DCLIP: A Robust Dual-Modal Complementary Backbone for Weakly Supervised Semantic Segmentation

1st Hongyang Chen
School of Computer and Big Data
Heilongjiang University
Harbin, China
2242790@s.hlju.edu.cn

2nd Lei Wang*
School of Computer and Big Data
Heilongjiang University
Harbin, China
wanglei@hlju.edu.cn

3rd Hengyang Wang*
School of Computer and Big Data
Heilongjiang University)
Harbin, China
2025002@hlju.edu.cn

***Abstract*—Weakly Supervised Semantic Segmentation (WSSS) with image-level labels aims to achieve pixel-level predictions using Class Activation Maps (CAMs). In recent years, Contrastive Language-Image Pretraining (CLIP) has been introduced into WSSS to provide more accurate and robust semantic guidance for class localization. However, most existing approaches focus on image-text alignment at the CAM level. Although CLIP excels at global semantic matching, it struggles to extract sufficiently fine-grained local semantics from images. To address this issue, we propose the DCLIP framework, which improves CLIP's image-text alignment process and compensates for its limited dense semantic modeling capability by integrating DINO's stronger unsupervised spatial perception ability while optimizing CLIP's own textual attribute descriptions. Specifically, we first introduce a Bimodal Complementary Learning module (BCL) that employs frozen DINO and frozen CLIP as parallel backbones, enabling more comprehensive yet low-cost semantic feature extraction through a shared decoder. Second, to further unlock CLIP's dense knowledge and improve alignment quality, we design a Knowledge Graph-based text augmentation module (KG) that extends textual semantics with fine-grained attributes, thereby enhancing the mapping accuracy between visual and textual modalities. Extensive experiments demonstrate the effectiveness of DCLIP. While maintaining an end-to-end architecture and low training cost, DCLIP achieves 79.6% mIoU on the PASCAL VOC 2012 test set and 51.4% mIoU on the MS COCO val set, significantly outperforming existing methods.**

***Index Terms*—Weakly Supervised Semantic Segmentation, Bimodal Complementary Learning, Text Semantic Enhancement.**

I. Introduction

Weakly supervised semantic segmentation (WSSS) aims to generate pixel-level predictions using only weak annotations—such as points [1], scribbles [2], bounding boxes [3], or image-level labels—thereby significantly reducing the annotation cost of fully supervised approaches [4]. Among all low-cost annotation types, most WSSS approaches leverage image-level labels, providing dense localization cues that link visual concepts to specific pixel regions [5]. The main problem for WSSS with image-level labels is the lack of location information. Class activation mapping (CAM) methods [6] creatively enable class-discriminative localization using only image-level class labels. However, CAMs typically only highlight the most discriminative parts of objects, resulting in incomplete and poor segmentation performance. To mitigate this limitation, Contrastive Language–Image Pretraining (CLIP) [7] has been introduced into the WSSS field in recent years. CLIMS [8] leverages image–text pairs to constrain visual relationships across different semantics, CLIP-ES [9] improves gradient quality via image-text alignment to produce more reliable GradCAMs. Although these methods have achieved remarkable progress, existing approaches predominantly rely on CLIP's global image-text alignment capability, as illustrated in Fig.1(a), leaving its inherent patch-level text alignment capability largely underexplored. To unlock the dense knowledge potential of CLIP in WSSS, ExCEL [10] proposed a patch-level text alignment paradigm that obtains higher-quality CAMs with lower training cost, as shown in Fig.1(b).

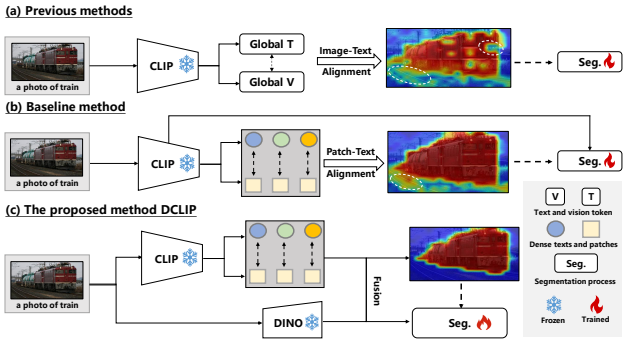

Fig. 1. Our motivation. (a) Previous methods leverage CLIP to generate CAMs with global image-text alignment, using a frozen CLIP model to produce pseudo-labels and subsequently train a separate ImageNet-pretrained segmentation model. (b) Baseline methods explore CLIP's dense knowledge through a novel patch-to-text alignment paradigm, yielding higher-quality CAMs with reduced training cost. (c) Our proposed DCLIP, which integrates frozen CLIP and frozen DINO to construct a novel single-stage approach.

However, CLIP struggles to preserve fine-grained spatial semantic information in dense prediction tasks. This limitation primarily stems from two key design aspects of its

*Corresponding author.

architecture and training objective. First, the image encoder of CLIP primarily relies on global pooled features or the [CLS] token to represent the entire image, while its text-side input typically consists only of a single class label, lacking fine-grained semantic constraints capable of guiding localized region responses. Second, CLIP's contrastive learning objective focuses predominantly on achieving image–text alignment at the global semantic level, rather than explicitly modeling pixel-level semantic consistency. In contrast, the DETR with Improved deNoising anchOr boxes (DINO) [11], as a self-supervised visual representation learning model, enforces consistency between the [CLS] token and patch tokens through a teacher–student distillation mechanism. This enables each local region to learn dense semantic representations with spatial resolution, resulting in superior performance in capturing fine-grained structures and spatial layouts compared to CLIP's global semantic alignment approach.

