



ARTICLE

Patch Collaboration with Contrastive Learning for Weakly Supervised Semantic Segmentation

Zhenhong Chen¹ and Wangyu Wu^{2,*}

¹Microsoft, Redmond, Washington, DC, USA

²School of Computer Science, University of Liverpool, Liverpool, UK

*Corresponding Author: Wangyu Wu. Email: wangyu.wu@liverpool.ac.uk

Received: 09 June 2026; Accepted: 22 September 2026; Published: 22 October 2026

ABSTRACT: Weakly Supervised Semantic Segmentation (WSSS), which aims to learn pixel-level segmentation using only image-level labels, has attracted increasing attention due to its low annotation cost and practical scalability. Existing methods mainly focus on enhancing class activation maps, improving inter-class discrimination, or introducing external cues to reduce noisy and incomplete activations. However, they often overlook the collaborative relationships among image patches, where semantically related regions may be spatially separated or locally under-activated. This limitation prevents the model from fully capturing object structures and leads to incomplete segmentation results. To address these issues, we propose Patch Collaboration with Contrastive Learning (PCCL), which unifies bidirectional contextual propagation and a newly designed spatially weighted patch contrastive objective. Unlike prior non-weighted patch contrast, PCCL embeds two-dimensional patch geometry directly into the learning signal: nearby positive patches receive stronger compactness constraints, whereas distant positive pairs retain weaker contributions. Applying this distance-aware objective to context-enhanced embeddings enables semantic similarity, spatial continuity, and directional context to be learned jointly. By jointly exploiting semantic consistency and contextual collaboration, PCCL produces more complete and reliable activation maps for weakly supervised segmentation. Experimental results on PASCAL VOC 2012 and MS COCO 2014 show competitive performance among the literature values considered in this study.

KEYWORDS: Weakly supervised; semantic segmentation; patch collaboration; contrastive learning

1 Introduction

Semantic segmentation assigns a semantic category to each image pixel and supports applications such as autonomous driving, medical image analysis, and remote sensing. Despite substantial progress, fully supervised methods rely on dense pixel-level annotations that are expensive and difficult to scale. Weakly Supervised Semantic Segmentation (WSSS) addresses this limitation by learning from less costly annotations, among which image-level labels are particularly attractive because of their availability and low annotation cost.

Most image-level WSSS methods follow a two-stage pipeline. A classification network first generates localization cues, typically Class Activation Maps (CAMs), which are then refined into pseudo masks for training a segmentation network. However, classification models tend to activate only the most discriminative object regions while overlooking less salient but semantically relevant parts. Consequently, the resulting pseudo labels are often incomplete, fragmented, or noisy.

Recent studies improve object localization through CAM refinement, saliency priors, affinity learning, external cues, or Transformer-based representations. Vision Transformers (ViTs) are especially promising because self-attention captures long-range relationships among image patches. Nevertheless, existing ViT-based WSSS methods mainly focus on class-specific localization, attention refinement, or token-level discrimination. Although global self-attention enables interactions among distant patches, it does not explicitly provide directional information propagation over the two-dimensional patch grid. Under image-level supervision, patch tokens corresponding to non-discriminative object regions may therefore remain noisy or weakly activated.

Different regions of the same object may be spatially separated, visually diverse, or inconsistently activated. Without explicit collaboration, useful semantic evidence from confident object patches may not be effectively transferred to ambiguous regions, leading to incomplete activation maps. This motivates a framework that complements the global interactions of ViT with structured horizontal and vertical propagation while simultaneously improving patch-level semantic consistency.

To address this issue, we propose **Patch Collaboration with Contrastive Learning (PCCL)**. Although PCCL shares the ViT and horizontal-vertical recurrent-refinement foundations of CPF-CTE, TKP-PCL, and APC, it introduces a different representation-learning mechanism: a distance-dependent positive-pair objective optimized on context-enhanced patch embeddings. This objective replaces the uniform treatment of prediction-mined positive pairs with patch-grid-aware weights and thereby establishes an explicit interaction among semantic agreement, spatial proximity, and bidirectional contextual propagation.

Specifically, the Patch LSTM Context Fusion module propagates information along the left-to-right, right-to-left, top-to-bottom, and bottom-to-top directions of the patch grid, allowing weakly activated regions to receive complementary evidence from confident regions. Collaborative Patch Contrastive Learning then imposes geometry-aware semantic regularization on these fused embeddings. The technical novelty lies in this coupling: contextual propagation first enriches each patch representation, and the proposed spatial weight subsequently controls the strength of positive-pair compaction according to patch-grid distance. This design goes beyond placing two existing modules side by side because spatial context directly changes how semantic pair relations are optimized.

As illustrated in [Fig. 1](#), the main contributions of this work are summarized as follows:

- We introduce a **spatially weighted positive-pair contrastive objective** for prediction-mined patches. Unlike uniform positive-pair constraints, it assigns stronger compactness constraints to nearby positive patches and weaker constraints to distant pairs, preserving local object continuity without forcing every same-class prediction to align equally.
- We develop a **semantic-spatial collaboration mechanism** in which the proposed objective operates on **contextual-fusion bidirectional long short-term memory (CF-BiLSTM)** representations. This ordered coupling allows directional context propagation and distance-aware semantic regularization to reinforce each other.
- We provide controlled component and weighting ablations together with evaluations on PASCAL VOC 2012 and MS COCO 2014. The complete collaboration framework improves lmean intersection over union (mIoU) from 65.1% to 71.1%, while replacing uniform positive-pair weights with the proposed spatial weights further improves mIoU from 70.6% to 71.1% under otherwise identical settings.

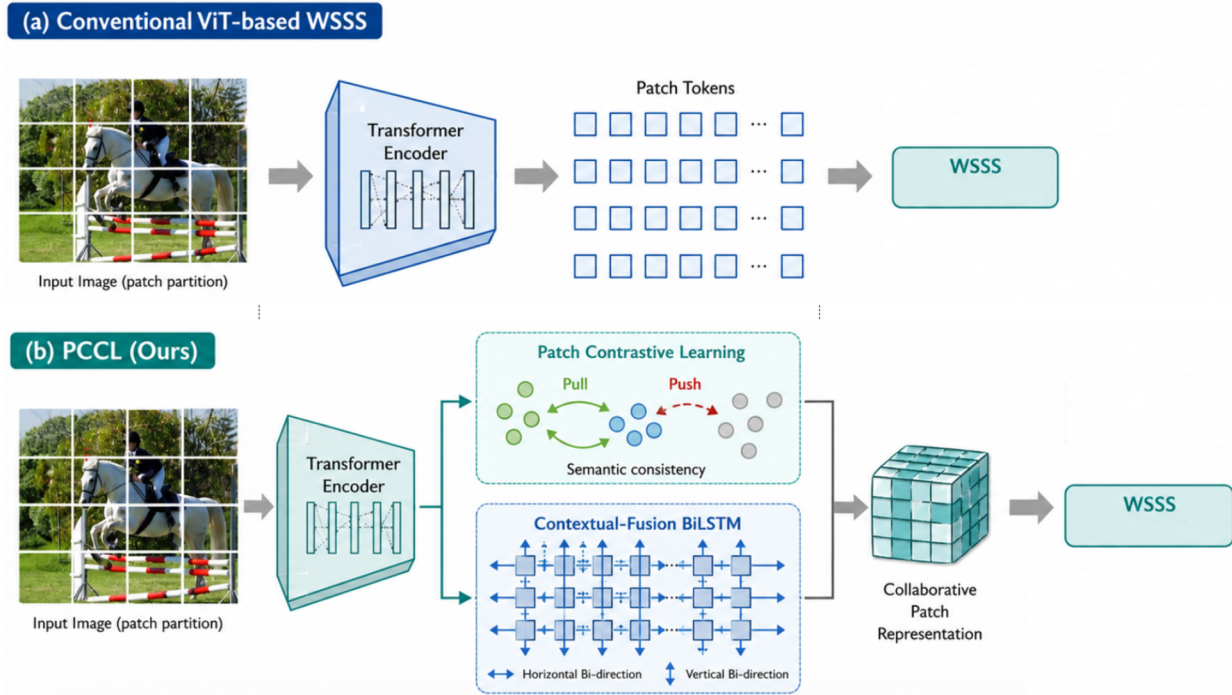


Figure 1: Main idea of the proposed PCCL framework. Conventional ViT-based WSSS methods mainly rely on direct patch-token representations, which may produce incomplete activation maps due to limited patch collaboration. In contrast, PCCL introduces patch contrastive learning to enhance semantic discrimination and a contextual-fusion bidirectional long short-term memory (BiLSTM) module to model bidirectional contextual dependencies among patches. The collaborative patch representation enables more complete and reliable activation maps for weakly supervised semantic segmentation.

2 Related Work

2.1 Weakly Supervised Semantic Segmentation

Weakly Supervised Semantic Segmentation (WSSS) aims to learn pixel-level segmentation models from weak annotations [1], among which image-level labels are particularly attractive due to their low annotation cost and easy scalability. A common paradigm in image-level WSSS first trains an image classification network and then extracts Class Activation Maps (CAMs) [2] as initial localization cues. These cues are further refined into pseudo segmentation masks, which are finally used to train a fully supervised segmentation network. Although this pipeline has achieved considerable progress, conventional CAMs usually highlight only the most discriminative object regions while ignoring less salient but semantically relevant parts. Consequently, the generated pseudo labels are often incomplete, noisy, and insufficient for training accurate segmentation models.

