InfoCLIP: Bridging Vision-Language Pretraining and Open-Vocabulary Semantic Segmentation via Information-Theoretic Alignment Transfer

Muyao Yuan[1,2], Yuanhong Zhang[1,3], Weizhan Zhang[1,2*], Lan Ma[4], Yuan Gao[4], Jiangyong Ying[5], Yudeng Xin[6]

[1]School of Computer Science and Technology, Xi'an Jiaotong University
[2]Ministry of Education Key Laboratory of Intelligent Networks and Network Security, Xi'an Jiaotong University
[3]Shaanxi Province Key Laboratory of Big Data Knowledge Engineering, Xi'an Jiaotong University
[4]China Telecom
[5]China Telecom E-surfing Vision Technology Co., Ltd
[6]Faculty of Engineering and Information Technology, University of Melbourne
{yuanmuyao,yuanhongzhang}@stu.xjtu.edu.cn, zhangwzh@xjtu.edu.cn,
{malan,gaoy97,yingjiangyong}@chinatelecom.cn, xinyudeng005@gmail.com

Abstract

Recently, the strong generalization ability of CLIP has facilitated open-vocabulary semantic segmentation, which labels pixels using arbitrary text. However, existing methods that fine-tune CLIP for segmentation on limited seen categories often lead to overfitting and degrade the pretrained vision-language alignment. To stabilize modality alignment during fine-tuning, we propose InfoCLIP, which leverages an information-theoretic perspective to transfer alignment knowledge from pretrained CLIP to the segmentation task. Specifically, this transfer is guided by two novel objectives grounded in mutual information. First, we compress the pixel-text modality alignment from pretrained CLIP to reduce noise arising from its coarse-grained local semantic representations learned under image-text supervision. Second, we maximize the mutual information between the alignment knowledge of pretrained CLIP and the fine-tuned model to transfer compact local semantic relations suited for the segmentation task. Extensive evaluations across various benchmarks validate the effectiveness of InfoCLIP in enhancing CLIP fine-tuning for open-vocabulary semantic segmentation, demonstrating its adaptability and superiority in asymmetric transfer.

Project — https://muyaoyuan.github.io/InfoCLIP-Page

Introduction

Open-vocabulary semantic segmentation (OVSS) aims to build a model that assigns pixel-level semantic labels based on arbitrary text descriptions, without being restricted to a closed category set. Vision-language foundation models (Radford et al. 2021; Jia et al. 2021; Li et al. 2022b), particularly CLIP (Radford et al. 2021), are commonly fine-tuned to enable such open-vocabulary recognition. This fine-tuning process primarily serves to transfer a vision-language model pretrained on image-text alignment to the pixel-level prediction task (Cho et al. 2024; Jiao et al. 2024).

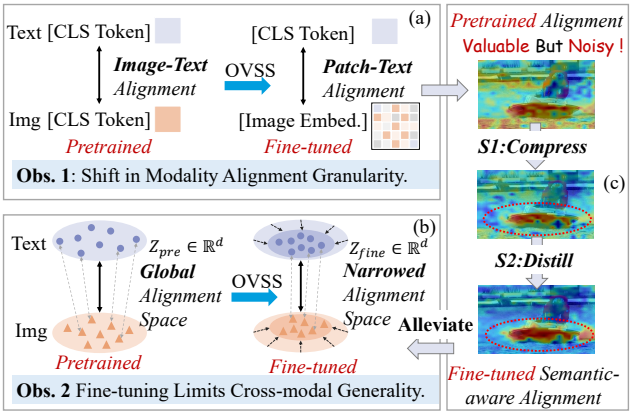


Figure 1: **Motivation of InfoCLIP.** To leverage the valuable yet noisy pixel-text modality alignment from the pretrained CLIP for enhancing OVSS, we: (1) denoise the pretrained alignment through compression to extract semantic-aware alignment; and (2) transfer more generalized semantic alignment via distillation to alleviate the narrowing of the modality alignment space.

Existing methods (Xu et al. 2023b; Jiao et al. 2023; Cho et al. 2024; Jiao et al. 2024) attempt to fine-tune CLIP on a benchmark dataset for open-vocabulary segmentation. As illustrated in Fig. 1(b), task-specific fine-tuning on limited training categories often impairs generalization by narrowing the modality alignment space. Most methods (Jiao et al. 2023, 2024; Cho et al. 2024) strive to limit alignment shift by simply freezing model parameters, which serves as an implicit preservation of the pretraining. However, this strategy is fragile (Cho et al. 2024), as modifying even a small subset of parameters can distort the feature distribution and further compromise the image-text alignment.

To preserve the pretrained modality alignment for better generalization, an intuitive approach is to extract the alignment relations from the pretrained model and transfer them to the downstream model, guiding the joint fine-tuning of

*Corresponding authors.

CLIP's image and text encoders. However, this approach faces a key challenge: as shown in Fig. 1(a), while the pretrained CLIP learns global image-text alignment, the segmentation task requires fine-grained pixel-text alignment. Specifically, CLIP constructs a coarse semantic space during image-text pretraining (Radford et al. 2021), yielding implicit patch-level semantic representations in its intermediate features. However, these representations are inherently noisy (Xie et al. 2024; Cho et al. 2024), resulting in imprecise and ambiguous pixel-text alignment, whereas semantic segmentation demands accurate and deterministic patch-level semantics. Therefore, transferring patch-level alignment information from the pretrained CLIP without denoising is ineffective and can even harm segmentation models.