To this end, we propose DCLIP, a more robust weakly supervised semantic segmentation (WSSS) framework. To address the architectural limitations of existing approaches, we introduce the Bimodal Complementary Learning module (BCL) to construct a higher-quality semantic feature backbone by effectively fusing the complementary strengths of CLIP and DINO, as illustrated in Fig.1(c). We incorporate the segmentation predictions from CLIP and DINO into a bimodal high-confidence complementary learning module, achieving complementary fusion by applying a high-confidence threshold, while sharing a unified decoder between the two branches.

Furthermore, as an image–text contrastive learning model, CLIP relies on predefined text prompts for semantic alignment. However, coarse-grained prompt templates often fail to fully capture the fine-grained attributes of target objects, thereby limiting the model's ability to extract detailed semantic information from images. Therefore, to overcome the textual limitations of prior methods, we propose a novel knowledge graph-based text augmentation module (KG). Specifically, we first use an LLM to generate multi-dimensional semantic prompts per category, encode them with the CLIP text encoder to build a cross-category knowledge base, and organize the resulting vectors into a concept–attribute knowledge graph. During augmentation, the original prompt queries the graph to retrieve the Top-K non-background constituent concepts. These are similarity-weighted, aggregated, and fused with the original vector to produce a more discriminative, fine-grained text embedding.

The main contributions of our work are listed as follows:

- We design a Bimodal Complementary Learning module (BCL). BCL performs bidirectional high-confidence complementary learning on image features with a confidence threshold to obtain more detailed semantic representations.
- We design a Knowledge Graph-based text augmentation module (KG). KG constructs a concept-attribute knowledge graph using an LLM and retrieves the most semantically relevant attributes from the graph to generate enhanced text embeddings that are more discriminative and fine-grained.
- We propose a novel bimodal framework, DCLIP, that effectively addresses CLIP's limitation in extracting fine-grained semantics for weakly supervised semantic segmentation (WSSS). Extensive experiments on PASCAL VOC 2012 and MS COCO 2014 demonstrate that our method substantially outperforms state-of-the-art approaches while requiring only 4.2 GB of GPU memory, significantly reducing training cost.

II. Related Works

A. Weakly Supervised Semantic Segmentation

Typically, the WSSS pipeline consists of three stages: (1) training a model with image-level class labels to obtain CAMs for each training image, (2) refining the CAMs into pseudo-labels, and (3) training a semantic segmentation model using the pseudo-labels. Weakly supervised semantic segmentation with image-level labels typically relies on Class Activation Maps (CAMs) to provide dense supervision [12]. Several WSSS methods have employed CNNs to generate CAMs as initial seeds [13]. However, recent research suggests that CNN-based architectures are limited in capturing global information effectively due to their restricted receptive fields. Recently, WSSS methods have adopted ViT as a backbone for generating localization maps. For instance, ToCo [14] employs token-level contrastive learning to obtain more accurate local responses, SeCo [15] adopts a "separate-and-conquer" strategy to effectively alleviate co-occurring category interference.

B. Vision-Language Pre-training

Contrastive Language–Image Pretraining (CLIP), renowned for its pretraining on billions of image-text pairs, has demonstrated remarkable transferability across numerous downstream tasks. Many approaches have sought to adapt it to new scenarios: CLIP-Adapter [16] introduces lightweight trainable modules to achieve efficient few-shot classification, DenseCLIP [17] leverage image-text alignment for dense prediction tasks. Recently, CLIP has also been progressively incorporated into WSSS. CLIMS treats image-text pairs as regularization to constrain visual semantics, CLIP-ES generates class-specific gradients via image-text alignment for GradCAM inference, WeCLIP further streamlines the pipeline by directly employing the CLIP vision encoder for segmentation. Although these methods have brought substantial performance gains, most of them remain confined to global image-level text alignment and fail to fully exploit CLIP's latent semantic capability at the patch level.

III. Methodology

A. Overview

Our DCLIP framework generates Class Activation Maps (CAMs) in a low-cost and highly efficient manner. The overall training pipeline is illustrated in Fig.2 and can be summarized as follows: (1) Text Semantic Enhancement via the KG Module. We employ the Knowledge Graph-based