To alleviate this limitation, many studies have focused on improving the quality of CAMs and pseudo labels. Early works introduce post-processing or refinement modules, such as dense Conditional Random Fields (denseCRF) [3], AffinityNet [4], and AdvCAM [5], to expand object regions and suppress noisy activations. Other methods exploit external cues or auxiliary information to enhance localization reliability. For example, data augmentation strategies [6] have been investigated to improve the robustness of WSSS models without substantially modifying the core architecture. Saliency maps are also incorporated in [7,8] to provide additional foreground priors and help distinguish object regions from background distractors. In addition, Chang et al. [9] introduced self-supervised sub-category discovery to improve object localization by

exploring intra-class visual variations. Although these methods are effective, their performance still largely depends on the quality of initial CAMs and auxiliary cues.

Another important research direction attempts to enhance semantic representations and region-level relationships for more complete object localization. Prototype-based methods construct class-aware representations from reliable seed regions. For instance, IS-CAM [10] estimates image-specific prototypes by aggregating structure-aware seed areas according to spatial feature similarities. Du et al. [11] further use top-ranked CAM responses to aggregate pixel features and build class prototypes, thereby improving feature discrimination. Meanwhile, affinity-based methods learn pairwise semantic relationships to propagate activation responses. AffinityNet [4] learns semantic affinities from pseudo labels and uses random walk propagation to refine CAMs. Subsequent methods exploit intrinsic feature affinities from classification networks [12,13] or combine saliency and segmentation cues to improve affinity estimation. These approaches demonstrate that modeling relationships among pixels or regions is crucial for generating high-quality pseudo masks.

Despite these advances, existing WSSS methods still face difficulties in recovering object regions that are spatially separated, weakly activated, or visually diverse. Many methods mainly focus on refining local activation responses or learning pairwise affinities, while the collaborative relationships among image patches are not sufficiently explored. In practical scenarios, different parts of the same object may appear in different spatial locations and may receive inconsistent activation strengths. If these patches are treated independently, useful semantic information from confident regions may not be effectively propagated to ambiguous regions. In this work, we address this issue by introducing patch-level contextual collaboration and contrastive semantic regularization. Specifically, our method models bidirectional contextual dependencies among image patches and encourages semantically consistent patches to form compact feature representations, leading to more complete and reliable pseudo labels for WSSS. Different from conventional affinity propagation, which mainly estimates pairwise relationships for mask refinement, our contrastive regularization is applied to context-enhanced patch embeddings and explicitly modulates positive patch relationships according to their spatial proximity.

2.2 Transformers for Visual Tasks

Vision Transformer (ViT) [14] represents an image as a sequence of patch tokens and models long-range dependencies through self-attention. This property has made ViT particularly suitable for weakly supervised semantic segmentation, where patch-level relationships are important for recovering object regions from image-level supervision. Early ViT-based localization methods, such as TS-CAM [15], generate class-discriminative localization maps from token attention.

Recent ViT-based WSSS methods further exploit class tokens, attention maps, and token relationships. MCTformer [16] introduces multiple class tokens to learn class-to-patch attention maps and uses patch-to-patch attention for mask refinement. MCTformer+ [17] incorporates class-token contrast to improve semantic discrimination. AFA [18] learns semantic affinities by supervising multi-head self-attention maps, while ToCo [19] applies patch-token and class-token contrast to improve foreground-background discrimination and semantic consistency. ViT-PCM [20] adopts a CAM-independent design that directly estimates pixel-level class probabilities from ViT patch representations.

Although these methods are all based on ViT representations, they impose different relations on the resulting tokens. MCTformer and MCTformer+ emphasize class-to-patch attention and attention-derived patch affinity, whereas ToCo supervises patch-token and class-token relations through its PTC and CTC objectives. The closest architectural precedents for PCCL are CPF-CTE [21], TKP-PCL [22], and APC [23]. CPF-CTE combines post-ViT learnable class tokens with CF-BiLSTM contextual fusion but does not employ

pairwise patch contrast. TKP-PCL and APC combine confidence-derived patch contrast with horizontal-vertical recurrent refinement, but their objectives do not distinguish positive pairs according to patch-grid geometry. PCCL advances this line by conditioning positive-pair compactness on two-dimensional spatial distance and applying the resulting objective to context-enhanced embeddings.

Table 1 summarizes this boundary, while the following discussion explains its technical implication. PCCL shares standard pipeline elements, including the ViT backbone, pooling, CRF refinement, and downstream segmentation, with prior WSSS systems. Its distinct methodological contribution is the spatially weighted positive-pair objective together with its semantic-spatial coupling to recurrent contextual fusion.

Table 1: Component-level positioning of PCCL relative to closely related ViT-based WSSS methods. The central comparison concerns semantic relation learning, spatial/context relation modeling, and the interaction between contrastive supervision and contextual patch collaboration. Pooling, CRF, and the downstream segmentation network are summarized as conventional pipeline configurations.

Method	ViT	Primary Relation Target	Contrastive Relation Learning	Spatial/Context Relation Modeling	Coupling of Contrast and Context	Class-Token Role	Conventional Setup
CPF-CTE [21]	✓	Patch context and class semantics	–	BiLSTM contextual patch fusion	Class-token enhancement followed by context fusion	Post-ViT learnable class tokens	Top- <i>k</i> /CRF/DeepLabv2
TKP-PCL [22]	✓	Patch embeddings	PCE with confidence-derived patch pairs	H/V BiLSTM refinement	Recurrent refinement with non-spatially weighted PCE	–	Top- <i>k</i> /CRF/DeepLabv2
APC [23]	✓	Patch embeddings	PCL with confidence-derived patch pairs	H/V BiLSTM refinement	Recurrent refinement with non-spatially weighted PCL	–	Adaptive- <i>k</i> /CRF/end-to-end decoder
ToCo [19]	✓	Patch and class tokens	Patch-token and class-token contrast	Relations derived from intermediate tokens	Token-relation supervision; no explicit recurrent fusion	Standard class token used in CTC	–/CRF/end-to-end decoder
MCTformer [16]	✓	Class-to-patch attention	–	Patch-attention affinity	Class-specific localization with affinity refinement	Multiple class tokens	–/CRF/DeepLabv1
MCTformer + [17]	✓	Class tokens	Class-token contrast	Patch-attention affinity	Class-token discrimination with affinity refinement	Multiple class tokens with CCT	–/CRF/DeepLabv1
AFA [18]	✓	Attention affinities	–	Learned attention affinity	Affinity-supervised attention propagation	–	–/CRF/end-to-end decoder

(Continued)

Table 1 (continued)

Method	ViT	Primary Relation Target	Contrastive Relation Learning	Spatial/Context Relation Modeling	Coupling of Contrast and Context	Class-Token Role	Conventional Setup
ViT-PCM [20]	✓	Patch predictions	–	Direct patch-to-pixel prediction	–	–	Max/CRF/DeepLabv2
PCCL (ours)	✓	Semantic and spatial patch relations	Spatially weighted positive-pair contrast	CF-BiLSTM-based H/V patch collaboration	Spatially weighted contrast over context-collaborated embeddings	–	Top- <i>k</i> /CRF/DeepLabv2

The distinction of PCCL should therefore be interpreted from the interaction between its two representation-learning mechanisms rather than from either component alone. In a non-weighted patch contrastive objective, all prediction-mined positive pairs contribute equally once they pass the confidence threshold. Under image-level supervision, however, equal confidence does not imply equal relational reliability: nearby same-class patches are more likely to describe a coherent object region, whereas distant predictions may correspond to different instances, disconnected parts, or correlated false activations. PCCL encodes this structural prior through a continuous spatial weight. Nearby pairs receive stronger compactness constraints, while distant pairs remain involved with reduced influence. The objective therefore avoids indiscriminate same-class contraction while preserving long-range contextual interaction through CF-BiLSTM.

CF-BiLSTM propagates contextual evidence horizontally and vertically before contrastive regularization is evaluated. Consequently, the proposed loss does not act on isolated ViT tokens but on representations that already encode directional context. The spatial weight then controls the local strength of semantic compaction over those context-enhanced representations. This ordered interaction is central to PCCL: it combines what the patches represent after collaboration with how strongly their semantic relations should be optimized according to geometry.

This formulation establishes an explicit novelty boundary. Relative to CPF-CTE, PCCL replaces post-ViT class-token enhancement with patch-pair contrast over context-enhanced embeddings. Relative to TKP-PCL and APC, it extends confidence-mined patch contrast by introducing distance-dependent positive-pair weighting. Relative to ToCo, it operates on prediction-mined patch pairs after explicit recurrent H/V fusion rather than contrasting patch and class tokens. Thus, Eqs. (14)–(17) introduce a geometry-aware optimization mechanism that is absent from these closely related formulations.

However, ViT self-attention alone may still be insufficient under image-level supervision. Patch tokens corresponding to non-discriminative object regions can remain noisy or weakly activated, while global self-attention does not explicitly enforce directional propagation across the two-dimensional patch grid. Consequently, weakly activated regions may fail to receive sufficient complementary information from more confident object patches, resulting in fragmented activation maps and incomplete object localization.

Operationally, PCCL first applies Patch LSTM Context Fusion to exchange information along the horizontal and vertical directions of the patch grid and then applies Collaborative Patch Contrastive Learning to the resulting embeddings. The objective separates spatially weighted positive-pair compactness from positive–negative feature separation. Recurrent fusion addresses contextual fragmentation, while the

proposed weight injects patch-grid geometry into semantic compaction; their ordered interaction constitutes the semantic-spatial collaboration mechanism of PCCL.

3 Methodology

In this section, we introduce the overall architecture and key components of the proposed Patch Collaboration with Contrastive Learning (PCCL) framework for weakly supervised semantic segmentation. First, Section 3.1 presents the overall pipeline of PCCL. Then, Section 3.2 describes the proposed Patch LSTM Context Fusion module, which models bidirectional contextual dependencies among image patches and promotes information propagation across spatially correlated regions. Next, Section 3.3 introduces the Collaborative Patch Contrastive Learning strategy, which enhances semantic consistency and discriminative patch representation based on the context-enhanced patch features. Finally, Section 3.4 presents the final prediction process and the overall training objective.

