \mathbf{v}_p^{(l)}, \tag{6}$$

where $\mathbf{v}_p^{(l)}$ denotes value features, and $A_p^{(l)}$ is a construction module.

3.3. Head-Selective Gating

Sec. 3.2 establishes the multi-layer value embeddings as a semantically consistent basis for dense feature construction. In pursuit of dense representations that facilitate semantic alignment while improving local discrimination, we examine *how* to integrate these embeddings. In the original CLIP architecture, multi-head self-attention (MHSA) jointly performs feature aggregation across spatial and head dimensions. To investigate its role in dense alignment, we explicitly decompose the process along the spatial and head dimensions and analyze each component separately.

The Necessity of Spatial Aggregation for A_p. As observed in Fig. 1, spatial aggregation would fuse unrelated semantics. To theoretically understand its effect on alignment, we consider the following setup. Let the student produce patch embeddings $\{v_j\}_{j=1}^N \in \mathbb{R}^D$, and let the teacher provide a target representation c_r for region r. Without loss of generality, we assume:

$$c_r = \sum_{l=1}^{L} M^{(l)} v_r^{(l)} + \varepsilon_r, \tag{7}$$

where $M^{(l)} \in \mathbb{R}^{D\times D}$ captures the teacher–student representation discrepancy, and ε_r is a small residual accounting for mismatched semantics or teacher noise. For simplicity, we present the derivation for a **single layer** of value features. The extension to multiple layers is straightforward and omitted. We analyze two construction strategies:

- **Spatial–head fusion (S)** aggregates features across spatial positions:

$$\tilde{v}_r^{(S)} = \sum_{j=1}^{N} S_{rj} P v_j, \tag{8}$$

  where S_{rj} is the spatial aggregation weight ($\sum_j S_{rj} = 1$) and $P \in \mathbb{R}^{D\times D}$ is a learnable projection.
- **Head-only Modeling (H)** transforms each patch independently:

$$\tilde{v}_r^{(H)} = W v_r, \tag{9}$$

  where W performs head-wise reweighting without spatial mixing.

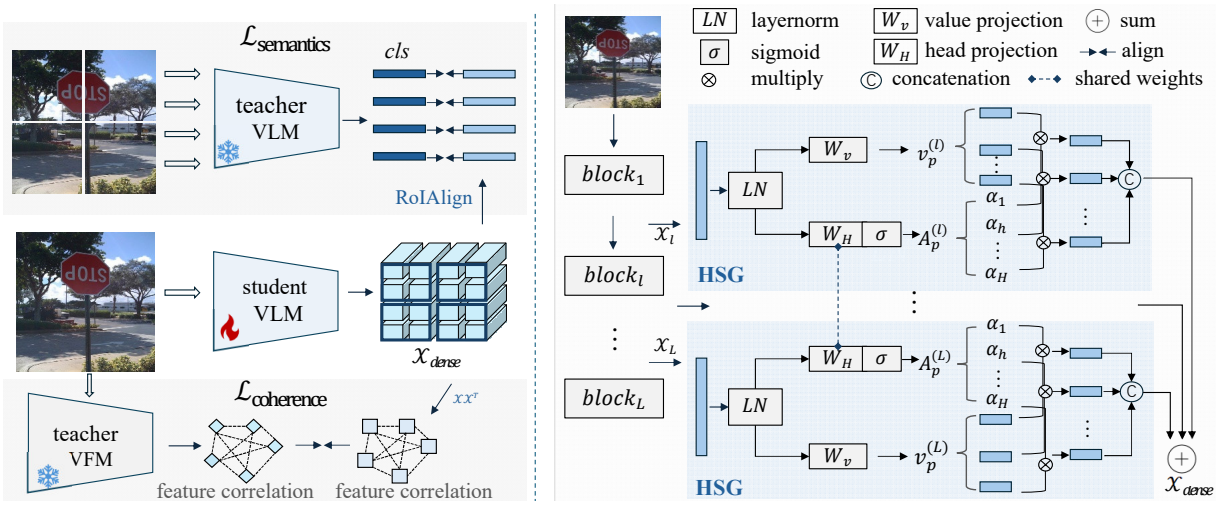

Figure 2. Overview of the proposed DenseRC framework. **Left**: DenseRC jointly imposes local semantic alignment and feature coherence distillation on dense representations $\mathcal{X}_{dense}$. **Right**: The complete pipeline for dense representation construction, featuring the Head-Selective Gating (HSG) module. W_H is applied **token-wise**, not pooled, producing $A_p \in \mathbb{R}^{N\times H}$.