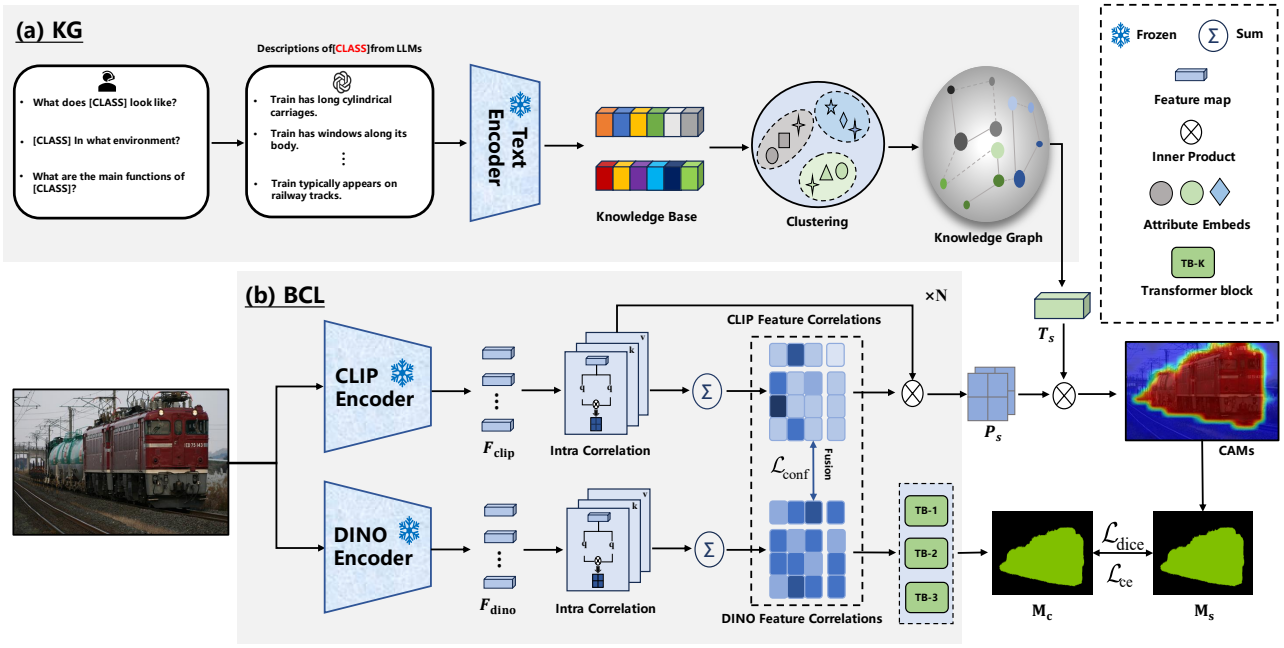

Fig. 2. DCLIP Architecture. (a) Knowledge Graph Text Augmentation Module (KG): Utilizes an LLM to construct a knowledge base, clusters it, and builds a knowledge graph. The final text representation T_s is augmented by identifying relevant attributes and suppressing non-target concepts. (b) Bimodal Complementary Learning Module (BCL): Freeze DINO and CLIP as backbone networks. Input images into the Frozen CLIP image encoder and Frozen DINO to generate two separate image feature maps. Perform complementary fusion learning through mutual supervision using CLIP and DINO segmentation results, enabling high-confidence prediction. Subsequently, decode using a unified decoder to generate semantic segmentation prediction P_s.

text augmentation module (KG) to enrich category semantics. Specifically, we first utilize GPT to generate multi-dimensional category descriptions, which are then encoded by the frozen CLIP text encoder to construct a global knowledge bank. Unsupervised clustering [18] is subsequently applied to extract cross-category shared attributes, forming a concept-attribute knowledge graph. Using global text prompts as queries, we retrieve the most semantically relevant attribute nodes for the target category from the graph, thereby generating the final enhanced text representation. (2) Bimodal Complementary Learning via the BCL Module. The input image is fed into the frozen CLIP and DINO encoders to extract global semantic features and fine-grained structural features, respectively, yielding two complementary image feature maps. We then perform bidirectional high-confidence complementary learning on the CLIP and DINO features, effectively enhancing semantic representation capability under weakly supervised conditions and further improving overall segmentation performance. (3) Feature Fusion and Decoding. Finally, the dual-branch image features are fused and passed through a unified decoder.

B. Text Semantic Enhancement

Knowledge Base Construction. To overcome the limitation of traditional global text templates that can only describe object presence and fail to provide dense semantic knowledge, we first employ a large language model (LLM) to generate multi-view and fine-grained category-specific text descriptions. As shown in Fig.1(a), the initial Class Activation Maps (CAMs) produced by CLIP based on the patch-to-text alignment mechanism tend to activate background regions that are semantically similar to the target category in complex scenes. Such semantic over-generalization introduces considerable noise into pseudo-labels, thereby degrading the reliability of downstream segmentation models. To mitigate this issue, we design a Knowledge Graph-based text augmentation module (KG) that automatically mines cross-category shared constitutive semantic attributes while explicitly suppressing background-related concepts, ultimately yielding more discriminative and category-specific text embeddings.

Specifically, for each category $s \in \mathcal{C}$, We first design prompt templates and leverage the LLM to generate fine-grained, multi-perspective descriptions, forming a category-specific description set: $\mathcal{D}_s = \{d_s^{(i)}\}_{i=1}^{M}$, where $d_s^{(i)}$ denotes the i-th LLM-generated description for category s, with a total of M descriptions. Subsequently, we employ the frozen CLIP text encoder $E_{\text{text}}(\cdot)$ to map these descriptions into a high-dimensional semantic space, obtaining the category text vector set: $\mathcal{T}_s = \{\mathbf{t}_s^{(i)}\}_{i=1}^{M}$, $\mathbf{t}_s^{(i)} = E_{\text{text}}(d_s^{(i)})$, where $\mathbf{t}_s^{(i)} \in \mathbb{R}^D$ represents the semantic embedding of description $d_s^{(i)}$, and D is the feature dimension of the CLIP text encoder. To unify textual knowledge across categories and facilitate subsequent semantic relationship analysis, we further construct a dataset-level global knowledge vector bank: $\mathcal{T} = \bigcup_{s\in\mathcal{C}} \mathcal{T}_s$.

