Investigating Mask-Text Contrastive Alignment in Semantic Segmentation

ECAI 2025
I. Lynce et al. (Eds.)

doi:10.3233/FAIA250887

Investigating Mask-Text Contrastive Alignment in Semantic Segmentation

Federico D'Asaro[a,b,*], **Andrea Bottino**[b] **and Giuseppe Rizzo**[a,b]

[a]LINKS Foundation – AI, Data & Space (ADS)
[b]Politecnico di Torino – Dipartimento di Automatica e Informatica (DAUIN)
ORCID (Federico D'Asaro): https://orcid.org/0009-0003-8727-3393, ORCID (Andrea Bottino): https://orcid.org/0000-0002-8894-5089, ORCID (Giuseppe Rizzo): https://orcid.org/0000-0003-0083-813X

Abstract. Vision-Language Pretrained Models (VLPs) have shown remarkable success in transferring knowledge to various downstream tasks, ranging from image-level tasks such as classification to pixel-level tasks such as Semantic Segmentation. However, a persistent challenge in the latter dense prediction tasks is the misalignment between pixel and text features. This mismatch hinders effective fusion between visual and textual representations, and leads to sub-optimal predictions. While some studies attribute this to a *Modality Gap*, where vision and language modalities form distinct clusters within the shared feature space, we argue that the key issue is semantic misalignment, where the pixel features do not accurately reflect the concepts encoded by the text features. To achieve a stronger semantic alignment between pixels and text embeddings, in this work we propose a Mask-Text Contrastive (MTC) module that explicitly enforces an alignment between image regions and their corresponding semantic concepts. This is achieved by projecting both pixel and text features into a common space where an InfoNCE-based loss promotes semantic correspondence, reducing the modality gap as a side effect. Our approach can be seamlessly integrated into state-of-the-art VLP-based segmentation architectures, requiring only a lightweight linear projection and introducing minimal computational overhead at inference time. Experiments show that the MTC module consistently improves segmentation performance in benchmarks such as ADE20K, COCO-Stuff 10k and Pascal Context. Further experiments with COCO show that MTC is also effective in other downstream dense tasks such as object detection and instance segmentation. The repository associated with this work is available at https://github.com/fedasaro62/mask-text-contrastive-fully-seg.

1 Introduction

In recent years, Vision-Language Pre-trained Models (VLPs), such as CLIP [31], have shown promising results on dense prediction tasks such as Semantic Segmentation [8]. These models are often adapted to segmentation pipelines by integrating the CLIP text encoder into an encoder-decoder architecture. After encoding, the visual and textual features are fused by a decoder to produce the final segmentation output. Previous work has shown that leveraging the semantic richness embedded in the text encoder can benefit the visual path and improve the segmentation accuracy [32, 43].

* Corresponding Author. Email: federico.dasaro@polito.it.

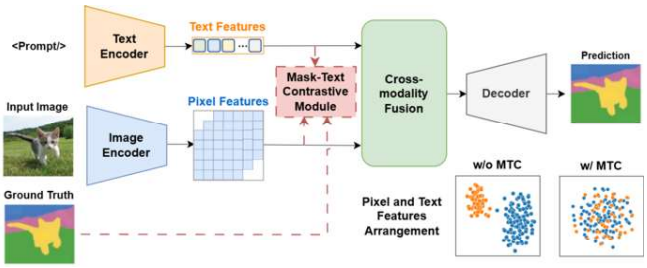


Figure 1: A generic CLIP-based encoder-decoder architecture for semantic segmentation, extended by our Mask-Text Contrastive (MTC) module (in red). Rather than addressing only the *Modality Gap*, MTC focuses on reducing the semantic misalignment between pixel and text features by aligning them in a common embedding space using a contrastive loss. This semantic alignment enables more effective fusion and leads to consistent improvements in segmentation accuracy.

When working in a multimodal setting, effective fusion between visual and textual information requires high-quality alignment [23]. In this context, *alignment* refers to the semantic correspondence between the modalities, while *fusion* refers to the integration of these aligned features. Nevertheless, existing segmentation approaches often overlook explicit mechanisms for aligning pixel-level features with their textual counterparts prior to fusion, despite its proven importance in various image-level tasks [22].

We first note that the *Modality Gap* [25] that occurs in VLPs at the image level, also extends to the pixel level (see Figure1) in most existing architectures such as DenseCLIP [32], ZegCLIP [43], and OTSeg [19]. However, we find that Modality Gap reduction alone is not sufficient to improve segmentation performance if it is not accompanied by semantic alignment.

To address this issue, we propose the *Mask-Text Contrastive (MTC)* module, which aligns pixel and text features by projecting them into a shared embedding space where a contrastive loss (InfoNCE) enforces semantic alignment (Figure 1). Inspired by [5], this alignment uses ground-truth pixel masks instead of individual pixels, resulting in a conceptually more accurate representation. Indeed, this class-wise supervision enforces concept-level consistency, ensuring that visual features accurately reflect the semantics encoded in text embeddings reducing as well the Modality Gap as a positive side effect. The main contributions of this work are the following:

- We analyze the effects of semantic misalignment between pixel and text features in VLP-based segmentation models.
- We introduce a novel *Mask-Text Contrastive* module that enhances semantic alignment between pixel and text features, leading to improved segmentation performance across multiple models and datasets. Additionally, it proves effective in object detection and instance segmentation.
- Through extensive ablation studies, we demonstrate that semantic alignment, rather than merely reducing the Modality Gap, is the primary driver of improvements in segmentation accuracy.

2 Related Work

Semantic Segmentation. Semantic Segmentation (SS) is a dense prediction task that involves assigning a semantic label to each pixel in an image [20]. Over the past two decades, SS has found widespread application in areas such as robotics [17], satellite imaging [1], and medical image analysis [30]. The field has evolved from early CNN-based models like U-Net [33], FCN [27], and DeepLab [3], to more recent Transformer-based architectures such as SegFormer [36] and SETR [41]. Hybrid models, such as SegNeXt [14], aim to leverage the strengths of both CNNs and Transformers by integrating CNN-based building blocks with Transformer-like multi-head attention mechanisms, thereby achieving a balance between computational efficiency and expressive power.

Most existing segmentation approaches rely on models pretrained on ImageNet [6], requiring fine-tuning of all parameters, which is computationally intensive. In contrast, recent works leverage Vision-Language Pretrained Models such as CLIP as backbones, significantly reducing the computational burden while maintaining competitive performance.

Vision-Language Pre-trained Models. Recently, VLPs such as CLIP [31] and ALIGN [18] have emerged as powerful general-purpose learners, showing remarkable success in zero-shot classification and retrieval tasks. These models are trained in a weakly supervised manner using large-scale web-sourced image-text pairs and a contrastive learning objective. Due to their architectural design, which employs separate branches for image and text encoding, they are commonly referred to as *Dual-Encoders*. The versatility of these models has motivated extensive efforts to transfer their capabilities to downstream vision tasks, including object detection [13], and semantic segmentation [32].