Using mean-squared error for the alignment process, the expected losses are (See the Appendix for details):

$$L^{(H)} = \mathbb{E}\|(W - M)v_r\|^2 + \mathbb{E}\|\varepsilon_r\|^2, \tag{10}$$

$$L^{(S)} = \mathbb{E}\|(S_{rr}P - M)v_r\|^2 + \underbrace{\sum_{j\in\mathcal{U}_r} S_{rj}^2\, \mathbb{E}\|Pv_j\|^2}_{\Delta L} + \mathbb{E}\|\varepsilon_r\|^2, \tag{11}$$

where $\mathcal{U}_r$ indexes spatial locations unrelated to region r. Since both W and $S_{rr}P$ are learnable with sufficient capacity, gradient-based optimization can in principle reduce $\mathbb{E}\|(W-M)v_r\|^2$ and $\mathbb{E}\|(S_{rr}P-M)v_r\|^2$ to similarly small values. The key difference arises from the additional interference term ΔL, which is strictly positive when spatial aggregation assigns weight to off-target tokens:

$$L^{(H)} < L^{(S)} \quad \text{if } \exists j \in \mathcal{U}_r,\ S_{rj} \neq 0. \tag{12}$$

This underscores an inherent limitation of spatial aggregation in this task: its inability to perfectly isolate relevant patches inevitably leads to the incorporation of off-target tokens, thereby increasing alignment error. Moreover, empirical observations (Fig. 3) show that spatial mixing affects patch gradient distribution. With spatial aggregation, a few dominant patches capture the majority of gradient flow, exhibiting extremely large magnitudes and leading to a low average magnitude (0.26). In contrast, head-only modeling produces more uniform gradients across patches (mean: 0.7), providing denser supervision and more stable learning.

Head-Selective Gating (HSG). Having established that spatial aggregation is unnecessary in A_p, we now explore the need for head-level modeling by analyzing the *head-wise characteristics* of the CLIP visual encoder. We focus on two aspects critical for dense representations, with both metrics computed over 2,000 randomly sampled images from the COCO dataset.

(1) Contribution to cross-modal semantics. We quantify each head's importance for cross-modal alignment by summing its attention weights toward the *cls* token layers. The resulting semantic importance scores for layers are shown in Fig. 4 (Upper).

(2) Locality preference. To characterize the spatial focus of each head, we compute its *mean attention distance* following [29]. This metric averages pairwise pixel distances weighted by attention scores. A smaller value indicates a strong local bias, whereas a larger value implies a global-receptive field. Results for representative layers are presented in Fig. 4 (Bottom).

As shown in Fig. 4, different heads exhibit substantially diverse behaviors in both semantic contribution and locality preference, demonstrating that representation heads are functionally heterogeneous. This suggests that uniformly aggregating all heads may be suboptimal and motivates explicit head-wise weighting for dense representation construction. To this end, we introduce a lightweight **Head-Selective Gating (HSG)** module that employs learnable gating to adaptively reweight head-wise features. Given input features $x \in \mathbb{R}^{N\times D}$ of a Transformer layer, the value embeddings and head gates are computed as:

$$v = W_v(\mathrm{LN}(x)) \in \mathbb{R}^{H\times N\times (D/H)}, \tag{13}$$

$$A_p = \sigma(W_H(\mathrm{LN}(x))) \in \mathbb{R}^{N\times H}, \tag{14}$$

where W_v is the original CLIP value projection, $W_H \in$

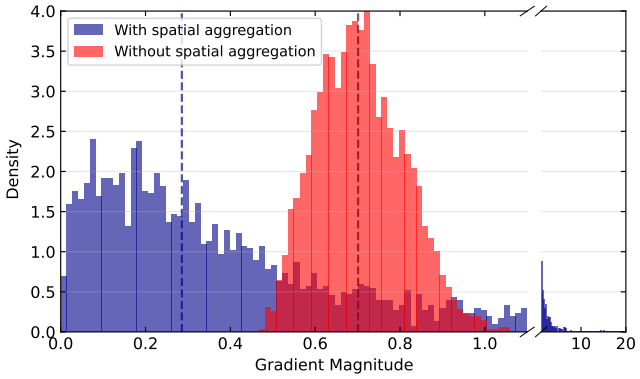

Figure 3. Effect of spatial aggregation on patch gradient distribution. With spatial aggregation, a few dominant patches capture the majority of gradient flow, exhibiting extremely large magnitudes (10) while the average gradient magnitude remains low (0.26). As a result, many patches that require optimization receive little supervision. In contrast, removing spatial aggregation yields a more uniform gradient distribution, providing denser signals for effective dense cross-modal alignment.