Hence, this raises a two-stage challenge: *1) how to extract refined pixel-level alignment relations from noisy pretrained representations? 2) how to effectively transfer them to guide downstream fine-tuning while preserving modality alignment ?*

In response to these challenges, we propose InfoCLIP, a novel information-theoretic distillation method specifically designed for asymmetric fine-tuning of CLIP. To address the first challenge, we introduce a Learnable Pixel-Text Alignment Module (LPAM) to extract alignment relations from the pretrained model. By compressing the mutual information between the extracted relations and embeddings of different modalities, the LPAM captures refined modality alignment information. This process denoises CLIP's local semantic representations, captures key object features, and enhances the overall segmentation perception. To address the second challenge, we maximize the mutual information between the modality alignment relations extracted by the LPAM of the pretrained and fine-tuned models. Leveraging mutual information's ability to preserve structural information (Zhang et al. 2025), we transfer compact local semantics to protect the modality alignment relations within the fine-tuned model.

Extensive experiments demonstrate that InfoCLIP sets a new state of the art in open-vocabulary semantic segmentation across three benchmarks by fine-tuning CLIP, highlighting its superiority in handling asymmetric vision-language transfer.

Overall, our contribution can be summarized as follows:

- We propose InfoCLIP, the first information-theoretic framework for CLIP fine-tuning, introducing a novel distillation strategy tailored to asymmetric transfer of pretrained knowledge, significantly improving performance in open-vocabulary semantic segmentation.
- InfoCLIP introduces dual complementary mechanisms for CLIP adaptation: a bottleneck that strategically compresses noisy information in CLIP's local semantic representations, and an adaptive cross-model mutual information maximization scheme that preserves essential modality alignment.
- Extensive evaluations on diverse benchmarks show that InfoCLIP consistently surpasses state-of-the-art open-vocabulary semantic segmentation methods.

Related Work

Open-Vocabulary Semantic Segmentation

Existing open-vocabulary semantic segmentation methods leverage the pretrained CLIP (Radford et al. 2021) to ensure generalization capability. However, fine-tuning CLIP on limited training classes impairs its ability to generalize to unseen categories (Zhou, Loy, and Dai 2022a). To mitigate this issue, mask-based methods (Ghiasi et al. 2022; Xu et al. 2022, 2023a) freeze the pretrained vision-language model and introduce an auxiliary mask generator (Cheng et al. 2022) for segmentation, thereby preserving open-vocabulary capability. However, without pixel-wise fine-tuning, the feature distribution in CLIP leads to misalignment between localized visual representations and corresponding text embeddings (Liang et al. 2023). In contrast, pixel-based methods (Xu et al. 2023b; Yu et al. 2023; Cho et al. 2024) offer a more promising solution by directly fine-tuning CLIP for pixel-level predictions with a carefully crafted parameter space. Although these methods improve segmentation performance, the pretrained vision-language alignment is inevitably degraded due to the paradigm shift from image-text to pixel-text alignment. Therefore, we propose an information-theoretic framework to preserve essential modality alignment and enhance generalization.

Knowledge Distillation

Knowledge distillation (KD) is designed to transfer knowledge from a large, well-trained teacher model to a smaller, lightweight student model (Hinton, Vinyals, and Dean 2015), and is later widely used as an effective tool for transferring knowledge from a source task to a target task (Mansourian et al. 2025). For open-vocabulary semantic segmentation, MAFT (Jiao et al. 2023) introduces a self-distillation loss between the frozen CLIP and the fine-tuned IP-CLIP to preserve CLIP's transferability and mitigate overfitting. MAFT+ (Jiao et al. 2024) further proposes a representation compensation strategy that distills multi-scale features, generated through adaptive pooling, from the frozen CLIP to the fine-tuned CLIP, aiming to maintain zero-shot capability during fine-tuning. However, these approaches are tailored for mask-based methods that distill knowledge at the image level, without addressing the pixel-text alignment problem, making them unsuitable for pixel-based methods.

Preliminaries

In information theory, Rényi's α-entropy $H_\alpha(X)$ generalizes Shannon's entropy and is defined for a continuous random variable X with PDF $p(x)$ over a finite set $\mathcal{X}$ as:

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \int_{\mathcal{X}} p^\alpha(x) \mathrm{d}x, \tag{1}$$

where H_α reduces to Shannon's entropy as $\alpha \to 1$. However, its reliance on the underlying data distribution limits its applicability in high-dimensional settings. To address this, matrix-based Rényi's α-entropy enables direct computation from data without density estimation.

Definition 1. Let $\kappa : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a real-valued positive kernel that is also infinitely divisible (Bhatia 2006). Given