However, a known limitation of Dual-Encoder VLPs is the presence of a *Modality Gap* [25], where image and text features tend to reside in distinct subspaces of the shared embedding space. This issue, rooted in the contrastive pre-training paradigm, has been shown to hinder tasks that require tight cross-modal alignment. Reducing this gap has proven beneficial in applications such as cross-modal retrieval [9], multimodal arithmetic [10], semantic communication through compact latent representations [12], and cross-modality classification, where text encodings are directly fed into classifiers trained on images [28].

In this work, we observe that VLP-based segmentation networks display a Modality Gap at the pixel level, indicating sub-optimal alignment quality. Throughout this study, we investigate the role of this gap in the semantic alignment between pixel and text features.

Pixel-Text Alignment. VLPs have been successfully applied to dense prediction tasks such as semantic segmentation. For instance, [32] adapt CLIP from an image-text to a pixel-text matching paradigm, showcasing the effectiveness of vision-language pre-training for transferring knowledge to segmentation. Building on the zero-shot classification capabilities of VLPs, recent work has also expanded semantic segmentation to an open-set setting, giving rise to the task of Open-Vocabulary Semantic Segmentation.

Effectively adapting VLPs to dense prediction requires transitioning from image-text alignment to pixel-text alignment. Early *two-stage* approaches first generate class-agnostic mask proposals and then classify each proposal using CLIP by matching region proposals with the textual features of class labels [7, 11, 38, 37]. Notably, MaskCLIP [8] and SAN [38] supervise segmentation via dice and binary cross-entropy losses for mask generation, and cross-entropy loss for the final classification, but do not explicitly enforce pixel-text alignment. More recent *one-stage* methods bypass the need for a mask proposal module by directly aligning pixel and text features using visual feature maps extracted from the image encoder [21, 32, 43, 19]. For example, DenseCLIP [32] performs pixel-text matching through a cross-entropy loss computed over intermediate similarity score maps. PPL [21] introduces a probabilistic pixel-text matching strategy to capture diverse and informative class-level attributes. OTSeg [19] proposes a Multi-Prompt Sinkhorn algorithm to align multiple textual prompts with semantic pixel features.

However, these methods often lack sufficiently strong semantic correspondence between class-specific pixel and text features. To address this, MTA-CLIP [5] propose using ground-truth masks to explicitly guide cross-modal alignment within a one-stage framework. Inspired by their approach, we introduce a *Mask-Text Contrastive (MTC)* module that similarly leverages mask supervision but differs in design. Unlike the architecture in [5] —which relies on customized decoders and prompt learning—our method offers a lightweight, plug-and-play alignment component that can be easily integrated into any CLIP-based encoder-decoder segmentation framework. This modularity highlights the flexibility of our solution and its effectiveness in promoting feature alignment prior to fusion in the decoding stages.

3 Preliminaries on the Modality Gap

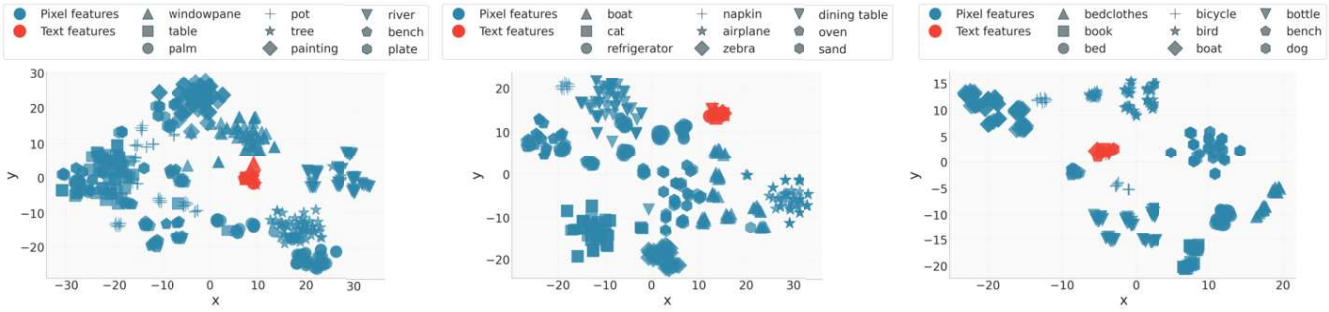
In this section, we present a preliminary discussion of our work by introducing the concept of the *Modality Gap*, detailing its measurement, and reviewing existing methods proposed for its reduction.

3.1 Introduction to the Modality Gap

Previous works have observed the phenomenon of the *Modality Gap* in CLIP-based multimodal learning methods [25, 34]. It refers to the separation of image and text features, which are clustered in different subregions of the shared feature space. Formally, let $E_I, E_T \in \mathbb{R}^{n \times d}$ denote the image and text embeddings, respectively, where n is the number of samples and d is the embedding dimension. The *Modality Gap* is defined as the difference between the centroids of the image and text embeddings:

$$\Delta_{\text{gap}} = \frac{1}{n}\sum_{i=1}^{n} E_I^{(i)} - \frac{1}{n}\sum_{i=1}^{n} E_T^{(i)}, \tag{1}$$

Authors from [10] demonstrate that this gap, common to various VLPs, arises due to random initialization and contrastive pretraining. These factors cause the features to lie on a lower-dimensional manifold in the latent space. Their experiments, along with those of [9], showed that reducing this gap can improve downstream performance in tasks such as cross-modal retrieval, both in zero-shot settings and with fine-tuning.



(a) ADE20K. (b) COCO-Stuff 10k. (c) Pascal Context.

Figure 2: T-SNE visualization of features extracted from DenseCLIP ViT-B, trained on ADE20K, COCO-Stuff 10k and Pascal Context, showing pixel and text representations. Ten categories are randomly sampled for display.

3.2 *Reducing the Modality Gap*

The Modality Gap phenomenon is a recent discovery, and there is still no universally accepted method to mitigate it. In the following, we examine some of the techniques proposed over the past few years, whose impact we aim to understand in the context of the semantic segmentation task.

Optimizing Alignment and Uniformity. Recent studies have explored the adaptation of alignment and uniformity principles to multimodal contrastive learning to minimize the gap [10]. *Uniformity* refers to the property of ensuring that features within the same modality (image or text) are evenly distributed in the feature space. The idea is that uniformity within each modality can help reduce the modality gap by avoiding the two-regions effect.