3.1 Overall Framework

The overall architecture of the proposed PCCL framework is illustrated in Fig. 2. Given an input image $I \in \mathbb{R}^{H \times W \times 3}$ with only image-level labels, we first divide it into a sequence of non-overlapping patches. These patches are fed into a Vision Transformer (ViT) encoder to obtain initial patch token representations. Formally, the patch embeddings extracted by the ViT backbone are denoted as:

$$\mathbf{F}^{\text{vit}} = \{\mathbf{f}_1^{\text{vit}}, \mathbf{f}_2^{\text{vit}}, \dots, \mathbf{f}_s^{\text{vit}}\} \in \mathbb{R}^{s \times d}, \quad (1)$$

where s is the number of image patches and d denotes the feature dimension of each patch token.

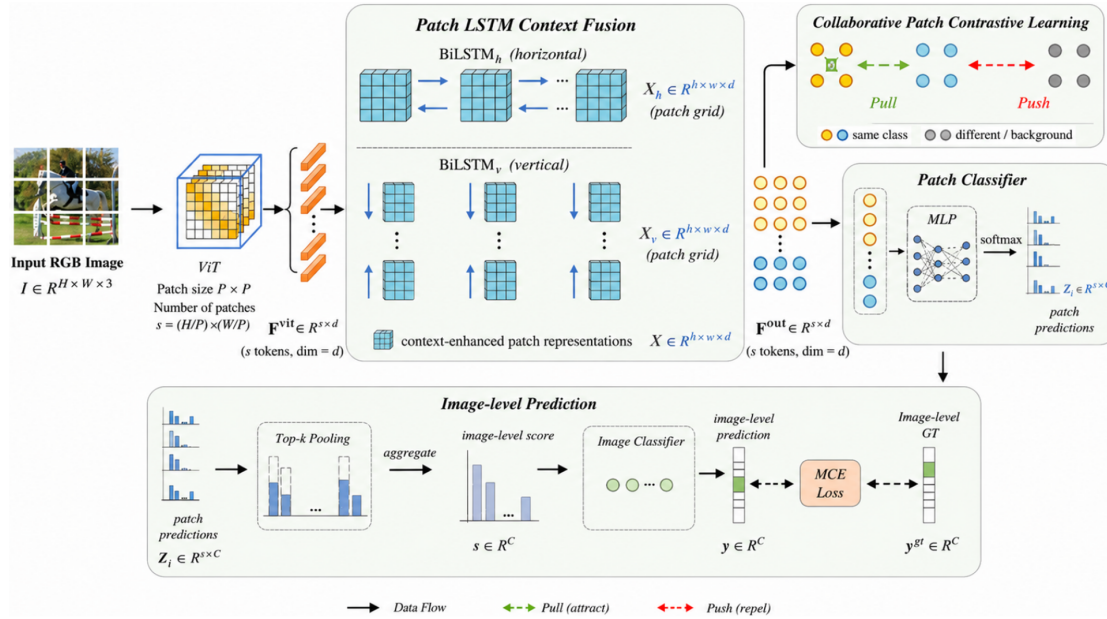


Figure 2: Overview of the proposed patch collaboration with contrastive learning (PCCL) framework. Given an input image, a ViT backbone first extracts patch-level token representations. The proposed Patch LSTM Context Fusion module then propagates bidirectional contextual information among spatially correlated patches to obtain context-enhanced patch embeddings. Based on these refined representations, collaborative patch contrastive learning pulls semantically consistent patches closer while pushing different semantic patches apart. The learned collaborative patch representations are further used to generate more complete and reliable activation maps for weakly supervised semantic segmentation.

Although ViT is effective in capturing global dependencies through self-attention, directly using its patch tokens for weakly supervised semantic segmentation may still lead to incomplete activation maps. This is because semantically related object regions can be spatially separated, locally ambiguous, or weakly activated. As a result, patch representations generated by the transformer alone may not fully exploit collaborative relationships among object regions.

To address this issue, PCCL introduces a two-stage patch collaboration mechanism. First, we design a Patch LSTM Context Fusion module to model bidirectional dependencies among patches and propagate contextual cues across spatially correlated regions. This module converts the original ViT patch tokens $\mathbf{F}^{vit}$ into context-enhanced patch embeddings $\mathbf{F}^{out}$. Second, we apply a Collaborative Patch Contrastive Learning strategy on $\mathbf{F}^{out}$ to further improve semantic consistency and patch-level discrimination. Specifically, semantically consistent patches are encouraged to be close in the feature space, while different semantic patches are pushed apart.

The complete training procedure of PCCL is summarized in Algorithm 1. The algorithm provides a concise overview of the main steps, including ViT feature extraction, two-dimensional patch-grid contextual propagation, patch-level contrastive learning, image-level prediction, and joint optimization of the classification and contrastive objectives.

Algorithm 1: Training procedure of PCCL

Require: Training image $\mathbf{I}$ with image-level label $\mathbf{y}$

Ensure: Optimized PCCL model and generated pseudo labels

- 1: Extract patch tokens $\mathbf{F}^{vit}$ using ViT encoder
 - 2: Reshape patch tokens into a two-dimensional patch grid
 - 3: Apply horizontal BiLSTM and vertical BiLSTM to obtain $\mathbf{F}^{out}$
 - 4: Predict patch-level class probabilities using the patch classifier
 - 5: Select positive and negative patch candidates based on prediction confidence
 - 6: Compute spatial weights $\omega_{i,j}$ for positive patch pairs
 - 7: Calculate collaborative patch contrastive loss $\mathcal{L}_{con}$
 - 8: Aggregate patch predictions using Top- k pooling
 - 9: Compute image-level classification loss $\mathcal{L}_{cls}$
 - 10: Optimize the joint objective:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{con} \mathcal{L}_{con}$$
 - 11: Generate pseudo labels from patch localization maps
-

After obtaining the refined patch representations, a lightweight patch classifier is used to predict patch-level class probabilities. These patch predictions are aggregated into image-level classification scores through Top- k pooling and supervised by image-level labels. Meanwhile, the patch-level prediction maps can be upsampled and refined to generate pseudo-labels, which are subsequently used to train a fully supervised segmentation network, such as DeepLabv2 [24]. Through the joint modeling of contextual collaboration and contrastive semantic consistency, PCCL produces more complete and reliable activation maps for weakly supervised semantic segmentation.

3.2 Patch LSTM Context Fusion

To enhance contextual collaboration among image patches, we propose a Patch LSTM Context Fusion module. The motivation is that patch tokens extracted by ViT may still suffer from spatial fragmentation when only image-level supervision is available. Although self-attention can capture global correlations, it

does not explicitly enforce sequential or spatial continuity among neighboring patches. Therefore, weakly activated object regions may fail to receive sufficient information from more confident regions of the same object.

Given the patch token sequence $\mathbf{F}^{\text{vit}} \in \mathbb{R}^{s \times d}$ extracted by the ViT encoder, we reshape it into a two-dimensional patch feature map:

$$\mathbf{X} = \text{Reshape}(\mathbf{F}^{\text{vit}}) \in \mathbb{R}^{h \times w \times d}, \quad (2)$$

where h and w denote the height and width of the patch grid, and $s = h \times w$. To model contextual dependencies from different spatial directions, we perform bidirectional sequence modeling along both horizontal and vertical dimensions.

For horizontal contextual modeling, each row of the patch feature map is treated as a sequence. A forward LSTM and a backward LSTM are applied to capture left-to-right and right-to-left dependencies:

$$\vec{\mathbf{h}}_{i,j}^H = \overrightarrow{\text{LSTM}}^H(\mathbf{x}_{i,j}, \vec{\mathbf{h}}_{i,j-1}^H), \quad (3)$$

$$\overleftarrow{\mathbf{h}}_{i,j}^H = \overleftarrow{\text{LSTM}}^H(\mathbf{x}_{i,j}, \overleftarrow{\mathbf{h}}_{i,j+1}^H), \quad (4)$$

where $\mathbf{x}_{i,j}$ represents the patch feature at spatial position (i, j) , and $\vec{\mathbf{h}}_{i,j}^H$ and $\overleftarrow{\mathbf{h}}_{i,j}^H$ denote the hidden states from the two horizontal directions.

Similarly, for vertical contextual modeling, each column of the patch feature map is regarded as a sequence. The vertical bidirectional LSTM captures top-to-bottom and bottom-to-top dependencies:

$$\vec{\mathbf{h}}_{i,j}^V = \overrightarrow{\text{LSTM}}^V(\mathbf{x}_{i,j}, \vec{\mathbf{h}}_{i-1,j}^V), \quad (5)$$

$$\overleftarrow{\mathbf{h}}_{i,j}^V = \overleftarrow{\text{LSTM}}^V(\mathbf{x}_{i,j}, \overleftarrow{\mathbf{h}}_{i+1,j}^V), \quad (6)$$

where $\vec{\mathbf{h}}_{i,j}^V$ and $\overleftarrow{\mathbf{h}}_{i,j}^V$ encode vertical contextual information from opposite directions.