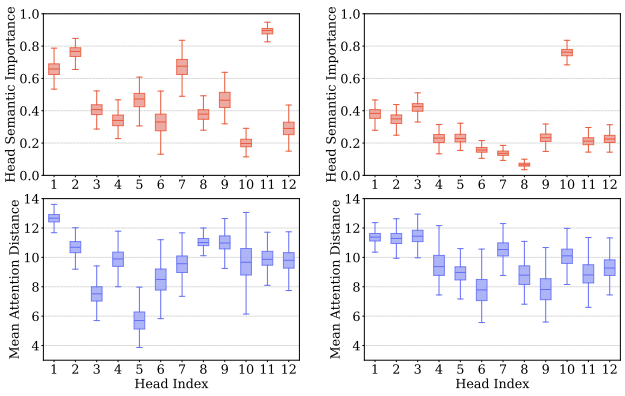

Figure 4. Head-wise characteristics of the CLIP visual encoder. The two columns present results for different layers and are not intended for cross-column comparison. **Upper**: Contribution of each head to the *cls* token, highlighting differences in cross-modal semantic importance. **Bottom**: Head-wise mean attention distance, indicating variability in locality preference across heads.

$\mathbb{R}^{D\times H}$ is a learnable head projection implemented as a single linear layer, and $\sigma(\cdot)$ is the sigmoid function producing per-head gating weights. W_H is shared across all Transformer layers, introducing minimal additional parameters. By operating solely along the head dimension, HSG modulates cross-modal semantics and local preference to form discriminative and alignment-friendly representations. The right part of Fig. 2 illustrates the overall architecture of the proposed HSG module.

Overall Objective. Our overall training objective combines a dense semantic alignment loss $\mathcal{L}_{\text{semantics}}$ with a feature coherence distillation loss $\mathcal{L}_{\text{coherence}}$, designed to jointly work on the dense representation $\mathcal{X}_{dense}$, as illustrated in the left of Fig. 2. The latter enhances the former by distilling patch-wise feature correlations from a frozen DINOv2 model, which encourages the dense representations to exhibit intra-class compactness and inter-class separability. The overall objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{semantics}} + \lambda\,\mathcal{L}_{\text{coherence}}, \tag{15}$$

where λ is a trade-off coefficient that balances semantic alignment and spatial structure distillation.

4. Experiments

4.1. Implementation Details

All experiments are conducted on 4 NVIDIA A100 GPUs with a batch size of 4 per GPU. We train the model for 6 epochs using the AdamW optimizer [24] with a learning rate of 1×10^{-5} and a weight decay of 0.1. Following standard practice of self-distillation methods [35, 37], distillation is performed on the COCO *train2017* split [1]. During training, each image is partitioned into $m \times n$ grids, where m and n are randomly sampled from $\{1,\ldots,6\}$ for local alignment distillation. To ensure a fair comparison with prior work [35, 37], we adopt cosine similarity for the semantic alignment loss $\mathcal{L}_{\text{semantics}}$, and employ the mean squared error (MSE) for the feature coherence distillation loss $\mathcal{L}_{\text{coherence}}$. The trainable parameters in the student VLM are kept the same as in [35]. Unless otherwise specified, input images are resized to 1024×1024 for distillation, and the loss balancing coefficient is set to $\lambda = 0.025$ in Eq. (15). For v_p, we extract the value embeddings from the last three Transformer layers, which empirically yields the most stable and accurate alignment (see Sec. 4.3 for analysis).

4.2. Open-vocabulary Dense Prediction

We evaluate our approach across multiple downstream OVDP tasks, including region classification, object detection, and semantic segmentation. In all settings, we simply replace CLIP with our DenseRC in the downstream pipelines, without modifying any of their hyperparameters (See the Appendix for more details).

Open-vocabulary Region Classification. We conduct dense-level zero-shot classification to assess region recognition capability, following the protocol in [37]. Region features are extracted from the COCO dataset using three standard settings: *(i) Boxes*: RoIAlign pooled from object bounding boxes; *(ii) Thing Masks*: mask pooling over instance regions; and *(iii) Stuff Masks*: mask pooling over stuff regions, where the latter two are derived from COCO Panoptic annotations. For a fair comparison, we also implement DenseRC using region proposals generated by an RPN pretrained on COCO *train2017*, following [37]. Results are

reported in Tab. 1 using Top-1 and Top-5 mean Accuracy (mAcc). As shown, DenseRC consistently outperforms existing baselines by a significant margin across all region settings and training protocols, demonstrating stronger open-vocabulary region discrimination and improved alignment.