Given the image and text embeddings $E_I, E_T \in \mathbb{R}^{n\times d}$, the uniformity loss over the image space is defined as:

$$L^I_{\text{Uniform}} = \log\left(\frac{1}{N}\sum_{j,k=1}^{N}\exp\left(-2\|E^j_I - E^k_I\|^2\right)\right). \tag{2}$$

Analogously, L^T_{Uniform} is defined for the text space. The overall uniformity loss is then:

$$L_{\text{Uniform}} = \frac{1}{2}\left(L^T_{\text{Uniform}} + L^I_{\text{Uniform}}\right). \tag{3}$$

The *Alignment* loss, which instead promotes semantic alignment between similar image-text pairs, is defined as:

$$L_{\text{Align}} = \frac{1}{N}\sum_{j=1}^{N}\|E^j_I - E^j_T\|^2. \tag{4}$$

The Uniformity and Alignment Loss is defined as:

$$\mathcal{L}_{\text{UA}} = \mathcal{L}_{\text{Uniform}} + \mathcal{L}_{\text{Align}}, \tag{5}$$

Swap Modalities. Authors from [39] show that swapping modalities during contrastive training helps in mitigating the Modality Gap. Specifically, they distinguish between *Hard Swapping* and *Soft Swapping*. Hard Swapping performs a random swap of $E_I[i,j]$ and $E_T[i,j]$ for each (i,j) independently with a probability p_{hs}, resulting in the swapped features E_I and E_T. Soft Swapping (SS) applies a probabilistic interpolation between the two modalities. For each (i,j), we randomly sample $\lambda_{ij} \in [0,1]$ and compute the new features as $E_I[i,j] = \lambda_{ij}E_I[i,j] + (1-\lambda_{ij})E_T[i,j]$ and $E_T[i,j] = \lambda_{ij}E_T[i,j] + (1-\lambda_{ij})E_I[i,j]$. Similar to Hard Swapping, the swapped features are used as inputs to the loss function for network optimization. Additionally, swapping is applied to a randomly selected portion p_{ss} of the training data.

Minimizing Central Moment Discrepancy. In multimodal sentiment analysis, [15] propose reducing the discrepancy between vision and text features using the Central Moment Discrepancy (CMD) [40] as a loss function. This metric quantifies distributional differences by comparing their respective moments. The CMD loss effectively promotes modality-invariant features, which, in other words, helps reduce the Modality Gap.

4 Pixel-Text Alignment in Semantic Segmentation

Despite recent adaptations of VLPs to dense prediction tasks like semantic segmentation, pixel-text alignment prior to fusion remains underexplored. To address this, we analyze feature space structure in terms of the Modality Gap and Semantic Alignment across similar and dissimilar concepts.

Modality Gap. Differently from classification or retrieval tasks, where the Modality Gap is assessed between image-level features and textual features, in semantic segmentation, the Modality Gap must be analyzed locally, between pixel-level features and textual ones. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$ and a set of K labels $C = \{c_1, \ldots, c_K\}$, each candidate label c_i can be either hard-coded or soft-coded. Depending on the specific model, labels follow either template-based or prompt-learning strategies, resulting in prompts $P = \{p_1, \ldots, p_K\}$. The image I and prompts P are encoded by the CLIP image encoder Φ_I and text encoder Φ_T, respectively:

$$F_I = \Phi_I(I) \in \mathbb{R}^{H'\times W'\times d}, \tag{6}$$

$$F_T = \Phi_T(P) \in \mathbb{R}^{K\times d}. \tag{7}$$

where H' and W' denote the spatial dimensions of the feature map produced by Φ_I, and d is the feature dimensionality.

This formulation allows us to study pixel-level features of F_I against text features F_T. Specifically, we begin by assessing the Modality Gap on DenseCLIP [32], ZegCLIP [43], and OTSeg [19], each trained on ADE20K [42], COCO-Stuff 10k [2] and Pascal Context [29].

Table 1 illustrates the gap between visual and textual features in terms of centroid distance. We observe that the modality gap (Eq. 1) in the CLIP model propagates to the pixel level, resulting in separated

Table 1: Semantic alignment and Modality Gap, measured by Δ_{PN} and Δ_{gap} respectively, across recent CLIP-based semantic segmentation models trained on ADE20K, COCO-Stuff 10k, and Pascal Context.

Model	ADE20K		COCO-Stuff 10k		Pascal Context	
	$\Delta_{\text{PN}} \uparrow$	$\Delta_{\text{gap}} \downarrow$	$\Delta_{\text{PN}} \uparrow$	$\Delta_{\text{gap}} \downarrow$	$\Delta_{\text{PN}} \uparrow$	$\Delta_{\text{gap}} \downarrow$
DenseCLIP ViT-B	0.43	0.88	0.42	0.89	0.52	0.89
DenseCLIP Res-50	0.27	0.97	0.31	1.08	0.40	0.94
ZegCLIP	0.01	0.97	0.01	0.90	0.01	0.99
OTSeg	0.02	0.96	0.02	0.93	0.02	0.93

features with $\Delta_{\text{gap}} \geq 0.88$ (DenseCLIP ViT-B on ADE20K), which is even higher than the image level gap previously observed (0.66 in [10]). Figure 2 further illustrates this phenomenon: although the pixel features are clustered by class, they remain distant from the text features, which are grouped in a separate region of the feature space.

Semantic Alignment. A pronounced Modality Gap suggests poor semantic alignment between modalities. Beyond measuring this gap, we also assess whether the joint pixel-text feature space exhibits meaningful semantic structure. Notably, a low Δ_{gap} does not guarantee proper alignment—semantic concepts may still be misaligned if modality distributions merely overlap in latent space.

Properly assessing semantic alignment requires establishing correspondences between object pixel features and their textual representations, ensuring availability of both positive and negative pairs. To this end, we leverage the ground truth mask $Y \in \mathbb{R}^{H \times W}$ to isolate features corresponding to each object class from the visual feature map F_I.

We first upsample F_I from $H' \times W'$ to match the resolution of Y, using bilinear interpolation to preserve spatial consistency. The upsampled features allow extraction of class-specific pixel features:

$$F_{I,c_i} = \{F_I(x,y) \mid Y(x,y) = c_i\}. \tag{8}$$

We then compute the mean feature representation per class:

$$\bar{F}_{I,c_i} = \frac{1}{|F_{I,c_i}|} \sum_{(x,y) \in Y_{c_i}} F_I(x,y), \tag{9}$$

which serves as the visual counterpart to the text embedding $F_T(c_i)$.