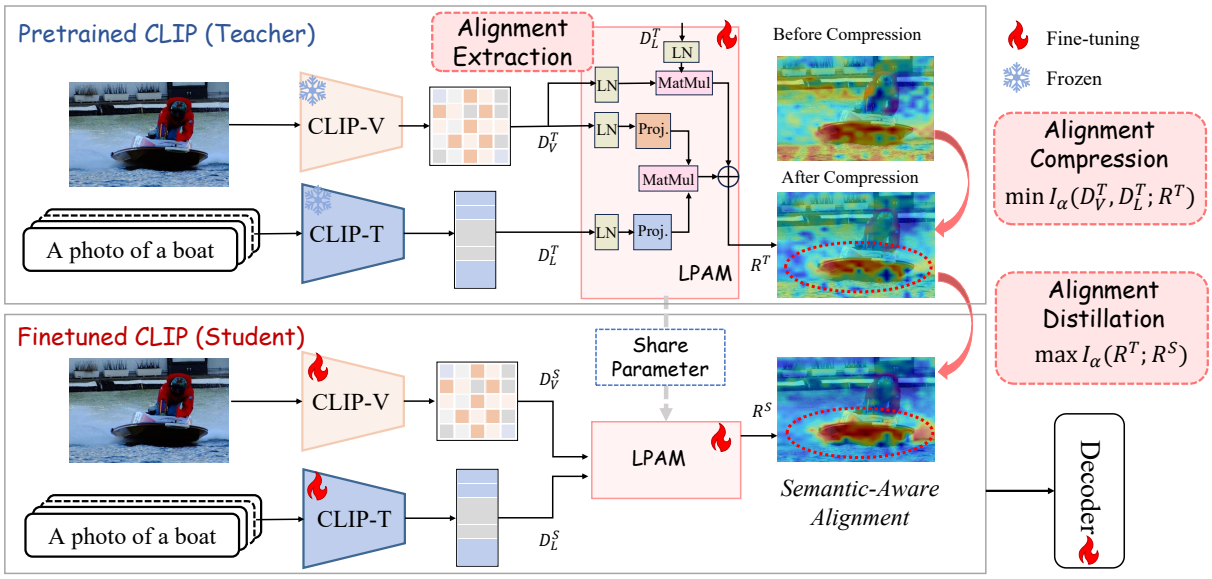


Figure 2: **Overview of InfoCLIP.** To exploit the valuable yet noisy pixel-text alignment from a pretrained foundation model (CLIP) for OVSS, InfoCLIP introduces an information-theoretic framework for asymmetric adaptation, comprising: (1) a Learnable Pixel-Text Alignment Module (LPAM) to extract fine-grained patch-text relations; (2) an information bottleneck loss to suppress noise and retain semantic-aware alignment; and (3) a mutual information transfer loss to preserve modality alignment by bridging pretrained and fine-tuned CLIP representations. The detailed formulation of LPAM is provided in the Appendix.

$\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$, each $\mathbf{x}_i$ being a real-valued scalar or vector, and the Gram matrix K obtained from $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$, a matrix-based analogue to Rényi's α-entropy can be defined as:

$$S_\alpha(A) = \frac{1}{1-\alpha} \log\left(\mathrm{tr}(A^\alpha)\right) = \frac{1}{1-\alpha} \log\left[\sum_{i=1}^n \lambda_i^\alpha(A)\right], \tag{2}$$

where $A_{ij} = \frac{1}{n} \frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}}$ is a normalized kernel matrix and $\lambda_i(A)$ denotes the i-th eigenvalue of A.

The kernel matrix $\mathbf{A}$ is positive semi-definite with $\mathrm{tr}(\mathbf{A}) = 1$, ensuring all eigenvalues $\lambda_i \in [0, 1]$. Under this setting, matrix-based Rényi's joint entropy, mutual information $I_\alpha(\mathbf{A}; \mathbf{B})$, and their multivariate extensions can be defined (Yu et al. 2019).

Definition 2. Let $\kappa_1 : \mathcal{X}^1 \times \mathcal{X}^1 \mapsto \mathbb{R}, \ldots, \kappa_L : \mathcal{X}^L \times \mathcal{X}^L \mapsto \mathbb{R}$ be positive, infinitely divisible kernels, and let $\{\mathbf{x}_i^1, \ldots, \mathbf{x}_i^L\}_{i=1}^n \subset \mathcal{X}^1 \times \cdots \times \mathcal{X}^L$ be a collection of n samples. A matrix-based analogue to Rényi's α-order joint entropy among L variables can be defined as:

$$S_\alpha(A_1, \ldots, A_L) = S_\alpha\left(\frac{A_1 \circ \cdots \circ A_L}{\mathrm{tr}(A_1 \circ \cdots \circ A_L)}\right), \tag{3}$$

where $A_1, \ldots, A_L$ are normalized kernel matrices, and $\circ$ denotes the Hadamard product. Within the settings, the mutual information between A_1 and A_2, denoted as $I_\alpha(A_1; A_2)$, is given by:

$$I_\alpha(A_1; A_2) = S_\alpha(A_1) + S_\alpha(A_2) - S_\alpha(A_1, A_2). \tag{4}$$

This formulation avoids the need for high-dimensional probability density estimation required by Shannon entropy, providing a more accurate and computationally efficient alternative (Yu et al. 2019).

Methodology

Overview of InfoCLIP

Given an image I and a set of candidate class categories $\mathcal{C} = \{T^{(n)}\}_{n=1}^{N_C}$, where $T^{(n)}$ denotes the textual description of the n-th category and N_C is the number of classes, open-vocabulary semantic segmentation aims to assign a class label to each pixel in image I (Cho et al. 2024). Unlike classical semantic segmentation tasks (He et al. 2019; Zhou et al. 2022) with a fixed label space, open-vocabulary segmentation poses an additional challenge: the category set $\mathcal{C}$ varies across different images and is defined by arbitrary text descriptions.

Fig. 2 illustrates the proposed InfoCLIP, which contains three synergistic components: (1) a Learnable Pixel-Text Alignment Module (LPAM) that extracts fine-grained alignment relations between visual patches and text embeddings from the pretrained CLIP; (2) an information bottleneck loss that compresses the extracted alignment relations to suppress noisy semantics and preserve key semantic-aware information; and (3) a mutual information transfer loss that bridges the pretrained and fine-tuned CLIP by maximizing the mutual information between their alignment representations, ensuring the preservation of modality alignment during asymmetric fine-tuning. This information-theoretic design enables InfoCLIP to retain the structured alignment knowledge of pretrained CLIP while enhancing its local semantic precision for open-vocabulary segmentation.

Learnable Pixel-Text Alignment Module (LPAM)

While the CLIP model is pretrained to capture global image-text alignment, its intermediate visual features implicitly encode rich patch-level semantics that are potentially valuable for pixel-wise prediction (Radford et al. 2021). Motivated by this observation, we design a Learnable Pixel-Text Alignment Module (LPAM) to explicitly extract fine-grained alignment relations between image regions and textual concepts.