Knowledge Graph Construction. Building upon the previously constructed dataset-level global knowledge vector bank, we further introduce an attribute-level knowledge graph.

Since the category descriptions generated by large language models inevitably contain redundant and noisy semantics, and significant shared constitutive attributes exist among different

categories, we perform unsupervised clustering on the global knowledge vector bank $\mathcal{T}$ to automatically discover cross-category shared semantic attribute prototypes. Let the clustering operation be denoted as $\mathbf{\Phi}(\cdot)$. The resulting attribute-level concept set is then obtained as: $\mathbf{Z} = \{\mathbf{z}_k\}_{k=1}^{|\mathbf{Z}|}, \quad \mathbf{z}_k = \Phi(\mathbf{T})$, where each $\mathbf{z}_k \in \mathbb{R}^D$ represents an aggregated attribute concept embedding. Based on this, we construct an attribute-level knowledge graph: $\mathcal{G} = (\mathcal{Z}, \mathcal{E})$, where $\mathcal{Z}$ denotes the set of attribute nodes, and $\mathcal{E}$ represents the latent semantic associations between nodes.

During inference, for a target category c, we first generate the base prompt embedding $\mathbf{q}_c$ (e.g., "a photo of a [category]"). We then retrieve the top-K most semantically relevant non-background concept nodes $\mathcal{Z}_c^{\text{top}} \subseteq \mathcal{Z}$ with respect to $\mathbf{q}_c$. Attention weights α_k are computed as:

$$\alpha_k = \frac{\exp\big(\text{sim}(\mathbf{q}_c, \mathbf{z}_k)\big)}{\sum_{j \in \mathcal{Z}_c^{\text{top}}} \exp\big(\text{sim}(\mathbf{q}_c, \mathbf{z}_j)\big)}, \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. These weights are used to aggregate the concept embeddings into a concept-enhanced representation $\mathbf{g}_c = \sum_k \alpha_k \mathbf{z}_k$. Finally, we fuse it with the original prompt:

$$\mathbf{T}_s = \lambda \mathbf{q}_c + (1 - \lambda)\mathbf{g}_c, \tag{2}$$

where $\lambda \in [0, 1]$ is a balancing hyperparameter. Through this mechanism, the augmented text embedding not only captures shared constitutive semantics across categories but also effectively eliminates background noise, providing more robust and discriminative text representations for weakly supervised semantic segmentation.

C. Bimodal Complementary Learning

Spatial Semantic Capture. In weakly supervised semantic segmentation, CLIP's visual features are biased toward global semantics, and its self-attention mechanism tends to produce over-smoothed class activation maps [9], thereby neglecting local structures and boundary details. To improve object localization accuracy, as illustrated in Fig.2(b), we simultaneously exploit frozen CLIP and DINO image encoders to extract complementary features and achieve fine-grained spatial semantic modeling through a multi-stage process. Given an input image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$, we first resize it to match the native input resolutions of the two encoders: $\mathbf{I}_{\text{clip}} \in \mathbb{R}^{3\times h_1\times w_1}$, $\mathbf{I}_{\text{dino}} \in \mathbb{R}^{3\times h_2\times w_2}$. The final-layer features are then extracted as: $\mathbf{F}_{\text{clip}}^0 = \mathbf{E}_{\text{clip}}(\mathbf{I}_{\text{clip}})$, $\mathbf{F}_{\text{dino}}^0 = \mathbf{E}_{\text{dino}}(\mathbf{I}_{\text{dino}})$, where $\mathbf{F}_{\text{clip}}^0 \in \mathbb{R}^{C\times h_1 w_1}$, $\mathbf{F}_{\text{dino}}^0 \in \mathbb{R}^{C'\times h_2 w_2}$. The conventional query-key self-attention mechanism suffers from progressive feature over-smoothing across stacked layers. To mitigate this issue, we adopt a non-parametric intra-correlation approach to compute the attention map for each spatial location: $\mathbf{S}^l = \sum_i \alpha_i \,\text{Corr}(\mathbf{O}_i^l)$, $\mathbf{O}_i^l \in \{\mathbf{q}^l, \mathbf{k}^l, \mathbf{v}^l\}$, where α_i are learnable contribution weights for different correlation maps, and $\text{Corr}(\cdot)$ represents a self-correlation operation [19] within the same spatial position. Finally, to align channel dimensions and spatial resolutions, each branch is processed independently by an MLP followed by a convolutional module: $\mathbf{F}_{\text{n}} = \text{Conv}(\text{MLP}(\mathbf{F}_{\text{n}}^0))$, where down-sampling operations ensure spatial size consistency, and each MLP consists of two fully connected layers with non-linear activation for channel mapping and feature enhancement.