Given visual features $\bar{F}_{I,c_i}$ and textual features $F_T(c_i)$ for each class $c_i \in C$, we define the **Positive-Negative Similarity Gap**:

$$\Delta_{\text{PN}} = S_{\text{pos}} - S_{\text{neg}}, \quad \Delta_{\text{PN}} \in [-2, 2] \tag{10}$$

$$S_{\text{pos}} = \frac{1}{|C|} \sum_{c_i \in C} \langle \bar{F}_{I,c_i}, F_T(c_i) \rangle \tag{11}$$

$$S_{\text{neg}} = \frac{1}{|C|(|C|-1)} \sum_{c_i \in C} \sum_{c_j \neq c_i} \langle \bar{F}_{I,c_i}, F_T(c_j) \rangle \tag{12}$$

Here, S_{pos} and S_{neg} represent the average cosine similarity of positive and negative pairs, respectively, both ranging in $[-1, 1]$. This formulation avoids degenerate cases where all features are similarly aligned. A larger Δ_{PN} indicates stronger semantic alignment—positive pairs are closer, and negative pairs are more distinct.

Table 1 reports the values of Δ_{PN}, highlighting that after training, ZegCLIP and OTSeg exhibit low semantic alignment before fusion, with scores around 0.01. These values are notably low compared to the theoretical maximum of 2, with the highest observed value being 0.52 for DenseCLIP ViT-B on Pascal Context. This suggests significant room for improving semantic alignment between pixel and text before their fusion in the later stages of the network.

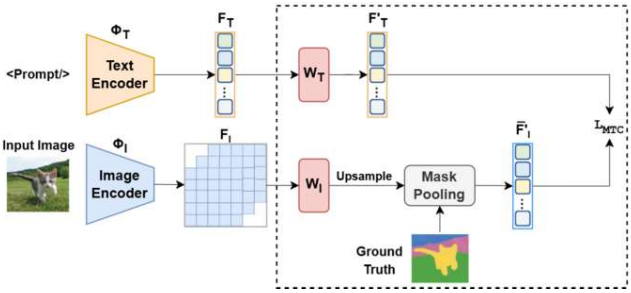


Figure 3: Our Alignment Module (in the box) is implemented within a generic CLIP-based Encoder-Decoder architecture for image semantic segmentation. Positioned immediately after the encoding stage, it enhances the alignment between pixel and text features, facilitating cross-modality fusion and improving prediction performance.

5 Mask-Text Contrastive Module

As observed, pixel and text features do not appear to be properly aligned in the shared semantic space across various CLIP-based encoder-decoder models. We hypothesize that this misalignment may hinder pixel-text fusion steps, thereby affecting segmentation performance of the model. To address this, we introduce our *Mask-Text Contrastive Module*, highlighted in the red blocks in Figure 3. The Alignment Module is positioned immediately after the image and text encoders to promote modality fusion in the subsequent stages of the network.

Specifically, during training, the model inputs consist of an image $I \in \mathbb{R}^{H \times W \times 3}$, a set of candidate labels $C = \{c_1, \ldots, c_K\}$, and the ground truth semantic map $Y \in \mathbb{R}^{H \times W}$. The image and class labels are encoded using the image encoder Φ_I (Equation 6) and text encoder Φ_T (Equation 7), resulting in image features $F_I \in \mathbb{R}^{H' \times W' \times d}$ and text features $F_T \in \mathbb{R}^{K \times d}$. The objective of our alignment module is to improve pixel-text alignment by enhancing the correspondence between visual and textual representations.

First, F_I and F_T are linearly projected using learnable transformation matrices W_I and W_T, respectively, to map them into a shared feature space:

$$F_I' = W_I F_I, \quad F_T' = W_T F_T. \tag{13}$$

Since the ground truth mask Y is available during training, we follow the same procedure used for computing Δ_{PN}: the feature map F_I' is upsampled and passed through Equations 8 and 9, resulting in the per-class visual features $\bar{F}'_{I,c_i}$. These embeddings serve as the visual counterparts to the corresponding text embeddings within a CLIP-based contrastive learning framework. We define this objective as the **Mask-Text Contrastive (MTC) loss**, which enforces semantic alignment between pixel-level features and their respective textual representations, both L2-normalized. The MTC contrastive loss function is formulated as:

$$\begin{aligned} L_{\text{MTC}} = &- \frac{1}{2K} \sum_{i=1}^{K} \log \left[\frac{\exp\left(\langle \bar{F}'_{I,c_i}, F'_T(c_i) \rangle / \tau\right)}{\sum_{j=1}^{K} \exp\left(\langle \bar{F}'_{I,c_i}, F'_T(c_j) \rangle / \tau\right)} \right] \\ &- \frac{1}{2K} \sum_{j=1}^{K} \log \left[\frac{\exp\left(\langle \bar{F}'_{I,c_j}, F'_T(c_j) \rangle / \tau\right)}{\sum_{i=1}^{K} \exp\left(\langle \bar{F}'_{I,c_i}, F'_T(c_j) \rangle / \tau\right)} \right]. \end{aligned} \tag{14}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product and τ is the contrastive temperature.

Table 2: Performance of the Alignment Module evaluated on Semantic Segmentation using DenseCLIP, ZegCLIP, and OTSeg as CLIP-based encoder-decoder models. Results are reported in terms of mIoU and mAcc, computed on the ADE20K, COCO-Stuff 10k, and Pascal Context datasets.

Model	Loss	ADE20K				COCO-Stuff 10k				Pascal Context			
		mIoU	mAcc	$\Delta_{\text{PN}} \uparrow$	$\Delta_{\text{gap}} \downarrow$	mIoU	mAcc	$\Delta_{\text{PN}} \uparrow$	$\Delta_{\text{gap}} \downarrow$	mIoU	mAcc	$\Delta_{\text{PN}} \uparrow$	$\Delta_{\text{gap}} \downarrow$
DenseCLIP ViT-B	Baseline	50.60	63.20	0.43	0.88	43.27	55.62	0.42	0.89	57.44	69.26	0.52	0.89
	MTC (ours)	**52.53**	**66.67**	**0.83**	**0.27**	**45.01**	**58.14**	**0.84**	**0.25**	**58.61**	**70.56**	**0.94**	**0.24**
DenseCLIP Res-50	Baseline	43.50	55.88	0.27	0.97	37.36	49.40	0.31	1.08	48.55	60.71	0.4	0.94
	MTC (ours)	**44.62**	**58.4**	**0.65**	**0.28**	**38.33**	**52.82**	**0.66**	**0.29**	**49.64**	**62.04**	**0.80**	**0.29**
ZegCLIP	Baseline	34.41	43.94	0.01	0.97	38.55	48.93	0.01	0.90	51.72	61.22	0.01	0.99
	MTC (ours)	**36.29**	**46.83**	**0.70**	**0.25**	**39.82**	**50.31**	**0.65**	**0.18**	**52.88**	**62.53**	**0.83**	**0.24**
OTSeg	Baseline	38.67	48.91	0.02	0.96	43.37	54.06	0.02	0.93	52.24	63.11	0.02	0.93
	MTC (ours)	**39.59**	**50.12**	**0.70**	**0.29**	**44.52**	**55.23**	**0.72**	**0.25**	**53.81**	**64.57**	**0.83**	**0.26**

This loss serves as a regularization mechanism, ensuring that image features corresponding to each class are maximally aligned with their respective text features while remaining distinct from non-matching text features.