Specifically, given an image I and a set of classes $\mathcal{C}$, we extract dense image embeddings $D_V = \Phi_V(I) \in \mathbb{R}^{(H\times W)\times d}$ and text embeddings $D_L = \Phi_L(T) \in \mathbb{R}^{N_C\times d}$, where $\Phi_V(\cdot)$ and $\Phi_L(\cdot)$ are the image and text encoders of CLIP, respectively. Following prior work (Zhou, Loy, and Dai 2022b; Cho et al. 2024), we modify the last attention layer of the image encoder to remove the pooling operation. Based on this, the semantic alignment map $R \in \mathbb{R}^{(H\times W)\times N_C}$ is computed by the proposed LPAM as:

$$R = f_{LPAM}(D_V, D_L), \tag{5}$$

where $f_{LPAM} : \mathbb{R}^{(H\times W)\times d} \times \mathbb{R}^{N_C\times d} \to \mathbb{R}^{(H\times W)\times N_C}$ is a learnable function that computes fine-grained similarities between image patches and class embeddings. As shown in Fig. 2, LPAM aligns each visual token with all class embeddings via a learned attention mechanism, producing a dense alignment map for pixel-level semantic prediction.

Notably, this module is applied to both the pretrained CLIP (teacher) and the fine-tuned CLIP (student), denoted as f^T_{LPAM} and f^S_{LPAM}, respectively, with shared parameters. Consequently, their resulting pixel-text alignments are represented as R^T and R^S.

Semantic Compression via Information Bottleneck

While LPAM extracts fine-grained pixel-text alignment maps from CLIP's intermediate features, these representations originate from a global image-text alignment objective and are not explicitly optimized for dense prediction. As a result, they often contain noisy or entangled semantics that lack the clarity and determinism required for accurate pixel-level segmentation (Xie et al. 2024; Cho et al. 2024).

To address this issue, we propose an information bottleneck mechanism that compresses the alignment map to extract essential semantic-aware semantics while suppressing irrelevant or noisy signals. Given the dense image embeddings D^T_V and text embeddings D^T_L from the pretrained CLIP, and their corresponding pixel-level alignment map R^T produced by LPAM, our first goal is to regulate the information flow by minimizing the mutual information $\mathbf{I}_\alpha(D^T_V, D^T_L;\, R^T)$. Here, $\mathbf{I}_\alpha(\cdot\,;\cdot)$ denotes the α-Rényi's mutual information, which measures the amount of information shared between the input embeddings and the alignment map.

In other words, this can be formulated as a regularization loss to supervise the learning of R^T:

$$\begin{aligned}\mathcal{L}_c =& \mathbf{I}_\alpha(D^T_V, D^T_L;\, R^T)\\ =& \mathbf{S}_\alpha(G^T_V, G^T_L) + \mathbf{S}_\alpha(G^T_R) - \mathbf{S}_\alpha(G^T_V, G^T_L, G^T_R).\end{aligned} \tag{6}$$

Notably, the teacher entropy term $\mathbf{S}_\alpha(G^T_V, G^T_L)$ in this loss can be excluded, as the teacher's weights remain fixed during fine-tuning. The $G^T_V, G^T_L, G^T_R \in \mathbb{R}^{N\times N}$ are the Gram matrices induced by a batch of normalized features D^T_V, D^T_L, and R^T with a polynomial kernel of degree 1 (Zhang et al. 2025). For instance, G^T_V can be defined as follows:

$$G^T_V = \frac{\kappa(D^T_V, D^T_V)}{\mathrm{tr}(\kappa(D^T_V, D^T_V))}, \tag{7}$$

where $\kappa(x, y) = x^\top y$ represents polynomial kernel function and $\mathrm{tr}(\cdot)$ denotes the trace of the matrix. Therefore, according to the Preliminaries and Eq. 7, we can reformulate the compression loss $\mathcal{L}_c$ as follows:

$$\begin{aligned}\mathcal{L}_c =& \mathbf{S}_\alpha(G^T_R) - \mathbf{S}_\alpha(G^T_V, G^T_L, G^T_R)\\ =& \frac{1}{1-\alpha}\left[\log_2 \sum_{i=1}^{n} \lambda_i^\alpha\left(G^T_R\right) - \log_2 \sum_{i=1}^{n} \lambda_i^\alpha\left(G^T_{VLR}\right)\right],\end{aligned} \tag{8}$$

where $G^T_{VLR} = G^T_V \circ G^T_L \circ G^T_R$ is derived from the marginal and joint entropy definitions in the Preliminaries, with $\circ$ denoting the Hadamard product.

As computing eigenvalues is expensive (Yu et al. 2019), we follow prior work (Miles, Rodriguez, and Mikolajczyk 2021; Zhang et al. 2025) and set $\alpha = 2$, enabling a Frobenius-norm-based approximation of Rényi's α-entropy: $\|\mathbf{A}\|_F^2 = \mathrm{tr}(\mathbf{A}\mathbf{A}^H) = \sum_{i=1}^{n} \lambda_i^2(\mathbf{A})$. Consequently, $\mathcal{L}_c$ can be reformulated as:

$$\mathcal{L}_c = -\log_2 \|G^T_R\|_F^2 + \log_2 \|G^T_{VLR}\|_F^2. \tag{9}$$

The first term suppresses redundant signals by minimizing the entropy of alignment features, while the second term encourages the alignment map to retain informative joint semantics by maximizing the joint entropy among image, text, and alignment features. Together, they form an information bottleneck that filters noise while preserving key semantics.