The horizontal and vertical contextual representations are then fused to obtain the context-enhanced patch representation:

$$\mathbf{g}_{i,j} = \text{Concat}(\vec{\mathbf{h}}_{i,j}^H, \overleftarrow{\mathbf{h}}_{i,j}^H, \vec{\mathbf{h}}_{i,j}^V, \overleftarrow{\mathbf{h}}_{i,j}^V). \quad (7)$$

To maintain the original feature dimension and ensure compatibility with subsequent modules, a learnable projection layer is applied:

$$\mathbf{f}_{i,j}^{\text{out}} = \mathbf{W}_p \mathbf{g}_{i,j} + \mathbf{b}_p, \quad (8)$$

where $\mathbf{W}_p$ and $\mathbf{b}_p$ are learnable projection parameters. The final context-enhanced patch embeddings are denoted as:

$$\mathbf{F}^{\text{out}} = \{\mathbf{f}_1^{\text{out}}, \mathbf{f}_2^{\text{out}}, \dots, \mathbf{f}_s^{\text{out}}\} \in \mathbb{R}^{s \times d}. \quad (9)$$

Relationship to and advance over closely related designs. PCCL shares the ViT-based H/V contextual-refinement foundation of TKP-PCL [22], APC [23], and CPF-CTE [21]. The methods nevertheless optimize different relations. CPF-CTE emphasizes post-ViT class-token enhancement and contextual fusion without pairwise patch contrast. TKP-PCL and APC construct patch pairs from confidence predictions but do not

encode the relative geometry of positive pairs in their contrastive objectives. PCCL introduces this missing relation after contextual fusion: Eqs. (14)–(17) optimize the fused embeddings with a distance-dependent positive-pair term while maintaining a separate positive–negative discrimination term. Consequently, PCCL transforms spatial layout from an implicit property of recurrent processing into an explicit factor in the contrastive learning objective.

Through this bidirectional contextual fusion process, each patch receives complementary information from surrounding patches and other spatially correlated patches. This is particularly beneficial for WSSS, where object regions are often partially activated. By propagating information from confident object patches to weakly activated regions, the proposed module helps recover incomplete object structures and improves the coherence of activation maps.

3.3 Collaborative Patch Contrastive Learning

After obtaining the context-enhanced patch embeddings $\mathbf{F}^{\text{out}}$, we further introduce Collaborative Patch Contrastive Learning to improve patch-level semantic discrimination. Different from contrastive strategies that only separate high-confidence and low-confidence patches, our method constructs contrastive relationships based on semantic consistency and contextual collaboration. The key idea is to encourage semantically related patches to form compact representations while pushing semantically different patches away in the feature space.

As shown in Fig. 2, the patch classifier takes $\mathbf{F}^{\text{out}}$ as input and produces patch-level semantic predictions:

$$\mathbf{Z} = \text{softmax}(\mathbf{F}^{\text{out}}\mathbf{W}_{\text{cls}}), \quad (10)$$

where $\mathbf{W}_{\text{cls}} \in \mathbb{R}^{d \times C}$ is the classifier weight matrix, and C denotes the total number of semantic classes, including one background class and $C - 1$ foreground classes. Accordingly, $\mathbf{Z} \in \mathbb{R}^{s \times C}$ represents the normalized categorical distribution of each patch over all semantic classes. The softmax operation is performed along the semantic-class dimension.

It is important to distinguish patch-level categorical prediction from image-level multi-label classification. In image-level WSSS, multiple foreground categories may simultaneously exist in one image. Therefore, the final image-level prediction is formulated as a multi-label classification problem and optimized using a binary cross-entropy loss over the $C - 1$ foreground classes. In contrast, each patch represents a localized image region and is assigned to one dominant semantic category, including the background category. This patch-level formulation is analogous to pixel-wise semantic segmentation, where each spatial unit is assigned a single semantic label. Therefore, softmax normalization is applied across all C patch-level categories to model the relative semantic response within each local region.

The resulting patch-level distributions are used for confident patch selection and contrastive sample construction. During pseudo-label generation, the semantic label of the i -th patch is obtained by selecting the category with the highest predicted probability:

$$\hat{c}_i = \arg \max_{c \in \{0, 1, \dots, C-1\}} Z_i^c, \quad (11)$$

where $c = 0$ denotes the background category and $c \in \{1, \dots, C - 1\}$ denotes a foreground category. Thus, a patch is assigned to the background category whenever Z_i^0 is the largest entry in its categorical distribution.

The prediction of the i -th patch is denoted as Z_i .

For each foreground class $c \in \{1, \dots, C-1\}$, we define a set of positive patch candidates according to the patch prediction score and the image-level label:

$$\mathcal{P}_c^+ = \{\mathbf{f}_i^{\text{out}} \mid Z_i^c > \varepsilon, y_c = 1\}, \quad (12)$$

where ε is a confidence threshold and y_c indicates whether foreground class c exists in the image-level label. These patches are regarded as semantically consistent candidates for class c . Meanwhile, patches with low response to class c are treated as negative candidates:

$$\mathcal{P}_c^- = \{\mathbf{f}_i^{\text{out}} \mid Z_{i,c} < 1 - \varepsilon\}, \quad (13)$$

where $1 - \varepsilon$ serves as an adaptive low-confidence threshold. By defining the negative threshold as the complement of the positive confidence threshold ε , we explicitly enforce a symmetric clearing margin between highly confident foreground regions and background/unrelated patches, thereby effectively reducing the number of hyperparameters and enhancing the robustness of the selection mechanism. For images containing multiple semantic categories, the patch mining process is performed independently for each foreground class indicated by the image-level labels. Therefore, patches belonging to different object categories can construct class-specific positive and negative relationships without interfering with each other. For background regions or patches unrelated to a target category, their low prediction responses naturally allow them to be selected as negative candidates. The class-specific validity rule excludes terms for which sufficient positive and negative patch relationships cannot be formed, thereby preventing ill-defined or statistically unstable pairwise updates. Prediction-driven patch mining can introduce confirmation bias: an incorrect high-confidence prediction may be selected as a positive candidate and subsequently reinforced by the contrastive objective, while a true object patch with an erroneously low response may be treated as a negative candidate. PCCL reduces, rather than completely eliminates, this risk through three complementary safeguards. First, the confidence margin selects only patches satisfying $Z_i^c > \varepsilon$ or $Z_i^c < 1 - \varepsilon$ and excludes uncertain intermediate responses. Second, the image-level presence constraint $y_c = 1$ prevents positive candidates from being constructed for categories absent from the image. Third, candidate predictions are obtained from CF-BiLSTM-enhanced representations, so their confidence incorporates horizontal and vertical contextual evidence instead of relying on isolated local appearance. We further assess candidate quality in the experimental analysis; pixel-level annotations are used only for post-hoc evaluation, not for training or candidate selection.

A class-specific term is valid only when $|\mathcal{P}_c^+| \geq 2$ and $|\mathcal{P}_c^-| \geq 1$, ensuring that at least one positive pair and one positive-negative pair can be constructed. The overall contrastive loss is averaged only over the valid foreground-class set $\mathcal{C}_{\text{valid}}^+$. Classes that do not meet these conditions are excluded from that image's contrastive average; if $\mathcal{C}_{\text{valid}}^+$ is empty for an image, $\mathcal{L}_{\text{con}}$ is set to zero for that image. This validity safeguard was enabled in all reported experiments, including the component ablations and the multi-seed evaluation; consequently, the reported performance reflects the practical operation of PCCL when some image-class combinations do not provide sufficient patch pairs.

The current training records do not include aggregate or per-class activation rates for this validity condition. Although this does not affect the execution of the proposed loss or the reported results, recording these statistics would provide additional insight into class-dependent contrastive-supervision coverage.

To measure the similarity between two patch embeddings, we adopt the cosine similarity:

$$\text{sim}(\mathbf{f}_i^{\text{out}}, \mathbf{f}_j^{\text{out}}) = \frac{(\mathbf{f}_i^{\text{out}})^T \mathbf{f}_j^{\text{out}}}{\|\mathbf{f}_i^{\text{out}}\|_2 \|\mathbf{f}_j^{\text{out}}\|_2}. \quad (14)$$

To further regulate the positive-pair relationships among the context-enhanced patch embeddings, we introduce a spatial weighting term. A uniform objective assumes that every selected same-class pair is equally reliable. This assumption is restrictive for prediction-driven mining because two high-confidence patches can be semantically compatible yet differ in local continuity, object instance, or prediction reliability. The proposed weighting strategy therefore assigns stronger compactness constraints to spatially close positive patches, which more often belong to a coherent local object region, while retaining weaker contributions from distant pairs. Importantly, spatial distance does not determine the semantic label or remove long-range pairs; it only modulates their contribution after semantic candidate selection.

The key insight is to separate semantic eligibility from relational reliability. Prediction confidence and image-level labels determine whether two patches can form a positive pair, whereas their patch-grid distance provides an additional soft estimate of how strongly that selected relation should be enforced. Thus, the spatial weight does not replace semantic selection or long-range contextual modeling; it only prevents all selected positive pairs from being compacted with identical strength.

For two positive patches located at positions $\mathbf{r}_i$ and $\mathbf{r}_j$, their spatial weight is defined as:

$$\omega_{i,j} = \exp\left(-\frac{\|\mathbf{r}_i - \mathbf{r}_j\|_2^2}{\tau_s}\right), \quad (15)$$

where $\mathbf{r}_i$ and $\mathbf{r}_j$ denote their two-dimensional patch-grid coordinates, and τ_s controls the decay rate of the weight with respect to spatial distance. We set $\tau_s = 1.0$ in our experiments. This setting provides a moderate spatial decay: nearby positive patches receive stronger compactness constraints, whereas spatially distant positive pairs remain involved with reduced weights. In this way, the proposed objective emphasizes local object continuity without forcing all same-class patches to align with identical strength. Meanwhile, long-range contextual information is captured by the preceding Patch LSTM Context Fusion module.

The collaborative patch contrastive loss for class c is formulated as:

$$\begin{aligned} \mathcal{L}_{\text{con}}^c = & \frac{1}{N_c^+} \sum_{\substack{\mathbf{f}_i, \mathbf{f}_j \in \mathcal{P}_c^+ \\ i \neq j}} \omega_{i,j} (1 - \text{sim}(\mathbf{f}_i, \mathbf{f}_j)) \\ & + \frac{1}{N_c^-} \sum_{\mathbf{f}_i \in \mathcal{P}_c^+} \sum_{\mathbf{f}_k \in \mathcal{P}_c^-} \text{sim}(\mathbf{f}_i, \mathbf{f}_k), \end{aligned} \quad (16)$$

where N_c^+ and N_c^- denote the numbers of positive-positive and positive-negative patch pairs for class c , respectively. The first term pulls semantically consistent positive patches closer, encouraging compact class-specific representations. The second term pushes different semantic patches away from foreground object regions, reducing noisy activations and improving semantic discrimination. Since the contrastive loss is applied after Patch LSTM Context Fusion, the learned representations are guided not only by local visual appearance but also by bidirectional contextual information.