The proposed Alignmend Module can be seamlessly integrated into any CLIP-based segmentation network following the image-text encoding stage. The total training objective is formulated as:

$$L_{\text{total}} = L_{\text{model}} + \lambda L_{\text{MTC}}. \tag{15}$$

where λ is a regularization hyperparameter that controls the contribution of the Mask-Text Contrastive loss.

6 Experiments

6.1 Datasets and Metrics

To assess the effectiveness of the proposed Mask-Text Contrastive Module in enhancing the semantic alignment between pixel and text features, we conducted extensive experiments on public semantic segmentation benchmarks.

ADE20K [42]: This dataset consists of more than 20K scene-centric images, each annotated with pixel-level object and object-part labels. It includes a total of 150 semantic categories, encompassing both "stuff" classes (e.g., sky, road, grass) and discrete "thing" classes (e.g., person, car, bed).

COCO-Stuff 10K [2]: This dataset comprises 10K images derived from the COCO training set. It is split into 9K training images and 1K validation images, covering 80 "thing" classes, 91 "stuff" classes, and one "unlabeled" class, which we exclude in this work.

Pascal Context [29]: The PASCAL Context dataset contains 10,103 images with pixel-wise annotations. It includes a subset of 59 frequent classes, divided into objects, stuff, and hybrids, commonly used for evaluation due to sparsity in other object categories.

To further evaluate the effectiveness of our Alignment Module, we conduct experiments on object detection and instance segmentation using the COCO 2017 dataset [26], which consists of 118K training images, 5K validation images, and 80 object categories.

We evaluate semantic segmentation using mean accuracy (mAcc) and mean intersection over union (mIoU), which measure per-class classification and overall mask overlap, respectively. Additionally, we assess pixel-text alignment quality via Semantic Consistency (Eq. 10) and Modality Gap (Eq. 1).

6.2 Experimental Details

Our experiments are implemented using the MMSegmentation toolbox.[1] We employ three main CLIP-based segmentation models: DenseCLIP [32], ZegCLIP [43], and OTSeg [19]. To isolate the contribution of the proposed Alignment Module within each architecture, we maintain the same hyperparameters as the baseline, including the number of iterations, batch size, learning rate, optimizer, and scheduler. Across all experiments, we set the regularization coefficient in Equation 15 to $\lambda = 0.4$. Whenever applied, the hard swapping and soft swapping probabilities p_{hs} and p_{ss} are set to 10^{-3} and 5×10^{-2}, respectively.

[1] MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. Available at https://github.com/open-mmlab/mmsegmentation

Table 3: Computational overhead introduced by MTC module compared to baselines

Model	FLOPs Increase (%)	Params Increase (M)	Params Increase (%)
DenseCLIP ViT-B	+0.015%	+0.5	+0.26%
ZegCLIP	+0.009%	+0.5	+3.40%
OTSeg	+0.005%	+0.5	+0.49%

6.3 Results

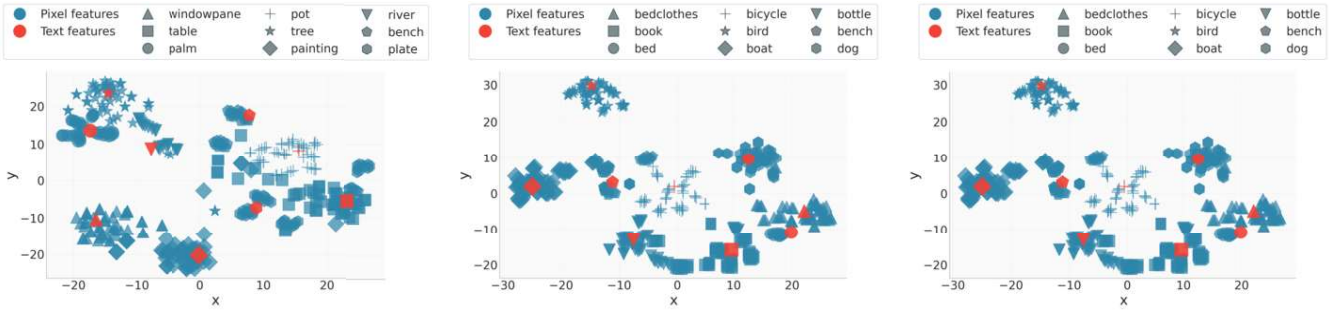
We begin by evaluating the performance of our Alignment Module against the baseline CLIP-based segmentation models DenseCLIP, ZegCLIP, and OTSeg, as reported in Table 2.

Our results show that the Mask-Text Contrastive Alignment consistently improves performance across all base models. Specifically, for DenseCLIP ViT-B, employing MTC leads to an improvement of 1.93% in mIoU on ADE20K (52.53% vs. 50.60%), and 1.74% in mIoU on COCO-Stuff 10k (45.01% vs. 43.27%). Similar improvements are observed across other segmentation models. For example, DenseCLIP Res-50 sees a 1.12% increase in mIoU on ADE20K (44.62% vs. 43.50%), and ZegCLIP experiences a 1.27% improvement on COCO-Stuff 10k (39.82% vs. 38.55%). Overall, performance gains range between 1% and 2% across all models and datasets.

The semantic alignment promoted by the MTC module can be observed through the Δ_{PN} values. For instance, on COCO-Stuff 10k, the Positive-Negative Similarity gap (Δ_{PN}) increases from 0.01 to 0.65 for ZegCLIP. Similarly, for DenseCLIP ViT-B, Δ_{PN} rises from 0.42 to 0.84. These results indicate that MTC fosters stronger semantic relationships between pixel and text, ultimately improving segmentation quality. Such improved alignment can also be observed by examining the Modality Gap Δ_{gap}, which, for COCO-Stuff 10k, decreases from 0.90 to 0.18 for ZegCLIP, and from 0.89 to 0.25 for DenseCLIP ViT-B. The diminished Modality Gap and improved Δ_{PN} are further illustrated in Figure 4, where pixel and text features appear closer in the feature space and are clustered according to their respective classes.

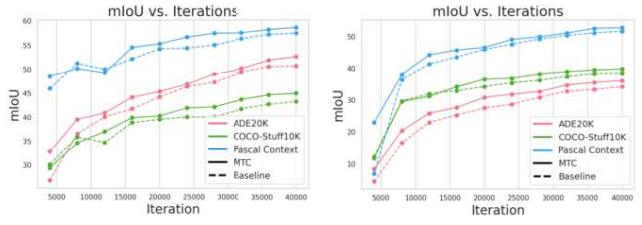
6.3.1 Training Details

Faster convergence. To emphasize the impact of the Mask-Text Contrastive Alignment during training, Figure 5 illustrates the trend



(a) ADE20K. (b) COCO-Stuff 10k. (c) Pascal Context.