Alignment Transfer via Mutual Information

To improve CLIP's semantic perception in open-vocabulary segmentation, we distill compact semantic-aware pixel-text alignment from the pretrained CLIP into its fine-tuned CLIP. Unlike conventional distillation losses that rely on distribution matching or soft targets, this formulation avoids expensive density estimation and provides a stable, differentiable objective. Specifically, we achieve this distillation by maximizing the mutual information between the extracted semantic alignment representations of the pretrained and fine-tuned models. The distillation loss can be defined as:

$$\begin{aligned}\mathcal{L}_d =& -\mathbf{I}_\alpha(R^T; R^S)\\ =& -\mathbf{S}_\alpha(G^T_r) - \mathbf{S}_\alpha(G^S_R) + \mathbf{S}_\alpha(G^T_R, G^S_R)\end{aligned} \tag{10}$$

where G^S_R denotes the Gram matrix of R^S. Following the same derivation as before, the alignment transfer via mutual information can be expressed as:

$$\mathcal{L}_d = \log_2 \|G^T_R\|_F^2 + \log_2 \|G^S_R\|_F^2 - \log_2 \|G^{TS}_R\|_F^2 \tag{11}$$

From an optimization perspective, the terms in $\mathcal{L}_d$ serve distinct roles: the first two terms act as regularizers that encourage each model to maintain informative and structured

Model	VLM	Add. Backbone	Training Dataset	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
Without Distillation									
SPNet (2019)	-	ResNet-101	PASCAL VOC	-	-	-	24.3	18.3	-
ZS3Net (2019)	-	ResNet-101	PASCAL VOC	-	-	-	19.4	38.3	-
LSeg (2022a)	CLIP ViT-B/32	ResNet-101	PASCAL VOC-15	-	-	-	-	47.4	-
LSeg+ (2022)	ALIGN	ResNet-101	COCO-Stuff	2.5	5.2	13.0	36.0	-	59.0
ZegFormer (2022)	CLIP ViT-B/16	ResNet-101	COCO-Stuff-156	4.9	9.1	16.9	42.8	86.2	62.7
ZegFormer (2022)	CLIP ViT-B/16	ResNet-101	COCO-Stuff	5.6	10.4	18.0	45.5	89.5	65.5
ZSseg (2022)	CLIP ViT-B/16	ResNet-101	COCO-Stuff	7.0	-	20.5	47.7	88.4	-
OpenSeg (2022)	ALIGN	ResNet-101	COCO Panoptic	4.4	7.9	17.5	40.1	-	63.8
OVSeg (2023)	CLIP ViT-B/16	ResNet-101c	COCO-Stuff	7.1	11.0	24.8	53.3	92.6	-
ZegCLIP (2023)	CLIP ViT-B/16	-	COCO-Stuff-156	-	-	-	41.2	93.6	-
SAN (2023b)	CLIP ViT-B/16	-	COCO-Stuff	10.1	12.6	27.5	53.8	94.0	-
SED (2024)	CLIP ConvNeXt-B	-	COCO-Stuff	11.4	18.6	31.6	57.3	94.4	-
CAT-Seg (2024)	CLIP ViT-B/16	-	COCO-Stuff	<u>12.0</u>	<u>19.0</u>	<u>31.8</u>	<u>57.5</u>	<u>94.6</u>	<u>77.3</u>
With Distillation									
MAFT (2023)	CLIP ViT-B/16	ResNet-101	COCO-Stuff	10.1	12.8	29.1	53.5	90.0	-
InfoCLIP (ours)	CLIP ViT-B/16	-	COCO-Stuff	**12.6**	**19.5**	**32.1**	**58.1**	**95.5**	**78.1**
Without Distillation									
LSeg (2022a)	CLIP ViT-B/32	ViT-L/16	PASCAL VOC-15	-	-	-	-	52.3	-
OpenSeg (2022)	ALIGN	Eff-B7	COCO Panoptic	8.1	11.5	26.4	44.8	-	70.2
OVSeg (2023)	CLIP ViT-L/14	Swin-B	COCO-Stuff	9.0	12.4	29.6	55.7	94.5	-
SAN (2023b)	CLIP ViT-L/14	-	COCO-Stuff	12.4	15.7	32.1	57.7	94.6	-
ODISE (2023a)	CLIP ViT-L/14	Stable Diffusion	COCO-Stuff	11.1	14.5	29.9	57.3	-	-
SED (2024)	CLIP ConvNeXt-L	-	COCO-Stuff	13.9	22.6	35.2	60.6	96.1	-
FC-CLIP (2023)	CLIP ConvNeXt-L	-	COCO Panoptic	14.8	18.2	34.1	58.4	95.4	-
CAT-Seg (2024)	CLIP ViT-L/14	-	COCO-Stuff	<u>16.0</u>	<u>23.8</u>	<u>37.9</u>	<u>63.3</u>	<u>97.0</u>	<u>82.5</u>
With Distillation									
MAFT (2024)	CLIP ViT-L/14	Mask2Former	COCO-Stuff	12.7	16.2	33.0	59.0	92.1	-
MAFT (2024)	CLIP ConvNeXt-L	Mask2Former	COCO-Stuff	13.1	17.0	34.4	57.5	93.0	-
MAFT+ (2024)	CLIP ConvNeXt-L	Mask2Former	COCO-Stuff	15.1	21.6	36.1	59.4	96.5	-
InfoCLIP (ours)	CLIP ViT-L/14	-	COCO-Stuff	**16.6**	**24.6**	**38.5**	**63.5**	**97.5**	**83.1**

Table 1: **Comparison of open-vocabulary semantic segmentation performance on standard benchmarks.** Bold indicates the best performance, and underlining denotes the second-best.

alignment representations, while the third term enforces consistency between the teacher and student via relation-level alignment. This distillation loss $\mathcal{L}_d$ effectively preserves fine-grained, semantically consistent alignment during open-vocabulary segmentation fine-tuning.