Open-vocabulary Detection. We evaluate our method for open-vocabulary object detection on the OV-COCO and OV-LVIS benchmarks using the two-stage detector F-ViT, following standard practice [37]. For OV-COCO, we report mean Average Precision (mAP) at an Intersection-over-Union (IoU) threshold of 0.5 on novel categories. OV-LVIS contains 1,203 categories, where only 461 common and 405 frequent classes are available for training, while evaluation is performed on common, frequent, and rare categories. Following prior works [37], we report mAP on the *rare* split averaged across IoU thresholds from 0.5 to 0.95.

As shown in Tab. 2, DenseRC outperforms the previous state-of-the-art method, DeCLIP, by 4.5% mAP on OV-COCO and 2.2% mAP on OV-LVIS [10] using a CLIP ViT-B/16 backbone, demonstrating stronger capability in recognizing novel objects. To further evaluate cross-dataset generalization, we directly transfer models trained on OV-LVIS to COCO and Object365 [30] validation sets without any additional fine-tuning. Results in Tab. 3 show that our method consistently surpasses existing approaches, validating its superior robustness and cross-domain transferability.

Open-Vocabulary Semantic Segmentation. Following prior work [27, 35, 37], we evaluate DenseRC on open-vocabulary semantic segmentation (OVSS) using the state-of-the-art framework CAT-Seg. The segmentation model is trained on COCO-Stuff and evaluated across multiple standard OVSS benchmarks. We adopt mean Intersection-over-Union (mIoU) as the evaluation metric, following prior OVSS works [4]. As shown in Tab. 4, DenseRC consistently outperforms existing methods on all datasets, setting a new state of the art. These results further demonstrate the strong dense recognition capacity and superior generalization ability of our approach.

4.3. Ablation Study

v_p and A_p. We first ablate the design of the v_p. DenseRC employs value embeddings from multiple Transformer layers (termed *multi-v*), while prior methods construct dense representations $\mathcal{X}_{dense}$ using either the final spatial features [37] of CLIP (denoted as $x{+}v$) or a single last-layer value embedding [35] (denoted as v_L). To ensure a fair comparison, no attention-based module is applied in this experiment. We evaluate all variants on both open-vocabulary detection (OVD) and open-vocabulary semantic segmentation (OVSS). As shown in Tab. 5, *multi-v* consistently outperforms alternative designs across all benchmarks, validating that employing multi-layer value embeddings facilitates more effective dense semantic alignment.

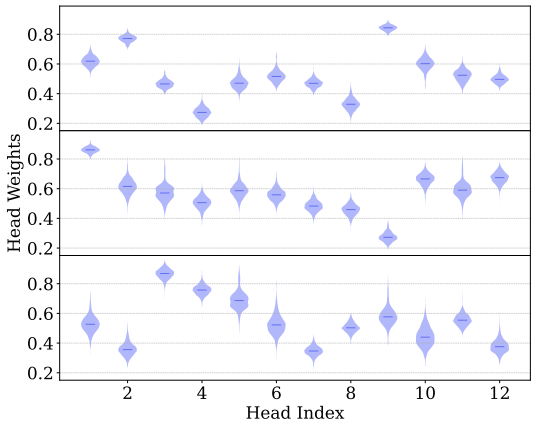

Figure 5. Head weights learned by HSG in the last three Transformer layers. The weights, computed over 2,000 COCO *val2017* samples, exhibit strong cross-sample consistency, revealing HSG effectively characterizes the intrinsic functional roles of different feature heads rather than image-specific patterns. The uneven distributions across heads validate the necessity of explicit head-wise reweighting for dense representation construction.

We further analyze the impact of the A_p by comparing four variants: (i) no attention (remove the attention block), (ii) vanilla self-attention (Q-K) as in CLIP, (iii) self-self attention (Q-Q) from DeCLIP, and (iv) our proposed HSG. All configurations are evaluated on identical OVD and OVSS benchmarks for a fair comparison. Results in Tab. 5 show that introducing spatial aggregation, either self-attention or self-self attention, leads to performance degradation compared to the no-attention baseline, corroborating our analysis in Sec. 3.3. In contrast, HSG selectively models head-level characteristics without entangling noisy spatial correlations, yielding the best performance across all benchmarks. These results highlight its effectiveness in constructing discriminative dense representations for open-vocabulary perception.