High-Confidence Complementary Learning. Although CLIP provides rich global semantic information, its perception of local boundaries and fine-grained regions remains limited, whereas DINO excels at capturing structured spatial features. To fully exploit their complementary strengths, we introduce a high-confidence complementary learning mechanism that achieves mutual supervision between CLIP and DINO segmentation predictions through reliable high-confidence regions, thereby improving pseudo-label accuracy and final segmentation precision.

The processed features are first passed through a shared decoder to produce preliminary segmentation predictions: $\mathbf{P}_{\text{n}} = \text{Decoder}(\mathbf{F}_{\text{n}})$, where the decoder is parameter-shared across both branches, mapping features from different sources into a unified spatial representation while preserving multi-source synergy. To leverage the complementary nature of CLIP and DINO, we propose a high-confidence region mutual supervision strategy. A confidence threshold τ is set, and high-confidence pixels are selected from each prediction: $\Omega_{\text{n}} = \{\mathbf{p} \mid \mathbf{P}_{\text{n}}(\mathbf{p}) > \tau\}$. On this basis, we design the high-confidence complementary loss $\mathcal{L}_{\text{conf}}$ to enable mutual supervision between the two predictions within the high-confidence region Ω. Finally, the predictions from CLIP and DINO are fused via simple averaging to produce the ultimate semantic segmentation output: $\mathbf{P} = \frac{\mathbf{P}_{\text{clip}} + \mathbf{P}_{\text{dino}}}{2}$.

D. Training Objectives

During training in this work, since both CLIP and DINO encoders are kept frozen during training, the optimization iss driven by leveraging their complementary predictions and pseudo-label supervision within a unified decoding framework. To this end, the overall optimization objective consists of three components: a segmentation prediction loss L_{seg}, a consistency loss L_{cons} between the CLIP and DINO outputs, and a complementary constraint loss L_{conf} based on high-confidence regions. These three terms jointly form the final training objective, formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{conf}}\mathcal{L}_{\text{conf}}, \tag{3}$$

where λ_{cons} and λ_{conf} are weighting hyperparameters that balance the contributions of the respective loss terms. The unified decoder fuses image features from CLIP and DINO and produces the final prediction map $\hat{\mathbf{Y}}$. To address the severe foreground-background imbalance commonly observed in weakly supervised semantic segmentation, we adopt a weighted combination of balanced cross-entropy and Dice loss as the base segmentation loss, which simultaneously improves per-pixel classification accuracy and structural consistency of foreground regions. This loss is defined as:

$$\mathcal{L}_{\text{seg}} = \alpha \cdot \mathcal{L}_{\text{CE-bal}} + (1 - \alpha) \cdot \mathcal{L}_{\text{Dice}}, \tag{4}$$

where $\alpha = 0.7$ is a constant weighting factor. The balanced cross-entropy is computed by separately calculating the cross-entropy for foreground and background regions and then averaging them, while the Dice loss strengthens the modeling of the overall shape of foreground objects.

Since the initial response maps provided by CLIP and DINO are complementary, their predictions for the same image should remain consistent, thereby providing stable and reliable cross-modal supervision to the decoder. To this end, we introduce a consistency loss that constrains the pseudo-label predictions from both branches to be close to each other:

$$\mathcal{L}_{\text{cons}} = \left\| \mathbf{P}^{\text{CLIP}} - \mathbf{P}^{\text{DINO}} \right\|, \tag{5}$$

where $\mathbf{P}^{\text{CLIP}}$ and $\mathbf{P}^{\text{DINO}}$ denote the response maps from the two branches, respectively. This loss effectively encourages the network to rely on regions where both branches agree and suppresses pseudo-label noise arising from their discrepancies. To further exploit the complementary strengths of the two branches in different regions, we introduce a confidence-based complementary constraint. First, high-confidence masks M^{CLIP} and M^{DINO} are generated according to activation strength, indicating the foreground regions where each branch is highly confident. On this basis, we design the high-confidence complementary loss as:

$$\begin{aligned} \mathcal{L}_{\text{conf}} = &\left\| M^{\text{CLIP}} \odot (\mathbf{P}^{\text{CLIP}} - \mathbf{P}^{\text{DINO}}) \right\| \\ &+ \left\| M^{\text{DINO}} \odot (\mathbf{P}^{\text{DINO}} - \mathbf{P}^{\text{CLIP}}) \right\| \end{aligned} \tag{6}$$

where $\odot$ denotes element-wise multiplication. By imposing constraints exclusively on the mutually high-confidence regions of both branches, this loss effectively avoids interference from unreliable pixels in the weakly supervised setting, enabling the two branches to mutually correct and converge toward consistency within their respective regions of expertise.

IV. Experiments and Results

A. Experimental Settings

Datasets and Metrics. The proposed DCLIP is evaluated on PASCAL VOC 2012 [20] and MS COCO 2014 [21]. VOC contains 21 categories (1 for background). Following prior methods [15], the augmented dataset with $10,582$, $1,449$, and $1,456$ images are used for training, validating, and testing, respectively. COCO includes 81 classes, in which $82,081$ images are used for training and $40,137$ images are for validating. Mean Intersection-Over-Union (mIoU) is used as the main evaluation metric.