Overall Training Process

Following prior work (Cho et al. 2024), we freeze some CLIP layers to reduce computational overhead. For the loss function, we incorporate the task loss (i.e., cross-entropy loss), denoted as $\mathcal{L}_{task}$, for the fine-tuning of CLIP. The overall loss function is derived as:

$$\mathcal{L} = \mathcal{L}_{task} + \lambda_1 \mathcal{L}_c + \lambda_2 \mathcal{L}_d \tag{12}$$

where λ_1 and λ_2 control the influence of the regularization terms to achieve an effective trade-off between task performance and knowledge preservation. The overall loss $\mathcal{L}$ combines three terms: the task loss $\mathcal{L}_{task}$ for segmentation, the compression loss $\mathcal{L}_c$ to preserve key object semantics alignment, and the distillation loss $\mathcal{L}_d$ to reinforce the effective semantic alignment learned by the pretrained CLIP into the fine-tuned model, helping to maintain semantic consistency and alleviate overfitting to textual inputs during fine-tuning.

Experiments

Experimental Setup

Datasets. Following prior work (Cho et al. 2024; Jiao et al. 2024), we train our model on COCO-Stuff (Caesar, Uijlings, and Ferrari 2018), which comprises 118k densely annotated training images across 171 categories. For evaluating open-vocabulary semantic segmentation performance, we test our model on ADE20K (Zhou et al. 2019), PASCAL VOC (Everingham et al. 2010), and PASCAL-Context (Mottaghi et al. 2014). Specifically, ADE20K contains 20k training and 2k validation images annotated with two category sets: A-150, consisting of 150 frequent classes, and A-847, covering 847 total classes. PASCAL-Context provides 5k training and validation images with dense annotations for 459 categories (PC-459), with a commonly used subset of the 59 most frequent classes (PC-59). PASCAL VOC comprises 1.5k train-

Method	$\mathcal{L}_c$	$\mathcal{L}_d$	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
w/o distillation	-	-	12.0	19.0	31.8	57.5	94.6	77.3
Distillation Methods								
KL (2015)	-	-	5.7	9.0	26.5	51.3	93.1	72.6
MAFT (2023)	-	-	11.5	17.8	31.8	56.4	95.2	76.7
MAFT+ (2024)	-	-	11.1	17.0	30.5	55.7	94.5	76.7
Ours								
InfoCLIP	✗	✓	11.3	18.1	30.6	57.2	94.6	77.2
	✓	✗	11.8	18.5	31.5	57.2	95.1	77.1
	✓	✓	**12.6**	**19.5**	**32.1**	**58.1**	**95.5**	**78.1**

Table 2: **Ablation study results of the two losses in InfoCLIP and comparison with other distillation methods.** Both base and teacher models are pretrained CLIP ViT-B/16. Existing distillation methods tend to degrade model performance, whereas InfoCLIP effectively transfers beneficial alignment from the pretrained model to the fine-tuned model.

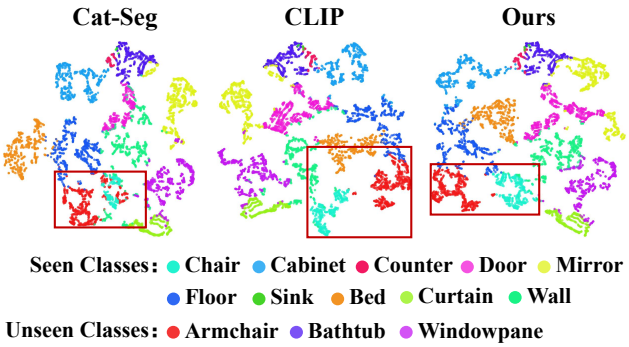


Figure 3: **Effectiveness of alignment distillation.** We present the t-SNE visualization of CLIP image embeddings. As highlighted in the red boxes, while a state-of-the-art method confuses the features of the seen class *chair* and the unseen class *armchair*, our method differentiates them and alleviates overfitting to seen classes, benefiting from the pretrained knowledge distilled from the teacher CLIP model.

ing and validation images labeled with 20 foreground object categories (PAS-20) and a background class. We also report results on PAS-20^b, where the background is redefined to include classes in PC-59 but not in PAS-20 (Cho et al. 2024).

Evaluation metric. To quantitatively evaluate performance, we follow standard practice (Cho et al. 2024; Jiao et al. 2024) and evaluate semantic segmentation results using mean Intersection over Union (mIoU).

Implementation Details. In line with previous studies (Cho et al. 2024), we evaluate our method using two Transformer-based CLIP variants, ViT-B/16 and ViT-L/14, with the same decoder as CAT-Seg. For training, the AdamW optimizer (Loshchilov and Hutter 2017) is employed with a learning rate of 2×10^{-4} for both the decoder (Cho et al. 2024) and our distillation module, and 2×10^{-6} for the CLIP backbone, along with a weight decay of 10^{-4}. We set the batch size to 4 and train the model for 80k iterations on a single NVIDIA A800 (80 GB) GPU. The key hyperparameters λ_1 and λ_2 are set to 1.