The overall collaborative contrastive loss is computed over the foreground classes for which valid pairs can be formed:

$$\mathcal{L}_{\text{con}} = \frac{1}{|\mathcal{C}_{\text{valid}}^+|} \sum_{c \in \mathcal{C}_{\text{valid}}^+} \mathcal{L}_{\text{con}}^c, \quad (17)$$

where $\mathcal{C}^+ = \{c \mid y_c = 1\}$ is the set of foreground classes present in the image and $\mathcal{C}_{\text{valid}}^+ = \{c \in \mathcal{C}^+ \mid |\mathcal{P}_c^+| \geq 2, |\mathcal{P}_c^-| \geq 1\}$. The zero-valid-class case uses $\mathcal{L}_{\text{con}} = 0$ rather than evaluating the fraction above. Through this design, PCCL explicitly models patch-level semantic consistency and contextual collaboration, enabling weakly activated object regions to be reinforced by semantically related patches.

The proposed objective addresses a specific question: whether patch-grid relationships should modulate the compactness constraint between prediction-mined positive patch pairs. Its pairwise formulation separates positive compactness from positive-negative separation and permits the spatial weight $\omega_{i,j}$ to act directly on the positive term. By contrast, InfoNCE introduces a temperature-scaled softmax denominator and a negative set, supervised contrastive learning generally constructs positive sets from explicit semantic labels, and margin or triplet objectives require additional margin and sampling strategies. Replacing the proposed objective with these alternatives would therefore vary several optimization and sampling factors simultaneously. The controlled comparison isolates the spatial weighting mechanism by changing only $\omega_{i,j}$ while keeping patch selection, pair construction, feature representation, and all other training settings unchanged.

For implementation-level reproducibility, [Appendix A](#) and [Algorithm A1](#) provide pseudocode for the complete spatially weighted contrastive-loss computation, including feature normalization, candidate masking, unordered positive-pair construction, spatial weighting, valid-class filtering, and image- and batch-level aggregation.

3.4 Final Loss

To train the proposed framework using only image-level labels, we first aggregate patch-level predictions into image-level classification scores. As shown in [Fig. 2](#), the patch classifier outputs patch-level predictions $\mathbf{Z}$, which are then aggregated through Top- k pooling. For class c , the image-level prediction score is computed as:

$$p_c = \frac{1}{k} \sum_{i \in \mathcal{T}_c^k} Z_i^c, \quad (18)$$

where C is the total number of patch-level categories, background is indexed by $c = 0$, and the number of foreground categories is $C_f = C - 1$ with indices $c \in \{1, \dots, C_f\}$. The set $\mathcal{T}_c^k$ contains the indices of the top k patches for foreground class c . This operation encourages the image-level prediction to focus on the most discriminative and reliable patch responses.

Because image-level annotations are available only for foreground presence or absence, the image-level multi-label classification loss excludes the background category and is defined as:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{C_f} \sum_{c=1}^{C_f} [y_c \log(p_c) + (1 - y_c) \log(1 - p_c)], \quad (19)$$

where y_c is the image-level presence label for foreground class c . No image-level label y_0 or BCE term is defined for background; background participates only in the C -way patch prediction and pseudo-label assignment.

The final training objective combines the image-level classification loss and the collaborative patch contrastive loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{con}} FL_{\text{con}}, \quad (20)$$

where λ_{con} is a balancing coefficient that controls the contribution of the contrastive objective.

During inference, the trained patch classifier produces patch-level prediction maps. These maps are reshaped into the spatial patch grid and upsampled to the original image resolution. To further improve boundary consistency and suppress noisy responses, Conditional Random Fields (CRF) [25] are applied as a post-processing step. We adopt the publicly available DenseCRF configuration of ViT-PCM [20] without additional parameter tuning for PCCL. Specifically, we perform 10 mean-field inference iterations. The Gaussian pairwise potential uses a compatibility weight of 3 and a spatial standard deviation of 1, while the bilateral pairwise potential uses a compatibility weight of 4, a spatial standard deviation of 67, and an RGB standard deviation of 3. CRF is used only as a conventional offline refinement step; it is not jointly trained with PCCL and is not included in the proposed representation-learning contribution. The resulting pseudo-labels are then used to train a fully supervised semantic segmentation model, such as DeepLabv2 [24], to obtain the final segmentation results.

By combining Patch LSTM Context Fusion and Collaborative Patch Contrastive Learning, PCCL effectively addresses two key limitations of existing ViT-based WSSS methods. The former strengthens contextual information propagation among spatially correlated patches, while the latter improves semantic compactness and discrimination in the patch feature space. As a result, the proposed framework generates more complete activation maps and more reliable pseudo-labels for weakly supervised semantic segmentation.

4 Experiment

4.1 Experimental Settings

Dataset. To validate the performance of our proposed model, we conduct extensive experiments on two widely adopted weakly supervised semantic segmentation benchmarks, including PASCAL VOC 2012 [26] and MS COCO 2014 [27]. PASCAL VOC 2012 contains 20 foreground object categories and one background class, covering diverse visual concepts such as vehicles, animals, and indoor objects. Following the standard protocol in weakly supervised semantic segmentation, we augment the original training set with additional annotations from the Semantic Boundaries Dataset (SBD) [28]. This results in 10,582 weakly labeled training images and 1,449 validation images.

In addition, we further evaluate our method on the more challenging MS COCO 2014 dataset, which contains 80 foreground object categories and one background class. Compared with PASCAL VOC 2012, MS COCO presents more complex scenes, larger object-scale variations, frequent object co-occurrence, and heavier background clutter, making it a more demanding benchmark for evaluating the robustness and generalization ability of WSSS methods. Following common practice, we use the COCO 2014 training images with image-level labels for model training and report segmentation performance on the validation set. Experiments on both datasets enable a comprehensive evaluation of the proposed method under different levels of scene complexity and category diversity.

Evaluation Metric. For quantitative assessment, we employ the standard mean Intersection-over-Union (mIoU) as our primary evaluation criterion. The mIoU metric serves as a rigorous benchmark in semantic segmentation by penalizing both false positives and false negatives at the pixel level. Specifically, the IoU score for an individual category c is determined by calculating the ratio of the intersection to the union between the prediction and ground truth:

$$\text{IoU}_c = \frac{|P_c \cap G_c|}{|P_c \cup G_c|}, \quad (21)$$

where P_c and G_c denote the predicted pixel set and the corresponding ground-truth mask for class c , respectively. To obtain a comprehensive overview of the model's overall efficacy, the final mIoU is computed by averaging the IoU values across all C semantic classes (including the background):

$$\text{mIoU} = \frac{1}{C} \sum_{c=0}^{C-1} \text{IoU}_c. \quad (22)$$

By treating each category with equal weight during the averaging process, mIoU effectively mitigates the impact of class imbalance, ensuring that minor classes contribute equally to the final score. Consequently, this metric provides a reliable and transparent reflection of our model's capacity to delineate precise object boundaries and maintain high taxonomic accuracy.

We report the primary results on the PASCAL VOC 2012 validation set and the MS COCO 2014 validation set. The final configuration, including $k = 6$, $\varepsilon = 0.8$, $\lambda_{\text{con}} = 0.01$, and the 960×960 inference resolution, was selected through a grid search based on pseudo-label quality on the PASCAL VOC training set. After this configuration was determined, all hyperparameters were fixed before evaluating the final model on the VOC validation set. The same fixed configuration was subsequently used for MS COCO 2014. Following the standard benchmark-reporting practice in WSSS, [Tables 2–4](#) compare PCCL with the results reported in the original publications, with each value linked to its cited source. Complete official code, checkpoints, and fully specified training configurations are not available for all selected methods; replacing their published results with partial unofficial reimplementations could therefore introduce additional reproduction bias. We accordingly use the published values for transparent benchmark-level positioning. The controlled evidence for the effects of the PCCL components is provided separately by the internal ablation studies, in which the training and evaluation settings are held fixed. The contrastive objective is active only during first-stage training, while Patch LSTM Context Fusion is used during first-stage training and offline pseudo-label generation. Neither component is retained in the downstream DeepLabv2 network; therefore, the architecture, parameter count, FLOPs, and deployment-time inference throughput of the final segmentation model are unchanged relative to the same DeepLabv2 configuration trained with other pseudo masks.

Table 2: Comparison with literature-reported pseudo-label results on the PASCAL VOC 2012 validation set (mIoU). Results for comparison methods are taken from their cited publications.

Method	Pub.	Backbone	mIoU (%)
CDA [29]	ICCV21	V1-RN38	66.4
Yao et al. [7]	CVPR21	V2-RN101	68.3
AuxSegNet [30]	ICCV21	V1-RN38	69.0
AdvCAM [5]	IT-PAMI24	V2-RN101	69.9
SIPE [10]	CVPR22	ResNet50	58.6
AFA [18]	CVPR22	ViT-B/16	66.0
FPR [31]	ICCV23	ResNet38	68.5
PGSD [32]	TCSVT24	ViT-B/16	68.7
CGM [33]	PR24	ViT-B/16	68.1
CPF-CTE [21]	CMES26	ViT-B/16	70.8
PCCL	Ours	ViT-B/16	71.7

Note: Boldface identifies the proposed method and the highest mIoU value among the methods listed in each table.