Fig. 5 visualizes the head weights generated by the HSG module from the last three Transformer layers of the student visual encoder, using 2,000 randomly sampled images from the COCO *val2017*. The results show strong cross-sample consistency in the learned head weights, indicating that HSG captures stable and semantically meaningful head preferences rather than image-specific patterns. This demonstrates that the head gating mechanism effectively characterizes the intrinsic functional roles of different feature heads, enabling reliable and structured feature construction. Moreover, the uneven weights distribution across heads reveals clear functional specialization, highlighting the necessity of explicit head-wise weighting instead of uniform aggregation in dense representation construction.

Table 1. **Zero-shot region classification of dense representation.** We report Top1 and Top5 mean accuracy.

Method	Backbone	RPN Proposals	Boxes		Thing Masks		Stuff Masks	
			Top1	Top5	Top1	Top5	Top1	Top5
EVA-CLIP[32]	ViT-B/16	-	18.2	33.2	20.6	36.5	18.4	43.5
CLIPSelf[37]	ViT-B/16	✗	72.1	91.3	74.4	91.8	46.8	80.2
R-SC-CLIPSelf[27]	ViT-B/16	✗	76.0	93.1	76.2	92.5	53.5	84.4
DenseRC (Ours)	ViT-B/16	✗	**76.7**	**93.9**	**78.1**	**94.1**	**55.8**	**85.9**
R-SC-RegionText[27]	ViT-B/16	✓	72.0	91.3	74.3	91.6	41.6	73.3
CLIPSelf[37]	ViT-B/16	✓	74.0	92.6	76.3	92.8	36.8	75.0
R-SC-CLIPSelf[27]	ViT-B/16	✓	77.3	94.0	78.9	94.2	52.6	83.9
DenseRC (Ours)	ViT-B/16	✓	**78.2**	**94.8**	**79.8**	**94.8**	**56.2**	**86.9**
EVA-CLIP[32]	ViT-L/14	-	56.7	78.0	59.0	79.8	20.8	41.9
CLIPSelf[37]	ViT-L/14	✗	77.1	93.3	78.7	93.7	44.4	78.3
R-SC-CLIPSelf[27]	ViT-L/14	✗	82.9	96.0	82.8	95.6	57.8	86.5
DenseRC (Ours)	ViT-L/14	✗	**83.2**	**96.6**	**85.1**	**96.8**	**58.2**	**87.5**
CLIPSelf[37]	ViT-L/14	✓	77.8	94.0	80.4	94.5	34.0	71.8
R-SC-CLIPSelf[27]	ViT-L/14	✓	81.7	95.8	82.9	95.9	52.5	83.9
DenseRC (Ours)	ViT-L/14	✓	**82.9**	**96.9**	**84.9**	**97**	**61**	**89.2**

Table 2. Comparison with state-of-the-art methods for open-vocabulary object detection. Caption supervision denotes that the model is trained with additional image–text pairs, while CLIP supervision indicates semantic transfer from the original CLIP model. FineCLIP leverages CC2.5M to generate region–text pairs as supervision.

(a) OV-COCO benchmark

Method	Supervision	Backbone	AP_{50}^{Novel}
ViLD [9]	CLIP	RN50	27.6
Detic [53]	Caption	RN50	27.8
OV-DETR [47]	CLIP	RN50	29.4
BARON-KD [36]	CLIP	RN50	34.0
SAS-Det [50]	CLIP	RN50	37.4
OV-DQUO [34]	CLIP	RN50	39.2
RegionCLIP [51]	Captions	RN50x4	39.3
CORA [39]	CLIP	RN50x4	41.7
F-ViT+FineCLIP [14]	CC2.5M	ViT-B/16	29.8
F-ViT+CLIPSelf [37]	CLIP	ViT-B/16	37.6
F-ViT [37]+R-SC-CLIPSelf[27]	CLIP	ViT-B/16	40.9
F-ViT [37]+DeCLIP[35]	CLIP	ViT-B/16	41.1
F-ViT [37]+**DenseRC (Ours)**	CLIP	ViT-B/16	**45.6**
RO-ViT [16]	CLIP	ViT-L/16	33.0
CFM-ViT [15]	CLIP	ViT-L/16	34.1
F-ViT[37]+FineCLIP [14]	CC2.5M	ViT-L/14	40.0
F-ViT[37]+CLIPSelf [37]	CLIP	ViT-L/14	44.3
F-ViT [37]+R-SC-CLIPSelf[27]	CLIP	ViT-L/14	48.1
F-ViT [37]+DeCLIP[35]	CLIP	ViT-L/14	46.2
F-ViT [37]+**DenseRC (Ours)**	CLIP	ViT-L/14	**54.8**