Implementation Details. We employ the ViT-B [35] variant of CLIP and DINO as the frozen CLIP-DINO backbones. For the KG module, we use GPT-4 to generate $n = 20$ descriptions per category. The number of attribute embeddings B is set to 112 for PASCAL VOC and 224 for MS COCO. During training, all images are resized and cropped to 320×320 to fit the frozen CLIP-DINO backbones. We train only the adapter and decoder modules using the AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-2}. During inference, multi-scale testing with scales $\{0.75, 1.0\}$ is applied. Following previous works, DenseCRF [36] is adopted as a post-processing step to refine the predictions.

TABLE I
SEGMENTATION COMPARISONS ON VOC AND COCO. NET. IS THE BACKBONE FOR SEGMENTATION. SUP. IS THE SUPERVISION TYPE. $\mathcal{I}$: IMAGE-LEVEL LABELS. $\mathcal{SA}$: SALIENCY MAPS. $\mathcal{L}$: LANGUAGE.

Method	Sup.	Net.	VOC		COCO
			Val	Test	Val
Multi-stage WSSS methods.					
L2G [22] CVPR'2022	$\mathcal{I}+\mathcal{SA}$	RN101	72.1	71.7	44.2
RCA [23] CVPR'2023	$\mathcal{I}+\mathcal{SA}$	RN38	72.2	72.8	36.8
OCR [24] CVPR'2023	$\mathcal{I}$	RN38	72.7	72.0	42.5
BECO [25] CVPR'2023	$\mathcal{I}$	RN101	73.7	73.5	45.1
CTI [26] CVPR'2024	$\mathcal{I}$	RN101	74.1	73.2	45.4
CLIMS [8] CVPR'2022	$\mathcal{I}+\mathcal{L}$	RN101	70.4	70.0	-
CLIP-ES [9] CVPR'2023	$\mathcal{I}+\mathcal{L}$	RN101	72.2	72.8	45.4
PSDPM [27] CVPR'2024	$\mathcal{I}+\mathcal{L}$	RN101	74.1	74.9	47.2
CPAL [28] CVPR'2024	$\mathcal{I}+\mathcal{L}$	RN101	74.5	74.7	46.8
Single-stage WSSS methods.					
AFA [29] CVPR'2022	$\mathcal{I}$	MiT-B1	66.0	66.3	38.9
ToCo [14] CVPR'2023	$\mathcal{I}$	ViT-B	71.1	72.2	42.3
DuPL [30] CVPR'2024	$\mathcal{I}$	ViT-B	73.3	72.8	44.6
SeCo [15] CVPR'2024	$\mathcal{I}$	ViT-B	74.0	73.8	46.7
DIAL [31] ECCV'2024	$\mathcal{I}+\mathcal{L}$	ViT-B	74.5	74.9	44.4
WeCLIP [32] CVPR'2024	$\mathcal{I}+\mathcal{L}$	ViT-B	76.4	77.2	47.1
PCRE [33] CVPR'2025	$\mathcal{I}+\mathcal{L}$	ViT-B	75.5	75.9	47.2
ExCEL [10] CVPR'2025	$\mathcal{I}+\mathcal{L}$	ViT-B	78.4	78.5	50.3
DCLIP (Ours)	$\mathcal{I}+\mathcal{L}$	**ViT-B**	**79.6**	**79.5**	**51.4**

TABLE II
TRAINING EFFICIENCY COMPARISONS ON VOC TRAIN (CAM) AND VAL SET (SEG). ALL EXPERIMENTS ARE CONDUCTED ON RTX 3090.

Method	Type	Training Time	GPU	Seg
CLIMS [8] CVPR'2022	$\mathcal{M}$	1068 mins	18.0 G	70.4
CLIP-ES [9] CVPR'2023	$\mathcal{M}$	420 mins	12.0 G	72.2
MCTformer+ [34] TPAMI'2024	$\mathcal{M}$	1496 mins	18.0 G	74.0
ToCo [14] CVPR'2023	$\mathcal{S}$	506 mins	17.9 G	71.1
DuPL [30] CVPR'2024	$\mathcal{S}$	508 mins	14.9 G	73.3
SeCo [15] CVPR'2024	$\mathcal{S}$	407 mins	17.6 G	74.0
WeCLIP [32] CVPR'2024	$\mathcal{S}$	**270 mins**	6.2 G	76.4
DCLIP (Ours)	$\mathcal{S}$	300 mins	**4.2 G**	**79.2**

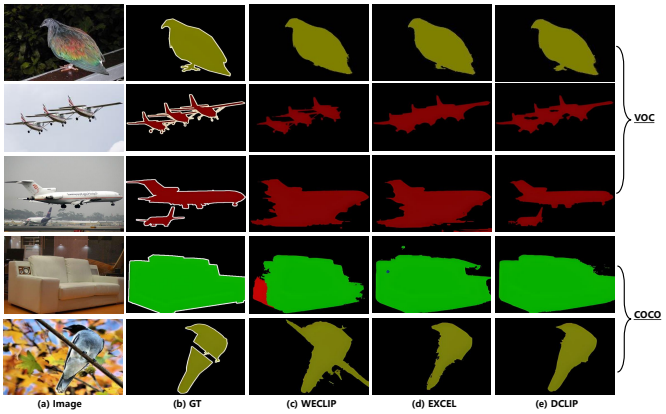

Fig. 3. Segmentation visualization of WeCLIP [32], ExCEL [10], and our method on PASCAL VOC and MS COCO. It can be observed that the online pseudo-labels generated by DCLIP are significantly more complete and accurate.