Figure 4: Pixel and text feature distributions obtained from DenseCLIP trained with our Alignment Module on ADE20K (a), COCO-Stuff 10k (b), Pascal Context (c). Ten categories are randomly sampled for display.



(a) DenseCLIP ViT-B. (b) ZegCLIP.

Figure 5: Comparison of the mIoU trends over training iterations between the Alignment Module (solid line) and the baseline (dashed line) on ADE20K (red), COCO-Stuff 10k (green), and Pascal Context (blue) datasets.

Table 4: Comparison of different upsampling methods on ADE20K and COCO-Stuff 10K datasets.

Model	Upsampling Method	ADE20K	COCO-Stuff 10K
DenseCLIP ViT-B	Learnable Upsampling	52.28	44.23
	Bilinear Upsampling	**52.53**	**45.01**
ZegCLIP	Learnable Upsampling	36.13	39.59
	Bilinear Upsampling	**36.29**	**39.82**

of the mIoU metric over training iterations, comparing our MTC approach with the baseline. We observe that enforcing alignment between pixel and text features consistently leads to faster convergence, already from the very early stages of training. This behavior indicates that our method enables the model to learn more discriminative feature representations with fewer training iterations, which is particularly advantageous in scenarios requiring computational efficiency.

Computational Overhead. Table 3 shows that the MTC module introduces minimal overhead, adding only +0.5M parameters. The relative increase remains below 0.02% in FLOPs and under 3.5% in parameters, confirming its lightweight and efficient design.

6.3.2 *Ablation Study*

Effect of Upsampling Method. We explore whether the bilinear interpolation used to align visual features with the mask resolution can be replaced by a Learnable Upsampling (LU) method, inspired by [24]. This method first extracts features using a convolutional layer, then progressively upsamples them through four stages. Each stage performs bilinear upsampling with a scale factor of 2, followed by a convolutional layer with ReLU activation, except for the final output layer. The results presented in Table 4 indicate that Learnable Upsampling yields comparable or lower performance compared to simple bilinear interpolation. Therefore, we prefer the latter due to its lack of additional parameters and faster computation.

Table 5: Comparison Between Pairwise and CLIP Loss in the Mask-Text Alignment Module.

Model	Loss	ADE20K	COCO-Stuff 10k
DenseCLIP ViT-B	Pairwise	51.24	42.82
	$\mathcal{L}_{\text{MTC}}$	**52.53**	**45.01**
ZegCLIP	Pairwise	35.07	30.95
	$\mathcal{L}_{\text{MTC}}$	**36.29**	**39.82**

Table 6: Effect of Random Masking (RM) and Label Confusion (LC) on the Mask-Text Contrastive module.

Model	Loss	ADE20K	COCO-Stuff 10k
DenseCLIP ViT-B	$\mathcal{L}_{\text{MTC}}$	**52.53**	**45.01**
	$\mathcal{L}_{\text{MTC}}$+RM	52.37	44.93
	$\mathcal{L}_{\text{MTC}}$+LC	52.28	44.76
ZegCLIP	$\mathcal{L}_{\text{MTC}}$	**36.29**	**39.82**
	$\mathcal{L}_{\text{MTC}}$+RM	36.11	39.50
	$\mathcal{L}_{\text{MTC}}$+LC	35.96	39.42

Effect of Contrastive Loss Choice. To evaluate the impact of different contrastive learning strategies, we implement two versions of our Mask-Text Alignment Module: one utilizing the $\mathcal{L}_{\text{MTC}}$ loss function, as defined in Equation 14, and another employing a Pairwise contrastive loss [35], which operates by directly minimizing the distance between positive sample pairs while maximizing the distance between negative pairs. Unlike $\mathcal{L}_{\text{MTC}}$ approach, which considers multiple negatives simultaneously, the pairwise method optimizes individual pairs in isolation. Table 5 highlights the importance of contrastive loss selection in influencing performance, as the pairwise loss consistently underperforms compared to $\mathcal{L}_{\text{MTC}}$.

Effect of Mask Perturbation. Since MTC heavily relies on ground truth masks, we evaluate the impact of mask quality on model performance. Table 6 reports the results of applying Random Masking (RM), which replaces ground truth pixels with ignore labels, and Label Confusion (LC), which randomly swaps labels with incorrect ones, during training. For each training sample, a random subset of pixels (with $r \sim \mathcal{U}(0.05, 0.5)$) is selected, and the perturbation is applied with probability $p = 0.2$ per selected pixel. The results suggest that both perturbation methods have a minimal effect on performance, with scores remaining above the baseline reported in Table 2.

6.3.3 *Comparison with State-of-the-Art*

Similarly to [5], Table 7 compares our method against prior works on ADE20K and COCO-Stuff 10k using ViT-B backbone. Specifically, we use DenseCLIP ViT-B integrated with our MTC alignment module for the comparison. The results demonstrate that our method outperforms previous approaches, achieving an mIoU of 52.5 on ADE20K and 45.0 on COCO-Stuff 10K. This suggests that MTC

Table 7: Comparison of mIoU scores with state-of-the-art methods on ADE20K and COCO-Stuff 10k. Gray blocks indicate missing results due to unavailable code; (*) indicates values computed by us.

Method	Pre-train	**ADE20K**	**COCO-Stuff 10k**
Mask2Former [4]	ImageNet	51.1	40.2*
Mask2Former [4]	CLIP	51.0	40.8*
PPL [21]	CLIP	51.6	
MTA-CLIP [5]	CLIP	52.3	
Semantic FPN [20]	ImageNet-21K	49.1	39.9*
CLIP + Semantic FPN [31]	CLIP	49.4	42.4*
DenseCLIP [32]	CLIP	50.6	43.2*
DenseCLIP + MTC (ours)	CLIP	**52.5**	**45.0**

Table 8: Comparison of different approaches to reduce the Modality Gap by exploiting various losses used in the Mask-Text Contrastive Module.