(b) OV-LVIS benchmark

Method	Supervision	Backbone	mAP_r
ViLD [9]	CLIP	RN50	16.3
OV-DETR [47]	CLIP	RN50	17.4
BARON-KD [36]	CLIP	RN50	22.6
RegionCLIP [51]	Caption	RN50x4	22.0
OV-SAM [46]	CLIP	RN50x16	24.0
$CORA^+$ [39]	Caption	RN50x4	28.1
F-VLM [17]	CLIP	RN50x64	32.8
Detic [53]	Caption	Swin-B	33.8
F-ViT+FineCLIP [14]	CC2.5M	ViT-B/16	10.4
F-ViT[37]+CLIPSelf [37]	CLIP	ViT-B/16	25.3
F-ViT [37]+R-SC-CLIPSelf[27]	CLIP	ViT-B/16	27.5
F-ViT [37]+DeCLIP[35]	CLIP	ViT-B/16	26.8
F-ViT [37]+**DenseRC (Ours)**	CLIP	ViT-B/16	**29.0**
RO-ViT [16]	CLIP	ViT-H/16	34.1
F-ViT+FineCLIP [14]	CC2.5M	ViT-L/14	20.2
F-ViT[37]+CLIPSelf [37]	CLIP	ViT-L/14	34.9
F-ViT [37]+R-SC-CLIPSelf[27]	CLIP	ViT-L/14	37.2
F-ViT [37]+DeCLIP[35]	CLIP	ViT-L/14	37.2
F-ViT [37]+**DenseRC (Ours)**	CLIP	ViT-L/14	**39.6**

Table 3. Zero-shot cross-dataset transfer evaluation of the LVIS-trained detector on COCO and Objects365.

Method	COCO			Objects365 [30]		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Supervised Baseline [9]	46.5	67.6	50.9	25.6	38.6	28.0
ViLD [9]	36.6	55.6	39.6	11.8	18.0	12.6
DetPro [6]	34.9	53.8	37.4	12.1	18.8	12.9
BARON [36]	36.2	55.7	39.1	13.6	21.0	14.5
F-VLM [17]	37.9	61.6	41.2	16.2	27.4	17.5
CoDet [25]	39.1	57.0	42.3	14.2	20.5	15.3
RO-ViT [16]	-	-	-	17.7	27.4	19.1
F-ViT[37]+FineCLIP [14]	33.6	52.7	36.1	12.1	19.8	12.6
F-ViT[37]+CLIPSelf [37]	40.5	63.8	44.3	19.5	31.3	20.7
F-ViT[37]+DeCLIP[35]	41.0	64.6	44.8	20.0	32.2	21.2
F-ViT[37]+**DenseRC (Ours)**	**43.4**	**66.6**	**47.6**	**20.8**	**32.4**	**22.6**

The layers for v_p. We conduct ablation studies on the layers for v_p, with results summarized in Tab. 6. Compared to the baseline, aggregating the last two or three layers improves performance. We find that using the last three Transformer layers yields the best results for dense feature construction, while incorporating more layers leads to degradation. This trend aligns with the observation in [23] that the fourth-to-last layer exhibits a notable decline in semantic alignment. Early layers tend to have a lower *signal-to-noise* ratio, making them less suitable for constructing dense representations. Moreover, aggregating these earlier layers, which capture heterogeneous features with the deep semantics, potentially results in gradient interference and reduced optimization stability. Based on these findings, we

Table 4. Comparison with state-of-the-art methods on open-vocabulary semantic segmentation across multiple benchmarks.