B. Comparisons with State-of-the-art Methods

Performance of Semantic Segmentation. Tab.1 presents the segmentation comparison between our DCLIP and state-of-

TABLE III
ABLATION STUDY OF DCLIP ON VOC VAL SET.

Conditions	KG	BCL	Precision	Recall	mIoU
Baseline			83.5	86.2	75.1
w/ KG	✓		85.6	88.2	76.5
w/ BCL		✓	86.3	89.8	77.7
DCLIP	✓	✓	**87.0**	**89.2**	**79.2**

the-art methods on PASCAL VOC and MS COCO. Our single-stage DCLIP achieves 79.6% mIoU on the VOC validation set and 79.5% mIoU on the test set, outperforming even complicated multi-stage approaches by at least 5.0% and 4.8% mIoU, respectively. Although ExCEL also conducts dense exploration of CLIP, its performance lags behind ours by at least 1.2%. On the more challenging COCO benchmark, DCLIP attains 51.4% mIoU on the val set, surpassing the previous CLIP-based SOTA method ExCEL by a substantial 1.1%.

Qualitative results on PASCAL VOC and MS COCO are shown in Fig.3. It can be observed that WeCLIP directly employs the CLIP vision encoder for segmentation and is therefore primarily constrained by CLIP's global image–text alignment mechanism, struggling to capture finer-grained local semantic structures. Although ExCEL explores CLIP's dense knowledge through a patch-level text alignment paradigm, its feature representation relies entirely on the frozen CLIP backbone, making it difficult to guarantee stability and accuracy in local semantic extraction. In contrast, DCLIP effectively integrates CLIP's global semantic perception with DINO's fine-grained spatial representation under the joint action of the Bimodal Complementary Learning module (BCL) and the Knowledge Graph-based text augmentation module (KG). Furthermore, structured text augmentation substantially improves vision–language alignment quality.

Training Efficiency Analysis. Tab.2 compares the training cost of our method with other state-of-the-art approaches on the PASCAL VOC 2012 dataset. As can be seen, DCLIP and WeCLIP exhibit considerably shorter training times, while the training overhead of the remaining methods generally exceeds 407 minutes. Although WeCLIP requires slightly less training time than DCLIP, its GPU memory consumption is at least 47.6% GB higher than ours. Moreover, in terms of performance, WeCLIP lags behind our method by at least 3.5% on the CAM score and 3.2% on the final segmentation score.

C. Ablation Studies

Efficacy of Key Components. Ablation studies for DCLIP are reported in Tab.3. With the ExCEL [10] under our baseline training setting, segmentation performance reaches only 75.1% mIoU. Introducing the KG module significantly enhances text–vision alignment by constructing a cross-category attribute relationship graph, filtering background interference semantics, and generating more discriminative category text representations, boosting mIoU to 76.5%. On the other hand, incorporating the BCL module alone leverages the complementarity of the frozen CLIP and DINO visual encoders and applies bidirectional consistency constraints on high-confidence regions, effectively mitigating instability caused by pseudo-label noise and further improving mIoU to 77.7%. When KG and BCL are jointly integrated into the DCLIP framework, the semantic enhancement on the text side and the complementary feature fusion on the vision side establish a mutually reinforcing optimization mechanism, pushing the final mIoU to 79.2%, with Precision and Recall reaching 87.0% and 89.2%.

TABLE IV
IMPACT OF HYPER-PARAMETERS. THE RESULTS ARE EVALUATED ON THE VOC VAL SET.

Parameters		Val
$\lambda_{\text{cons}} = 0$	$\lambda_{\text{conf}} = 0$	77.5
	$\lambda_{\text{conf}} = 0.25$	77.8
	$\lambda_{\text{conf}} = 0.5$	78.7
$\lambda_{\text{cons}} = 0.25$	$\lambda_{\text{conf}} = 0$	78.9
	$\lambda_{\text{conf}} = 0.25$	79.2
	$\lambda_{\text{conf}} = 0.5$	**79.5**
$\lambda_{\text{cons}} = 0.5$	$\lambda_{\text{conf}} = 0$	77.9
	$\lambda_{\text{conf}} = 0.25$	79.0
	$\lambda_{\text{conf}} = 0.5$	78.5

Parameter Analysis. Overall, both λ_{cons} and λ_{conf} play important roles in model performance. When $\lambda_{\text{cons}} = 0$, performance gradually improves as λ_{conf} increases, demonstrating that the high-confidence constraint effectively enhances pseudo-label quality. However, the absence of the consistency constraint still limits overall performance. Introducing an appropriate consistency constraint ($\lambda_{\text{cons}} = 0.25$) leads to optimal performance, with the highest mIoU of 79.5% achieved at $\lambda_{\text{conf}} = 0.5$. This indicates a strong synergistic effect between the two constraints. Nevertheless, further increasing λ_{cons} to 0.5 results in performance degradation, suggesting that an overly strong consistency constraint may hinder the model's ability to utilize complementary information.