Main Results

Comparison on Standard Benchmarks. Here, we compare our proposed InfoCLIP with several state-of-the-art methods over six test sets across three benchmarks, as summarized in Table 1. Overall, InfoCLIP achieves the best performance. Among these methods, CAT-Seg (Cho et al. 2024) achieves comparable performance to ours. However, it primarily emphasizes local features and employs a complex decoder to reduce redundancy, yet overlooks the degradation of alignment induced by the shift from image-text to pixel-text correspondence. Furthermore, unlike the MAFT series (Jiao et al. 2023, 2024) that distill visual features, InfoCLIP focuses on pixel-text alignment and achieves superior performance by explicitly mitigating alignment shifts during fine-tuning. Specifically, compared to MAFT, InfoCLIP achieves significant improvements of 8.4% on the PC-459 dataset and 4.5% on the PC-59 dataset when using ViT-L/14 as the backbone. These results demonstrate the effectiveness of our method in preserving generalization while learning pixel-level alignment semantic segmentation knowledge. Furthermore, we present an efficiency analysis in the Appendix.

Qualitative Analysis. To further validate the effectiveness of InfoCLIP, we present the t-SNE (Maaten and Hinton 2008) visualization of the dense image embeddings from CLIP on the A-150 (Zhou et al. 2019) dataset. Following prior work (Cho et al. 2024), we color the embeddings according to the predicted text classes. As shown in Fig. 3, outlined in red boxes, without preserving the pretrained alignment knowledge, the state-of-the-art method tends to overfit to seen classes during fine-tuning. Specifically, as the seen class *chair* is reinforced during fine-tuning, Cat-Seg often misclassifies similar objects, such as the unseen class *armchair*, as *chair*, resulting in entangled features between the two classes. Fortunately, the pretrained CLIP model establishes a distinctive feature space through large-scale vision-language pretraining. Our proposed InfoCLIP effectively transfers this alignment knowledge to the fine-tuned model, successfully disentangling the feature spaces of *chair* and *armchair*. Furthermore, we visualize the prediction results and provide detailed illustrations in the Appendix.

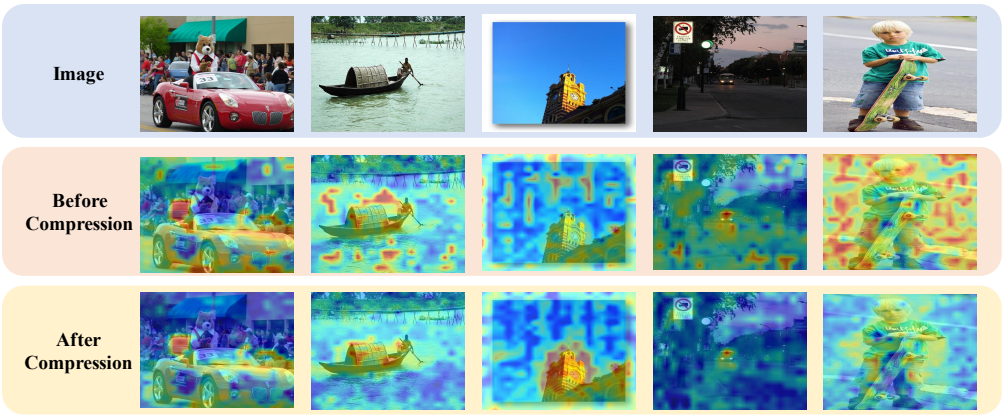
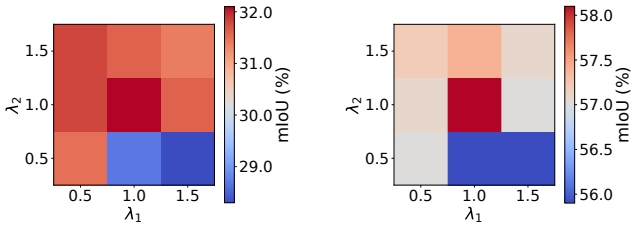


Figure 4: **Effectiveness of semantic alignment extraction and compression.** Semantic compression denoises the pixel-text alignments extracted from the pretrained model, resulting in a sharper focus on the semantic center. From left to right: examples corresponding to the classes *car*, *boat*, *building-other*, *bus*, and *person*.



(a) A-150 (b) PC-59

Figure 5: **Hyperparameter sensitivity analysis of λ_1 and λ_2 balancing $\mathcal{L}_c$ and $\mathcal{L}_d$** on the A-150 and PC-59 datasets.

Ablation Studies

Ablation of Main Components. Here, we conduct an ablation study to demonstrate the benefits of each component of our proposed InfoCLIP: semantic alignment compression loss $\mathcal{L}_c$ and alignment distillation loss $\mathcal{L}_d$, as shown in Table 2. We use the ViT-B/16 version of CLIP as the backbone. Additionally, we implement several representative distillation methods, including Kullback-Leibler (KL) divergence-based distillation and the distillation methods from the MAFT series (Jiao et al. 2023, 2024), as shown in rows 2 to 4. Notably, existing distillation-based methods even underperform compared to their non-distilled counterparts. In contrast, InfoCLIP extracts generalized semantic-level pixel-text alignment from the pretrained CLIP and suppresses noise through information compression, enabling effective knowledge transfer to the fine-tuned CLIP and leading to improved performance. To ablate the proposed loss functions, row 5 applies a mutual information-based distillation loss on cost volumes (Cho et al. 2024), while row 6 replaces the mutual information loss $\mathcal{L}_d$ with a KL divergence loss. The results show that combining both losses yields additional improvements, with a 1.4% gain on the PC-459 dataset. This highlights the necessity of extracting clean and informative signals from the pretrained CLIP to effectively suppress the impact of noise.

Discussion of Semantic Alignment Extraction and Compression. Furthermore, we visualize and evaluate the effectiveness of the Learnable Pixel-Text Alignment Module (LPAM) and the subsequent information bottleneck compression imposed on its outputs via the loss function $\mathcal{L}_c$. As shown in Fig. 4, the second row, marked with an orange background, presents the heatmap of the cost volume computed via cosine similarity (Cho et al. 2024), whereas the third row illustrates the outputs of the proposed InfoCLIP's learnable alignment module with compression. The results indicate that our approach more effectively captures prompt-guided object semantic information, yielding clearer and more accurate pixel-level alignment.