Method	Backbone	Training Set	A-847	PC-459	A-150	PC-59
ZegFormer [5]	ViT-B/16	COCO-Stuff	5.6	10.4	18.0	45.5
ZSseg [43]	ViT-B/16	COCO-Stuff	7.0	-	20.5	47.7
CAT-Seg [4]	ViT-B/16	COCO-Stuff	12.0	19.0	31.8	57.5
CAT-Seg[4] +FineCLIP[14]	ViT-B/16	COCO-Stuff	12.2	-	32.4	56.0
CAT-Seg[4] +CLIPSelf[37]	ViT-B/16	COCO-Stuff	9.3	-	29.0	58.0
CAT-Seg[4] +R-SC-CLIPSelf[37]	ViT-B/16	COCO-Stuff	12.2	-	32.0	57.2
CAT-Seg[4] +DeCLIP[35]	ViT-B/16	COCO-Stuff	15.3	21.4	36.3	60.6
CAT-Seg[4] +**DenseRC (Ours)**	ViT-B/16	COCO-Stuff	**15.9**	**22.7**	**37.6**	**61.3**
OVSeg [20]	ViT-L/14	COCO-Stuff	9.0	12.4	29.6	55.7
SAN [44]	ViT-L/14	COCO-Stuff	13.7	17.1	33.3	60.2
ODISE [42]	ViT-L/14	COCO-Panoptic	11.1	14.5	29.9	57.3
MAFT [12]	ConvNeXt-L	COCO-Stuff	13.1	17.0	34.4	57.5
FC-CLIP [45]	ConvNeXt-L	COCO-Panoptic	14.8	18.2	34.1	58.4
FrozenSeg [3]	ConvNeXt-L	COCO-Panoptic	14.8	19.7	34.4	-
CAT-Seg [4]	ViT-L/14	COCO-Stuff	16.0	23.8	37.9	63.3
CAT-Seg[4] +FineCLIP[14]	ViT-L/14	COCO-Stuff	14.1	-	36.1	59.9
CAT-Seg[4] +CLIPSelf[37]	ViT-L/14	COCO-Stuff	12.4	-	34.5	62.3
CAT-Seg[4] +R-SC-V[27]	ViT-L/14	COCO-Stuff	16.6	-	38.4	63.6
CAT-Seg[4] +DeCLIP[35]	ViT-L/14	COCO-Stuff	17.6	25.9	40.7	63.9
CAT-Seg[4] +**DenseRC (Ours)**	ViT-L/14	COCO-Stuff	**18.4**	**26.7**	**41.5**	**64**

Table 5. Ablation study on v_p and A_p. Each component is varied independently while keeping the other fixed. mAP is reported for OV-COCO, and mIoU for segmentation benchmarks. Our default method is highlighted in blue.

v_p	A-847	PC-459	A-150	OV-COCO
$x + v$	15.2	21.7	36.6	41.3
v	15.4	22.1	36.7	41.9
multi-v	15.7	22.3	37.2	44.4
A_p	A-847	PC-459	A-150	OV-COCO
no attention	15.7	22.3	37.2	44.4
self attention	14.7	21.3	36.0	39.1
self-self attention	15.4	21.9	36.7	42.3
HSG	15.9	22.7	37.6	45.6

Table 6. Ablation study on layers in v_p.

l	A-847	PC-459	A-150	OV-COCO
L	15.6	22.4	37.0	42.3
$L-1$	15.7	22.6	37.3	44.2
$L-2$	15.9	22.7	37.6	45.6
$L-3$	15.7	22.6	37.3	43.8
$L-4$	15.7	22.3	37.1	43.6

Table 7. Ablation study on λ.

λ	A-847	PC-459	A-150	PC-59
0.01	15.8	22.6	37.3	61.1
0.025	15.9	22.7	37.6	61.3
0.04	15.7	22.3	37.3	61.4
0.05	15.4	22.2	36.9	61.3
0.1	15.4	22.1	36.8	61.0

adopt the last three value embeddings in v_p.

λ. We ablate the effect of the balancing coefficient λ, which controls the trade-off between semantic alignment and feature coherence distillation in DenseRC. We evaluate performance on OVSS, with results reported in Tab. 7. The best overall performance is achieved at $\lambda = 0.025$.

5. Conclusion

In this paper, we present DenseRC, a self-distillation framework that advances dense representation construction for adapting CLIP to open-vocabulary dense perception. We investigate the internal semantics encoded in the *cls* token and reveal that multi-layer value embeddings offer a semantically consistent basis for building dense features. By decoupling the attention mechanism along spatial and head dimensions, we theoretically show that spatial aggregation introduces interference from unrelated regions, leading to increased alignment error during semantic transfer. Motivated by the functional heterogeneity across feature heads, we further propose a lightweight head-selective gating module to adaptively model head-wise characteristics, enabling more discriminative dense representations. Extensive experiments demonstrate that DenseRC achieves significant and consistent performance improvements across multiple open-vocabulary dense downstream tasks, validating the effectiveness and generalizability of our design.