V. CONCLUSION

This paper identifies CLIP's limitation in preserving fine-grained spatial semantics for dense prediction tasks and introduces DCLIP, a simple yet effective framework for weakly supervised semantic segmentation (WSSS). To address the architectural and textual shortcomings of existing approaches, we propose the Bimodal Complementary Learning module (BCL) and the Knowledge Graph-based text augmentation module (KG). Extensive experiments demonstrate that DCLIP achieves state-of-the-art performance on PASCAL VOC 2012 and MS COCO 2014, while offering new perspectives on utilizing dense representations from cross-modal pretrained models in weakly supervised settings.

VI. ACKNOWLEDGMENTS

This work was supported in part by the Guidance Project for Key Research and Development Plan of Heilongjiang Province under Grant No. GZ20230015, and in part by the Research Funds for Universities of Heilongjiang Province under grant NO. 2025-KYYWF-ZR0426.

REFERENCES

[1] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 549–565. Springer, 2016.

[2] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, pages 3159–3167, 2016.

[3] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, pages 1635–1643, 2015.

[4] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, pages 1713–1721, 2015.

[5] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *CVPR*, pages 1354–1362, 2018.

[6] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2017.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[8] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. Clims: cross language image matching for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4483–4492, 2022.

[9] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. *arXiv preprint arXiv:2212.09506*, 2022.

[10] Zhiwei Yang, Yucong Meng, Kexue Fu, Feilong Tang, Shuo Wang, and Zhijian Song. Exploring clip's dense knowledge for weakly supervised semantic segmentation, 2025.

[11] Yuanchen Wu, Xiaoqiang Li, Jide Li, Pinpin Zhu, Shaohua Zhang, et al. Dino is also a semantic guider: Exploiting class-aware affinity for weakly supervised semantic segmentation. In *ACM Multimedia 2024*.

[12] Xinqiao Zhao, Feilong Tang, Xiaoyang Wang, and Jimin Xiao. Sfc: Shared feature calibration in weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7525–7533, 2024.

[13] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2204–2213, 2019.

[14] Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2303.01267*, 2023.

[15] Zhiwei Yang, Kexue Fu, Minghong Duan, Linhao Qu, Shuo Wang, and Zhijian Song. Separate and conquer: Decoupling co-occurrence via decomposition and representation for weakly supervised semantic segmentation. In *CVPR*, pages 3606–3615, June 2024.

[16] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

[17] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen 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.

[18] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[19] Yuhang Yang, Jinhong Deng, Wen Li, and Lixin Duan. Resclip: Residual attention for training-free dense vision-language inference, 2024.

[20] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111:98–136, 2015.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[22] Peng-Tao Jiang, Yuqi Yang, Qibin Hou, and Yunchao Wei. L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16886–16896, 2022.

[23] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *CVPR*, pages 4299–4309, 2022.

[24] Zesen Cheng, Pengchong Qiao, Kehan Li, Siheng Li, Pengxu Wei, Xiangyang Ji, Li Yuan, Chang Liu, and Jie Chen. Out-of-candidate rectification for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23673–23684, 2023.

[25] Shenghai Rong, Bohai Tu, Zilei Wang, and Junjie Li. Boundary-enhanced co-training for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19574–19584, 2023.

[26] Sung-Hoon Yoon, Hoyong Kwon, Hyeonseong Kim, and Kuk-Jin Yoon. Class tokens infusion for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3595–3605, 2024.

[27] Xinqiao Zhao, Ziqian Yang, Tianhong Dai, Bingfeng Zhang, and Jimin Xiao. Psdpm: Prototype-based secondary discriminative pixels mining for weakly supervised semantic segmentation. In *CVPR*, pages 3437–3446, June 2024.

[28] Feilong Tang, Zhongxing Xu, Zhaojun Qu, Wei Feng, Xingjian Jiang, and Zongyuan Ge. Hunting attributes: Context prototype-aware learning for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3324–3334, 2024.

[29] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: end-to-end weakly-supervised semantic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16846–16855, 2022.

[30] Yuanchen Wu, Xichen Ye, Kequan Yang, Jide Li, and Xiaoqiang Li. Dupl: Dual student with trustworthy progressive learning for robust weakly supervised semantic segmentation. In *CVPR*, pages 3534–3543, June 2024.

[31] Soojin Jang, Jungmin Yun, Junehyoung Kwon, Eunju Lee, and Youngbin Kim. Dial: Dense image-text alignment for weakly supervised semantic segmentation. *arXiv preprint arXiv:2409.15801*, 2024.

[32] Bingfeng Zhang, Siyue Yu, Yunchao Wei, Yao Zhao, and Jimin Xiao. Frozen clip: A strong backbone for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3796–3806, 2024.

[33] Xiangfeng Xu, Pinyi Zhang, Wenxuan Huang, Yunhang Shen, Haosheng Chen, Jingzhong Lin, Wei Li, Gaoqi He, Jiao Xie, and Shaohui Lin. Weakly supervised semantic segmentation via progressive confidence region expansion. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9829–9838, 2025.

[34] Lian Xu, Mohammed Bennamoun, Farid Boussaid, Hamid Laga, Wanli Ouyang, and Dan Xu. Mctformer+: Multi-class token transformer for weakly supervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2024.

[35] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[36] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *NeurIPS*, 24, 2011.