Analysis of Hyperparameters λ_1 and λ_2. In Fig. 5, we present a key hyperparameter sensitivity analysis of λ_1 and λ_2, which balance the compression loss $\mathcal{L}_c$ and the distillation loss $\mathcal{L}_d$ across multiple benchmarks. Each subfigure displays the mIoU heatmap under varying hyperparameter settings, illustrating the impact on the overall loss. Based on the heatmap, we select $\lambda_1 = 1$ and $\lambda_2 = 1$ as the best-performing configuration. Please refer to the Appendix for further results and detailed analysis.

Conclusion

We present InfoCLIP, an information-theoretic framework featuring two complementary modules for extracting and transferring semantic alignment, tailored for the asymmetric fine-tuning of CLIP. InfoCLIP extracts denoised, task-relevant pixel-text alignment from pretrained CLIP representations and transfers this refined knowledge to downstream segmentation tasks. By compressing and preserving mutual information, InfoCLIP bridges the gap between CLIP's coarse-grained pretraining objectives and the fine-grained requirements of semantic segmentation, leading to more accurate and robust alignment. Extensive experiments demonstrate that InfoCLIP consistently outperforms prior methods across multiple benchmarks, establishing a new state of the art and highlighting the promise of information-driven alignment transfer in vision-language models.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 62192781, 62172326, 62137002, and 62302384, the Key Research and Development Project in Shaanxi Province No. 2023GXLH-024, and the Project of China Knowledge Centre for Engineering Science and Technology.

References

Bhatia, R. 2006. Infinitely divisible matrices. *The American Mathematical Monthly*, 113(3): 221–235.

Bucher, M.; Vu, T.-H.; Cord, M.; and Pérez, P. 2019. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32.

Caesar, H.; Uijlings, J.; and Ferrari, V. 2018. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1209–1218.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Cho, S.; Shin, H.; Hong, S.; Arnab, A.; Seo, P. H.; and Kim, S. 2024. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4113–4123.

Ding, J.; Xue, N.; Xia, G.-S.; and Dai, D. 2022. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11583–11592.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338.

Ghiasi, G.; Gu, X.; Cui, Y.; and Lin, T.-Y. 2022. Scaling open-vocabulary image segmentation with image-level labels. In *European conference on computer vision*, 540–557. Springer.

He, J.; Deng, Z.; Zhou, L.; Wang, Y.; and Qiao, Y. 2019. Adaptive pyramid context network for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7519–7528.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Jiao, S.; Wei, Y.; Wang, Y.; Zhao, Y.; and Shi, H. 2023. Learning mask-aware clip representations for zero-shot segmentation. *Advances in Neural Information Processing Systems*, 36: 35631–35653.

Jiao, S.; Zhu, H.; Huang, J.; Zhao, Y.; Wei, Y.; and Shi, H. 2024. Collaborative vision-text representation optimizing for open-vocabulary segmentation. In *European Conference on Computer Vision*, 399–416. Springer.

Li, B.; Weinberger, K. Q.; Belongie, S.; Koltun, V.; and Ranftl, R. 2022a. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7061–7070.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov): 2579–2605.

Mansourian, A. M.; Ahmadi, R.; Ghafouri, M.; Babaei, A. M.; Golezani, E. B.; Ghamchi, Z. Y.; Ramezanian, V.; Taherian, A.; Dinashi, K.; Miri, A.; et al. 2025. A Comprehensive Survey on Knowledge Distillation. *arXiv preprint arXiv:2503.12067*.

Miles, R.; Rodriguez, A. L.; and Mikolajczyk, K. 2021. Information theoretic representation distillation. *arXiv preprint arXiv:2112.00459*.

Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.-G.; Lee, S.-W.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014. 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*, 891–898.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Xian, Y.; Choudhury, S.; He, Y.; Schiele, B.; and Akata, Z. 2019. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8256–8265.

Xie, B.; Cao, J.; Xie, J.; Khan, F. S.; and Pang, Y. 2024. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3426–3436.

Xu, J.; Liu, S.; Vahdat, A.; Byeon, W.; Wang, X.; and De Mello, S. 2023a. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2955–2966.

Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; and Bai, X. 2023b. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2945–2954.

Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; and Bai, X. 2022. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, 736–753. Springer.

Yu, Q.; He, J.; Deng, X.; Shen, X.; and Chen, L.-C. 2023. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36: 32215–32234.

Yu, S.; Giraldo, L. G. S.; Jenssen, R.; and Principe, J. C. 2019. Multivariate Extension of Matrix-Based Rényi's α-Order Entropy Functional. *IEEE transactions on pattern analysis and machine intelligence*, 42(11): 2960–2966.

Zhang, Y.; Yuan, M.; Zhang, W.; Gong, T.; Wen, W.; Ying, J.; and Shi, W. 2025. InfoSAM: Fine-Tuning the Segment Anything Model from An Information-Theoretic Perspective. *arXiv preprint arXiv:2505.21920*.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3): 302–321.

Zhou, C.; Loy, C. C.; and Dai, B. 2022a. Extract free dense labels from clip. In *European conference on computer vision*, 696–712. Springer.

Zhou, C.; Loy, C. C.; and Dai, B. 2022b. Extract free dense labels from clip. In *European conference on computer vision*, 696–712. Springer.

Zhou, T.; Wang, W.; Konukoglu, E.; and Van Gool, L. 2022. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2582–2593.

Zhou, Z.; Lei, Y.; Zhang, B.; Liu, L.; and Liu, Y. 2023. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11175–11185.