Model	Loss	**ADE20K**			**COCO-Stuff 10k**		
		mIoU	$\Delta_{\text{PN}} \uparrow$	$\Delta_{\text{gap}} \downarrow$	mIoU	$\Delta_{\text{PN}} \uparrow$	$\Delta_{\text{gap}} \downarrow$
DenseCLIP ViT-B	L_{MTC}	**52.53**	0.83	0.27	**45.01**	0.84	0.25
	UA	52.26	0.76	0.21	43.68	0.79	0.15
	CMD	50.80	0.0	0.01	42.42	0.0	0.01
DenseCLIP RES-50	L_{MTC}	**44.62**	0.65	0.28	**38.33**	0.66	0.29
	UA	44.09	0.54	0.27	36.21	0.52	0.15
	CMD	42.52	0.0	0.01	35.04	0.0	0.01
ZegCLIP	L_{MTC}	**36.29**	0.70	0.25	**39.82**	0.66	0.18
	UA	35.43	0.65	0.23	39.15	0.65	0.17
	CMD	32.88	0.0	0.02	37.62	0.0	0.02
OTSeg	L_{MTC}	**39.59**	0.70	0.29	**44.52**	0.72	0.25
	UA	38.39	0.65	0.26	43.63	0.71	0.23
	CMD	37.39	0.0	0.02	43.07	0.0	0.01

is a highly effective component that can be seamlessly integrated into any CLIP-based encoder-decoder architecture, enhancing feature alignment and overall segmentation accuracy.

6.3.4 *Does a Lower Modality Gap Lead to Higher Performance?*

Comparing CLIP Loss with Alternative Gap Reduction Objectives. A key question in this study is whether reducing the Modality Gap directly improves segmentation performance, or if the gains observed in previous sections primarily stem from the semantic alignment enforced by the MTC module. To explore this, we first compare various objective functions that promote Modality Gap reduction and semantic alignment to different extents. We adopt the techniques introduced in Section 3.2—specifically, *Optimizing Uniformity and Alignment (UA)* and *Minimizing Central Moment Discrepancy (CMD)*—as alternative objectives to the L_{MTC} loss within our MTC module. The UA objective encourages a more uniform distribution of pixel and text features, effectively reducing the Modality Gap while still promoting alignment among positive pairs. In contrast, CMD focuses on minimizing the distributional discrepancy between modalities by aligning their central moments, without explicitly enforcing pairwise alignment between pixel and text features.

Table 8 presents a comparison between L_{MTC}, UA, and CMD in terms of mIoU, Δ_{PN}, and Δ_{gap}. The results indicate that the L_{MTC} loss consistently achieves the highest mIoU scores across all models. For example, on ADE20K, L_{MTC} achieves 52.53% mIoU for DenseCLIP ViT-B and 44.62% for DenseCLIP Res-50, outperforming both UA (52.26% and 44.09%, respectively) and CMD (50.80% and 42.52%). In terms of Modality Gap reduction, UA—and even more so CMD—effectively minimize Δ_{gap}. UA achieves values as low as 0.21 on ADE20K and 0.15 on COCO-Stuff 10k for DenseCLIP ViT-B, while CMD consistently maintains a near-zero value of 0.01. However, this reduction does not lead to improved segmentation performance, as the mIoU remains lower compared to that achieved using the L_{MTC} loss.

Table 9: Impact of Modality Gap reduction on L_{MTC} loss within the Mask-Text Contrastive Module using different techniques.

Model	Loss	**ADE20K**			**COCO-Stuff 10k**		
		mIoU	$\Delta_{\text{PN}} \uparrow$	$\Delta_{\text{gap}} \downarrow$	mIoU	$\Delta_{\text{PN}} \uparrow$	$\Delta_{\text{gap}} \downarrow$
DenseCLIP ViT-B	L_{MTC}	**52.53**	0.83	0.27	**45.01**	0.84	0.25
	+ HS	52.25	0.83	0.25	44.47	0.83	0.21
	+ SS	51.91	0.82	0.23	43.73	0.84	0.18
DenseCLIP RES-50	L_{MTC}	**44.62**	0.65	0.28	**38.33**	0.66	0.29
	+ HS	44.51	0.65	0.25	37.48	0.67	0.26
	+ SS	43.97	0.66	0.24	37.63	0.66	0.23
ZegCLIP	L_{MTC}	**36.29**	0.70	0.25	**39.82**	0.65	0.18
	+ HS	35.91	0.71	0.23	39.34	0.66	0.16
	+ SS	35.84	0.69	0.21	39.01	0.65	0.15
OTSeg	L_{MTC}	**39.59**	0.70	0.29	**44.52**	0.72	0.25
	+ HS	38.64	0.70	0.24	43.81	0.72	0.19
	+ SS	38.67	0.70	0.22	43.54	0.71	0.16

We hypothesize that this behavior stems from the differing degrees of semantic alignment enforced by each loss function. Specifically, while L_{MTC} maintains a relatively high Δ_{PN} (0.83 on ADE20K and 0.84 on COCO-Stuff 10k for ViT-B), UA enforces weaker alignment, leading to a lower Δ_{PN} (0.76 and 0.79, respectively). CMD, in contrast, completely disregards alignment, resulting in $\Delta_{\text{PN}} = 0.0$ across all cases. This suggests that merely reducing the Modality Gap without preserving alignment quality fails to improve segmentation, as the pixel-text features lose semantic consistency. While UA still enforces some level of alignment by constraining positive pairs, CMD does not account for alignment at all, leading to a collapse in meaningful feature associations. Thus, we believe that the impact of Modality Gap reduction should be analyzed while simultaneously ensuring robust semantic alignment between pixel and text features.

Integrating Modality Gap Reduction into CLIP Loss. Since semantic alignment between pixel and text features is crucial and the L_{MTC} loss effectively achieves it, we investigate whether *explicitly* reducing the Modality Gap within L_{MTC} offers additional benefits. To this end, we employ the *Modality Swapping* (HS/SS) technique (Section 3.2), as it preserves the L_{MTC} objective and allows evaluating the isolated effect of Modality Gap reduction.

Table 9 shows the results for each model using L_{MTC} and L_{MTC}+HS/SS losses. Analyzing the Δ_{gap} values, we find that both HS and SS effectively reduce the Modality Gap. However, this reduction does not improve segmentation performance. L_{MTC} alone achieves the best performance. We again observe the Δ_{PN} values to explain this phenomenon. These values remain comparable across both L_{MTC} loss and +HS/SS configurations, suggesting that the quality of semantic alignment remains approximately constant. In this context, the Modality Gap reduction can be assessed as an isolated effect, where explicitly enforcing a lower gap does not seem to contribute to improved performance in semantic segmentation. This may be due to the fact that the semantic alignment enforced by the L_{MTC} between pixel and text features is already sufficiently strong, such that additional reduction of the Modality Gap does not enhance the separation between positive and negative pairs, which is essential for semantic segmentation.

6.3.5 *MTC Module Performance in Object Detection and Instance Segmentation Tasks*

To further evaluate the effectiveness of our Mask-Text Contrastive Module, we conducted experiments on Object Detection and Instance Segmentation using DenseCLIP within the Mask R-CNN framework [32]. We report the standard Average Precision (AP), AP at IoU=0.5 and IoU=0.75, as well as the mean Average Precision

Table 10: Results on object detection and instance segmentation tested on COCO val2017 using Mask R-CNN [16] framework.