Acknowledgements

This work is partly supported by the National Key Research and Development Plan (2024YFB3309302), National Natural Science Foundation of China (82441024), the Beijing Natural Science Foundation (L251073), the Research Program of State Key Laboratory of Complex and Critical Software Environment, and the Fundamental Research Funds for the Central Universities.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 5, 2

[2] Fangyi Chen, Han Zhang, Zhantao Yang, Hao Chen, Kai Hu, and Marios Savvides. Rtgen: Generating region-text pairs for open-vocabulary object detection. *arXiv preprint arXiv:2405.19854*, 2024. 2

[3] Xi Chen, Haosen Yang, Sheng Jin, Xiatian Zhu, and Hongxun Yao. Frozenseg: Harmonizing frozen foundation models for open-vocabulary segmentation. *arXiv preprint arXiv:2409.03525*, 2024. 8

[4] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024. 2, 6, 8

[5] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 8

[6] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 7, 2

[7] 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(2):303–338, 2010. 2

[8] Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. In *The Twelfth International Conference on Learning Representations*, 2023. 4

[9] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2, 7

[10] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 6, 1, 2

[11] Joonhyun Jeong, Geondo Park, Jayeon Yoo, Hyungsik Jung, and Heesu Kim. Proxydet: Synthesizing proxy novel classes via classwise mixup for open-vocabulary object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2462–2470, 2024. 2

[12] Siyu Jiao, Yunchao Wei, Yaowei Wang, Yao Zhao, and Humphrey Shi. Learning mask-aware clip representations for zero-shot segmentation. *Advances in Neural Information Processing Systems*, 36:35631–35653, 2023. 8

[13] Siyu Jiao, Hongguang Zhu, Jiannan Huang, Yao Zhao, Yunchao Wei, and Humphrey Shi. Collaborative vision-text representation optimizing for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 399–416. Springer, 2025. 2

[14] Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Wei Wei, Huiwen Zhao, Zhiwu Lu, et al. Fineclip: Self-distilled region-based clip for better fine-grained understanding. *Advances in Neural Information Processing Systems*, 37:27896–27918, 2024. 1, 2, 7, 8

[15] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Contrastive feature masking open-vocabulary vision transformer. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15556–15566, 2023. 7, 2

[16] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11144–11154, 2023. 7, 2

[17] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022. 7, 2

[18] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *European Conference on Computer Vision*, pages 143–160. Springer, 2024. 1, 2, 4

[19] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv e-prints*, pages arXiv–2304, 2023. 4

[20] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 8

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2

[22] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 3

[23] Yajie Liu, Guodong Wang, Jinjin Zhang, Qingjie Liu, and Di Huang. Unveiling the knowledge of clip for training-free open-vocabulary semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5649–5657, 2025. 1, 7

[24] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[25] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 7

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez,

Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[27] Congpei Qiu, Yanhao Wu, Wei Ke, Xiuxiu Bai, and Tong Zhang. Refining clip's spatial awareness: A visual-centric perspective. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 6, 7, 8

[28] 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 *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3

[29] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021. 4

[30] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 6, 7, 1, 2

[31] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 3

[32] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 7

[33] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 3

[34] Junjie Wang, Bin Chen, Bin Kang, Yulin Li, Weizhi Xian, Yichi Chen, and Yong Xu. Ov-dquo: Open-vocabulary detr with denoising text query training and open-world unknown objects supervision. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7762–7770, 2025. 7, 2

[35] Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. Declip: Decoupled learning for open-vocabulary dense perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14824–14834, 2025. 1, 2, 5, 6, 7, 8

[36] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15254–15264, 2023. 7, 2

[37] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. CLIPSelf: Vision transformer distills itself for open-vocabulary dense prediction. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 3, 5, 6, 7, 8

[38] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Wentao Liu, and Chen Change Loy. Clim: Contrastive language-image mosaic for region representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6117–6125, 2024. 2

[39] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7031–7040, 2023. 7, 2

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

[41] Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment. In *Forty-second International Conference on Machine Learning*, 2025. 1, 2

[42] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 8

[43] 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*, pages 736–753. Springer, 2022. 8

[44] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 8

[45] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36, 2024. 8

[46] Haobo Yuan, Xiangtai Li, Chong Zhou, Yining Li, Kai Chen, and Chen Change Loy. Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively. In *ECCV*, 2024. 7

[47] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, pages 106–122. Springer, 2022. 7, 2

[48] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 3

[49] Heng Zhang, Qiuyu Zhao, Linyu Zheng, Hao Zeng, Zhiwei Ge, Tianhao Li, and Sulong Xu. Exploring region-word alignment in built-in detector for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16975–16984, 2024. 2

[50] Shiyu Zhao, Samuel Schulter, Long Zhao, Zhixing Zhang, Vijay Kumar B G, Yumin Suh, Manmohan Chandraker, and Dimitris N. Metaxas. Taming self-training for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13938–13947, 2024. 7, 2

[51] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 7, 2

[52] 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(3):302–321, 2019. 2

[53] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 7, 2