Table 3: Comparison with literature-reported semantic segmentation results on the PASCAL VOC 2012 validation set (mIoU). Results for comparison methods are taken from their cited publications.

Method	Pub.	Backbone	mIoU (%)
AFA [18]	CVPR22	ViT-B/16	63.8
SIPE [10]	CVPR22	ResNet50	58.6
MCTformer [16]	CVPR22	DeiT-S	61.7
TSCD [34]	AAAI23	MiT-B1	67.3
PGSD [32]	TCSVT24	ViT-B/16	68.7
CGM [33]	PR24	ViT-B/16	67.8
CPEWS [35]	CMC25	ViT	69.8
CPF-CTE [21]	CMES26	ViT-B/16	69.5
PCCL	Ours	ViT-B/16	71.1

Note: Boldface identifies the proposed method and the highest mIoU value among the methods listed in each table.

Table 4: Comparison with literature-reported semantic segmentation results on the MS COCO 2014 validation set (mIoU). Results for comparison methods are taken from their cited publications.

Model	Pub.	Backbone	mIoU (%)
MCTformer [16]	CVPR22	DeiT-S	42.0
SIPE [10]	CVPR22	ResNet38	43.6
TSCD [34]	AAAI23	MiT-B1	40.1
SAS [36]	AAAI23	ViT-B/16	44.5
FPR [31]	ICCV23	ResNet38	43.9
SFC [37]	AAAI24	ViT-B/16	44.6
IACD [6]	ICASSP24	ViT-B/16	44.3
ToCo [19]	CVPR23	ViT-B	42.3
DeFB-SV [38]	Neurocomputing24	ViT	43.5
CPEWS [35]	CMC25	ViT	41.4
CPF-CTE [21]	CMES26	ViT-B/16	45.4
PCCL	Ours	ViT-B/16	45.9

Note: Boldface identifies the proposed method and the highest mIoU value among the methods listed in each table.

4.2 Experimental Implementation Details

In our experiments, we adopt ViT-B/16 as the visual backbone to extract patch-level representations from input images. During training, all images are resized to 384×384 , following common practice in weakly supervised semantic segmentation [1]. Each image is divided into 24×24 non-overlapping patches, where each patch has a spatial size of 16×16 . The resulting patch tokens are first processed by the ViT encoder and then fed into the proposed Patch LSTM Context Fusion module to obtain context-enhanced patch embeddings. These refined patch representations are further optimized by the Collaborative Patch Contrastive Learning objective, which encourages semantically consistent patches to be compact while separating category-inconsistent patches.

The model is trained with a batch size of 16 for 30 epochs on two NVIDIA RTX 4090 GPUs. We use the Adam optimizer for network optimization. The learning rate is initialized as 1×10^{-3} for the first two epochs to accelerate early convergence and is then reduced to 1×10^{-4} for the remaining training epochs. This

two-stage learning rate schedule provides stable optimization for both the patch classifier and the proposed contextual collaboration modules.

During the training phase, as illustrated in Fig. 3, we employ a Top- k pooling strategy with $k = 6$ to selectively retain the most informative patch-level predictions. This approach ensures that the model focuses on confident patch responses while reducing the influence of uncertain regions. The values of $k = 6$, $\varepsilon = 0.8$, $\lambda_{\text{con}} = 0.01$, and the 960×960 inference resolution were selected through a grid search based on pseudo-label quality on the PASCAL VOC 2012 training set. After the final configuration was determined, all settings were fixed before evaluation on the VOC validation set and were subsequently kept unchanged for MS COCO 2014.

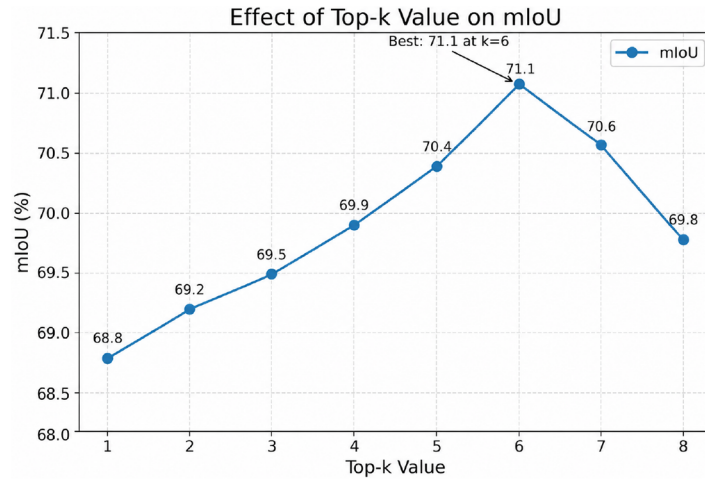


Figure 3: Impact of different Top- k values on WSSS performance.

Reusing this configuration on MS COCO was a deliberate no-retuning transfer protocol rather than an assumption that the VOC-selected values are universally optimal. The two datasets use the same ViT-B/16 backbone, 384×384 training input, and 24×24 patch grid; therefore, k is defined on the same token geometry. The confidence threshold ε acts on normalized class probabilities, and positive candidates are additionally restricted to categories present in the image-level label. Moreover, $\mathcal{L}_{\text{con}}^c$ is normalized by the numbers of valid positive–positive and positive–negative pairs, and the overall contrastive loss is averaged over valid foreground classes. These design choices reduce direct dependence of ε and λ_{con} on dataset size, category frequency, and the total number of categories. The fixed 960×960 resolution also preserves an identical offline pseudo-label-generation protocol across datasets. Applying the frozen configuration to the substantially more diverse MS COCO dataset yields 45.9% mIoU, providing empirical evidence of reasonable transferability without COCO-specific tuning. Nevertheless, no separate COCO grid search was performed, and we do not claim that this configuration is optimal for COCO.

The potential zero-retuning gap arises because COCO differs substantially from VOC in object scale, category co-occurrence, class frequency, and scene complexity. Consequently, the VOC-selected k may not provide the optimal pooling range for COCO images, ε may induce a different candidate precision–coverage balance, λ_{con} may not be the optimal loss balance under the 80-category distribution, and the fixed 960×960 pseudo-label-generation resolution may remain limiting for small objects. A COCO-specific grid search could therefore potentially improve the reported result. However, because no such retuning was performed, the magnitude of this gap cannot be estimated reliably from the frozen experiment and we avoid assigning it a speculative numerical value. The three frozen-configuration runs obtain 45.6%, 45.9%, and 46.2% mIoU

(45.90 ± 0.30), showing that the zero-retuning result is repeatable. We therefore interpret 45.9% as a stable, conservative transfer result rather than a COCO-optimized upper bound.

For the Collaborative Patch Contrastive Learning module, the predefined high-confidence threshold ε for selecting positive patch candidates is set to 0.8, which automatically yields an adaptive low-confidence threshold of $1 - \varepsilon = 0.2$ for identifying negative candidates. Here, negative candidates are defined with respect to the target category rather than being treated as global background regions. Specifically, a patch with a low response to class c indicates that it is unlikely to belong to class c , while it may still contain another foreground category in multi-object images. This coupled parameter design simplifies the optimization landscape while guaranteeing that only reliable patch representations participate in the contrastive alignment. The final joint objective is regularized by setting the balancing coefficient $\lambda_{\text{con}} = 0.01$, which ensures a stable trade-off between image-level multi-label classification and patch-level semantic consistency optimization.

The validation-set results in Figs. 3–5 are post-hoc sensitivity analyses that illustrate the effect of varying individual parameters after the training-set grid search had determined the final configuration. These analyses were not used to alter the selected hyperparameters or inference resolution.

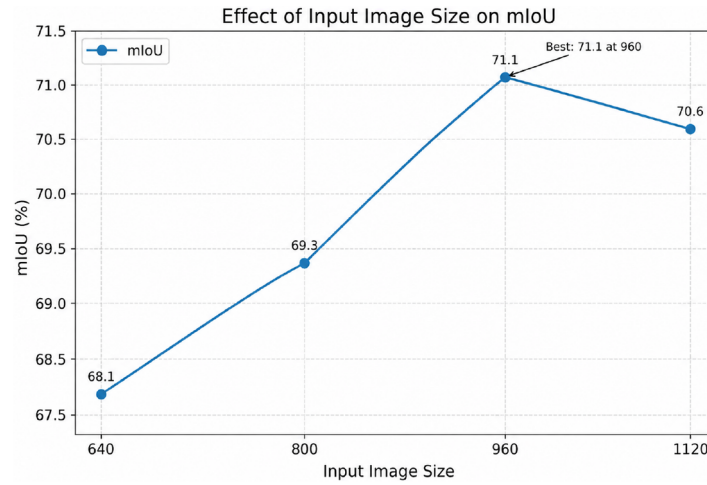


Figure 4: The impact of different input image sizes on the final performance of WSSS.

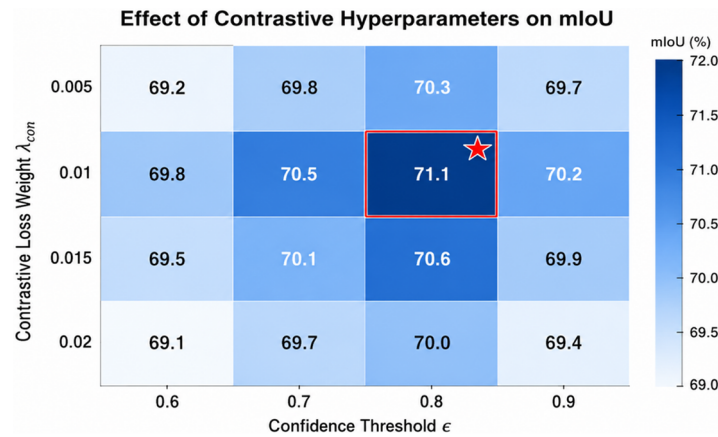


Figure 5: Impact of contrastive hyperparameters on segmentation performance. The heatmap reports the mIoU results under different combinations of the contrastive loss weight λ_{con} and the confidence threshold ε . The best performance is achieved when $\lambda_{\text{con}} = 0.01$ and $\varepsilon = 0.8$, yielding an mIoU of 71.1%. * indicates the best-performing hyperparameter configuration.