Model	Loss	AP^b	AP^b_{50}	AP^b_{75}	AP^b_S	AP^b_M	AP^b_L	AP^m	AP^m_{50}	AP^m_{75}	AP^m_S	AP^m_M	AP^m_L
DenseCLIP Res-50	Baseline	40.2	63.2	43.9	26.3	44.2	51.0	37.6	60.2	39.8	20.8	40.7	53.7
	MTC (ours)	**41.4**	**64.3**	**45.1**	**27.5**	**45.3**	**52.2**	**38.4**	**61.0**	**40.6**	**21.6**	**41.5**	**54.5**
DenseCLIP Res-101	Baseline	42.6	65.1	46.5	27.7	46.5	54.2	39.6	62.4	42.4	21.4	43.0	56.2
	MTC (ours)	**43.8**	**66.2**	**47.7**	**28.9**	**47.6**	**55.4**	**40.4**	**63.2**	**43.2**	**22.2**	**43.8**	**57.0**

(mAP) for both object detection and instance segmentation, since these tasks are performed concurrently.

Results in Table 10 demonstrate that integrating our MTC module into DenseCLIP consistently improves object detection and instance segmentation. For DenseCLIP with ResNet-50 and ResNet-101, overall AP increases by +1.2% (from 40.2% to 41.4% and 42.6% to 43.8%, respectively). Gains are also seen for AP at IoU=0.5 and 0.75, indicating enhanced localization and classification. Instance segmentation improves by 0.8%–1.0% mAP. Performance gains across backbones highlight that MTC reduces pixel-text misalignment, enhancing feature discrimination and benefiting various dense prediction tasks beyond segmentation.

7 Conclusion and Future Work

In this paper, we observe that existing CLIP-based semantic segmentation models often lack the proper semantic alignment between pixel and text features and suffer from a significant Modality Gap. To address this problem, we propose a *Mask-Text Contrastive Module* that uses ground truth masks to better align visual features of target objects with their textual representations. Experiments on segmentation, object detection and instance segmentation on different architectures confirm the effectiveness of our method and emphasize the importance of semantic alignment. Our approach fits seamlessly into state-of-the-art VLP-based segmentation models, requiring only a lightweight linear projection and minimal inference overhead. Finally, our analysis shows that Modality Gap reduction is subordinate to semantic alignment: gap reduction alone can degrade performance, and once alignment is fixed, further gap reduction does not provide any additional benefit.

Future work includes exploring the properties of the pixel-text feature space, such as Semantic Alignment and Modality Gap, in other dense prediction tasks like Open-Vocabulary and Few-Shot Semantic Segmentation.

Acknowledgements

This work has received funding from the European Union's Horizon Research and Innovation Programme under grant agreement No. 101132389, as part of the REEVALUATE project (https://reevaluate.eu/). We gratefully acknowledge support from Fondazione Compagnia di San Paolo through the DARE project (Developing AI for Risk management in the insurancE industry).

References

[1] E. Arnaudo, F. Cermelli, A. Tavera, C. Rossi, and B. Caputo. A contrastive distillation approach for incremental semantic segmentation in aerial images. In *International Conference on Image Analysis and Processing*, pages 742–754. Springer, 2022.

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

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

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

[5] A. Das, X. Hu, L. Jiang, and B. Schiele. Mta-clip: Language-guided semantic segmentation with mask-text alignment. In *European Conference on Computer Vision*, pages 39–56. Springer, 2024.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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

[8] Z. Ding, J. Wang, and Z. Tu. Open-vocabulary universal image segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022.

[9] S. Eslami and G. de Melo. Mitigate the gap: Investigating approaches for improving cross-modal alignment in clip. *arXiv preprint arXiv:2406.17639*, 2024.

[10] A. Fahim, A. Murphy, and A. Fyshe. Its not a modality gap: Characterizing and addressing the contrastive gap. *arXiv preprint arXiv:2405.18570*, 2024.

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

[12] E. Grassucci, G. Cicchetti, and D. Comminiello. Closing the modality gap enables novel multimodal learning applications. In *Second Workshop on Representational Alignment at ICLR 2025*.

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

[14] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, 35: 1140–1156, 2022.

[15] D. Hazarika, R. Zimmermann, and S. Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131, 2020.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[17] Z.-W. Hong, C. Yu-Ming, S.-Y. Su, T.-Y. Shann, Y.-H. Chang, H.-K. Yang, B. H.-L. Ho, C.-C. Tu, Y.-C. Chang, T.-C. Hsiao, et al. Virtual-to-real: Learning to control in visual semantic segmentation. *arXiv preprint arXiv:1802.00285*, 2018.

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

[19] K. Kim, Y. Oh, and J. C. Ye. Otseg: Multi-prompt sinkhorn attention for zero-shot semantic segmentation. In *European Conference on Computer Vision*, pages 200–217. Springer, 2024.

[20] A. Kirillov, R. Girshick, K. He, and P. Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019.

[21] H. Kwon, T. Song, S. Jeong, J. Kim, J. Jang, and K. Sohn. Probabilistic prompt learning for dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6768–6777, 2023.

[22] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[23] S. Li and H. Tang. Multimodal alignment and fusion: A survey. *arXiv preprint arXiv:2411.17040*, 2024.

[24] Y. Li, X. Chen, Z. Zhu, L. Xie, G. Huang, D. Du, and X. Wang. Attention-guided unified network for panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7026–7035, 2019.

[25] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625, 2022.

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[27] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[28] V. Maiorca, L. Moschella, A. Norelli, M. Fumero, F. Locatello, and E. Rodolà. Latent space translation via semantic alignment. *Advances in Neural Information Processing Systems*, 36:55394–55414, 2023.

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

[30] D. Nie, L. Wang, E. Adeli, C. Lao, W. Lin, and D. Shen. 3-d fully convolutional networks for multimodal isointense infant brain image segmentation. *IEEE transactions on cybernetics*, 49(3):1123–1136, 2018.

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

[32] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18082–18091, 2022.

[33] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[34] P. Shi, M. C. Welle, M. Björkman, and D. Kragic. Towards understanding the modality gap in clip. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023.

[35] Z. Wang, Y. Lu, Q. Li, X. Tao, Y. Guo, M. Gong, and T. Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022.

[36] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021.

[37] M. Xu, Z. Zhang, F. Wei, Y. Lin, Y. Cao, H. Hu, and X. 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.

[38] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. 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.

[39] C. Yaras, S. Chen, P. Wang, and Q. Qu. Explaining and mitigating the modality gap in contrastive multimodal learning. *arXiv preprint arXiv:2412.07909*, 2024.

[40] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811*, 2017.

[41] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

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

[43] Z. Zhou, Y. Lei, B. Zhang, L. Liu, and Y. Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023.