For the inference stage, as illustrated in Fig. 4, we preprocess the input images by upscaling them to a resolution of 960×960 . The higher inference resolution provides finer patch-level predictions and can help preserve small object regions and boundary details, which is beneficial for generating more accurate pseudo-labels.

Table 5 quantifies the overhead introduced in the first-stage pseudo-label-generation network. CF-BiLSTM, including its fusion projection, increases the parameter count from 86.11 M to 89.65 M (4.11%) and the 384×384 forward computation from 110.99 to 115.06 GFLOPs (3.67%). Its peak training memory on a single GPU increases from 11.1 to 11.7 GB. The average time for generating a pseudo-label for one augmented VOC training image increases from 305.0 ms/image for the ViT baseline to 341.6 ms/image for Full PCCL. Because the contrastive objective is disabled during pseudo-label generation, this latency difference primarily reflects the inference-time overhead introduced by CF-BiLSTM. The latency increase (11.99%) is larger than the FLOP increase (3.67%), which is consistent with the lower parallel efficiency of recurrent operations compared with the highly optimized matrix operations in the ViT backbone, although implementation-level factors such as memory access and kernel-launch overhead may also contribute.

Table 5: Computational complexity and resource overhead of the first-stage PCCL network on PASCAL VOC 2012. Parameters and FLOPs are calculated for one 384×384 input; FLOPs count multiplication and addition as separate operations. Peak training memory is measured on a single NVIDIA RTX 4090 GPU. PCCL pseudo-label-generation latency is measured on the same single GPU with batch size 1 at 960×960 and averaged over the 10,582 images in the augmented VOC training set.

Configuration	Params (M)	FLOPs (G)	Peak Training Memory (GB)	Pseudo-Label-Generation Latency (ms/Image)
ViT baseline	86.11	110.99	11.1	305.0
+ CF-BiLSTM	89.65	115.06	11.7	341.6 $\leq$
Full PCCL	89.65	115.06	12.5	341.6

Note: $\leq$ Here, latency refers to the offline first-stage forward pass used to generate pseudo-labels for the augmented VOC training images, not to inference of the downstream DeepLabv2 model on the validation set. The CF-BiLSTM-only configuration and Full PCCL have the same pseudo-label-generation graph because the spatially weighted contrastive objective is used only during training. Full PCCL was timed directly; therefore, the CF-BiLSTM-only latency is reported from the measured Full PCCL profile rather than from a redundant additional timing run.

Collaborative Patch Contrastive Learning introduces no trainable parameters and is not executed during pseudo-label generation. Consequently, Full PCCL has the same parameter count, forward FLOPs, and pseudo-label-generation graph as the CF-BiLSTM configuration. Its additional cost occurs only during training: constructing positive–positive similarity, spatial-distance, spatial-weight, and positive–negative similarity matrices increases peak memory from 11.7 to 12.5 GB. For high-resolution inputs, the horizontal and vertical recurrent sequences become longer and may moderately reduce throughput. Nevertheless, their state computation grows linearly with the number of patches, whereas ViT self-attention grows quadratically; the measured 11.99% pseudo-label-generation latency increase at 960×960 therefore indicates a moderate overhead rather than a dominant high-resolution bottleneck.

In the final semantic segmentation stage, the refined pseudo-labels are used to train DeepLabv2 [24] in a fully supervised manner. This two-stage training protocol follows the standard WSSS pipeline, where the proposed PCCL framework focuses on generating high-quality pseudo-labels from image-level annotations, and DeepLabv2 is subsequently trained to produce the final dense segmentation results.

For downstream segmentation, we follow the public DeepLabv2 implementation at <https://github.com/kazuto1011/deeplab-pytorch>, using a ResNet-101 backbone initialized with ImageNet-pretrained weights from <https://github.com/johnnylu305/deeplab-imagenet-pytorch>. The training code is unchanged except that PCCL-generated pseudo masks are used as segmentation supervision. Following the public VOC configuration, the model is trained for 20,000 iterations using 321×321 random crops, an effective batch size of 10, and SGD with an initial learning rate of 2.5×10^{-4} , momentum of 0.9, weight decay of 5×10^{-4} , and polynomial learning-rate decay. Data augmentation includes random scaling, cropping, and horizontal flipping. During inference, predictions at scales $\{0.5, 0.75, 1.0\}$ are fused, and flip testing is not used. DeepLabv2 is employed only as a standard downstream segmentation model and is not part of the proposed methodological contribution.

The complexity profile in Table 5 concerns only the first-stage network used for training and offline pseudo-label generation. Neither CF-BiLSTM nor the contrastive objective is retained in the downstream DeepLabv2 model. Therefore, the architecture, parameter count, FLOPs, and deployment-time latency of DeepLabv2 remain unchanged relative to the same segmentation network trained with alternative pseudo masks.

4.3 Final Segmentation Performance

Table 6 shows limited variation across independent random seeds, indicating that PCCL is stable with respect to random initialization. This multi-seed analysis evaluates the repeatability of PCCL, while Tables 2–4 provide comparison with the published point estimates of existing methods.

Table 6: Statistical results of PCCL over three independent runs with different random seeds on the PASCAL VOC 2012 and MS COCO 2014 validation sets. The last column reports the mean and sample standard deviation ($n - 1$).

Metric	Run 1	Run 2	Run 3	Mean $\pm$ Std
VOC pseudo-label mIoU	71.5	72.0	71.6	71.70 ± 0.26
VOC final segmentation mIoU	70.7	71.3	71.3	71.10 ± 0.35
COCO mIoU	45.6	45.9	46.2	45.90 ± 0.30

As shown in Table 2, the proposed PCCL achieves competitive pseudo-label performance on the PASCAL VOC 2012 dataset. PCCL obtains a mean mIoU of 71.7%, which is numerically higher than the cited literature values listed in the table, including 70.8% for CPF-CTE [21].

The competitive pseudo-label performance is associated with the proposed patch collaboration mechanism. The Patch LSTM Context Fusion module enables bidirectional contextual information propagation among spatially correlated patches, which helps recover weakly activated object regions. In addition, the Collaborative Patch Contrastive Learning strategy enhances patch-level semantic consistency by pulling semantically related patches closer while pushing different semantic patches apart. These two components jointly improve the completeness and discriminative quality of activation maps, resulting in more accurate pseudo-label generation under image-level supervision.

Improvements in Final Segmentation Results. To further evaluate the effectiveness of the generated pseudo labels, we train DeepLabv2 using the masks produced by PCCL and compare the final segmentation performance with previous methods on the PASCAL VOC 2012 validation set. As reported in Table 3, PCCL obtains a mean mIoU of 71.1%. This is numerically higher than the literature values listed in the table, including 69.5% for CPF-CTE. We further include the recent CPEWS method published in CMC 2025, which

reports 69.8% mIoU on the VOC validation set. PCCL is numerically 1.3 percentage points higher than this recently published ViT-based method.

These results demonstrate that the pseudo labels generated by PCCL are not only more accurate at the activation-map level but also more effective for training the downstream segmentation network. The improvement can be attributed to the combination of contextual patch modeling and contrastive semantic regularization. The former strengthens information exchange between related patches, while the latter improves feature compactness and suppresses noisy background activations. As a result, PCCL can generate more complete object masks and achieve stronger segmentation performance under weak supervision.

Generalization on MS COCO 2014. We further evaluate PCCL on the more challenging MS COCO 2014 dataset. As shown in Table 4, PCCL obtains a mean mIoU of 45.9%, which is competitive with and numerically higher than the cited values included in the table. In particular, it is 0.5 percentage points higher than the reported CPF-CTE value. We further include CPEWS, a recent state-of-the-art WSSS method published in CMC 2025. CPEWS reports 41.4% mIoU on the COCO validation set, whereas PCCL obtains 45.9%, corresponding to an improvement of 4.5 percentage points. Together with the VOC result, this comparison further demonstrates the competitive performance and cross-dataset effectiveness of PCCL against a recent WSSS approach.

Since MS COCO contains more object categories, more complex object co-occurrences, and heavier background clutter than PASCAL VOC, this result provides complementary evidence that PCCL can be applied beyond the VOC setting.

Visualization of Segmentation Results. As shown in Fig. 6, we present qualitative visualization results to demonstrate the segmentation performance of the proposed method. It can be observed that our method generally produces more complete and coherent segmentation masks in many weakly activated regions. Nevertheless, fine boundaries and structurally complex scenes may remain challenging, as discussed below. Overall, these qualitative results indicate that PCCL can enhance semantic localization and generate more reliable segmentation results. Furthermore, as shown in Fig. 7, we report the class-wise performance of our method across different semantic categories, which provides a more detailed evaluation of its segmentation capability.

Failure Cases and Error Analysis. Fig. 8 presents representative failure cases involving small objects, thin structures, and occluded multi-object scenes. These examples illustrate several remaining limitations of PCCL. Small objects may occupy only a few patch tokens and therefore provide insufficient semantic evidence. Thin structures and fine boundaries can be lost because multiple foreground and background pixels are represented within the same patch. In heavily occluded or crowded scenes, disconnected object regions and nearby instances may also introduce ambiguous contextual relationships, resulting in incomplete masks, merged regions, or local semantic leakage.

These failure cases are closely related to the fixed granularity of patch-based representations. Although CF-BiLSTM propagates contextual information among neighboring and distant patches, it cannot fully recover spatial details that have already been lost during patch embedding. Similarly, patch-level contrastive learning may still be affected when visually ambiguous or spatially disconnected regions are assigned unreliable semantic relations. Therefore, contextual collaboration improves overall mask completeness but does not completely resolve errors caused by limited spatial resolution and complex object interactions.

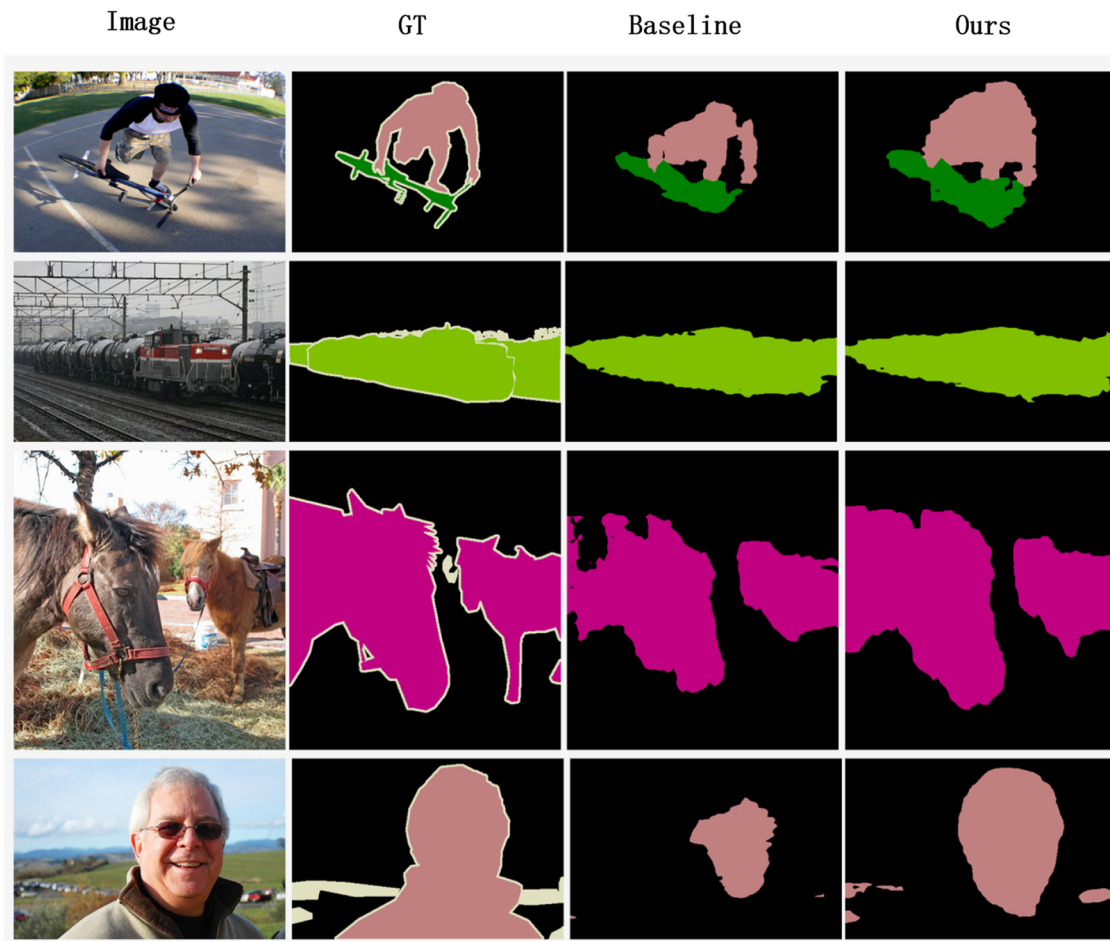


Figure 6: Visualization of segmentation results.

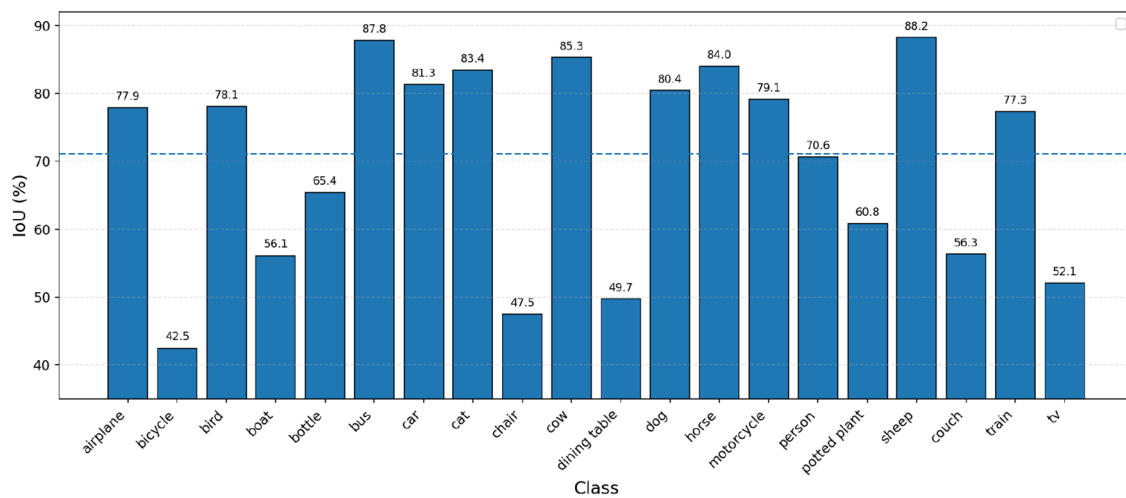


Figure 7: Segmentation performance on the validation set across different categories. The figure illustrates the per-class IoU scores.

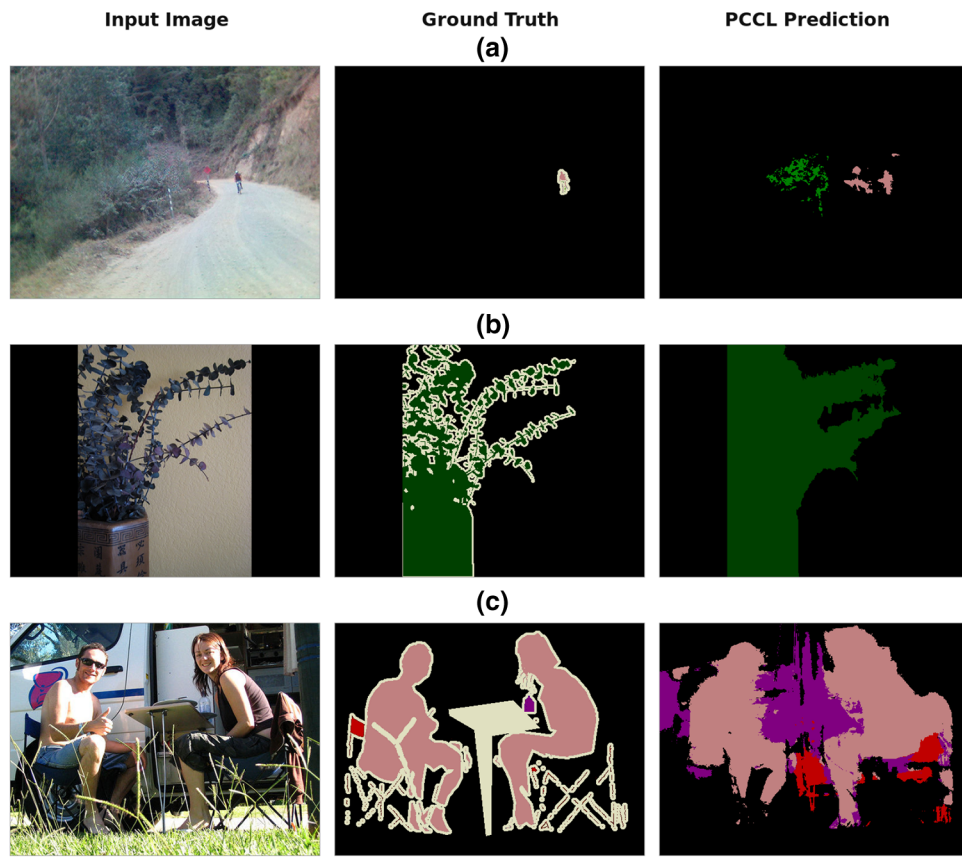


Figure 8: Representative PCCL failure cases. Each row shows the input image, ground truth, and PCCL prediction. The examples correspond to challenging scenes involving (a) small objects, (b) thin structures, and (c) occlusion and multiple adjacent objects.

Mechanistically, the fixed ViT-B/16 tokenizer represents each 16×16 image region with a single patch token, and CF-BiLSTM produces an updated representation on the same token grid without increasing its spatial resolution. A small object occupying only part of one patch must therefore compete with the surrounding background within a shared token, while a thin structure may cross several mixed foreground-background patches without any token explicitly representing its sub-patch boundary. Contextual propagation can strengthen semantic confidence when reliable foreground evidence exists in related patches, but it cannot subdivide a mixed token or create missing pixel-level and boundary evidence. When neighboring context is dominated by background or adjacent objects, propagation may also preserve or spread the dominant ambiguous response. Subsequent upsampling and CRF refinement operate on the resulting coarse prediction map and can sharpen existing responses, but they cannot reliably reconstruct an object or thin topology that is absent from the patch-level activations. CF-BiLSTM should therefore be understood as semantic context propagation rather than spatial super-resolution.

In future work, we will investigate multi-scale or adaptive patch representations for small objects, boundary-aware feature learning for thin structures, and occlusion-aware patch grouping for crowded scenes. Coarse-to-fine feature fusion and stronger local-detail modeling may further reduce these failure cases while retaining the semantic-spatial collaboration of PCCL.

4.4 Ablation Study

To verify the individual contribution and synergistic effect of each core component in our proposed method, we conduct a comprehensive ablation study on the PASCAL VOC 2012 dataset. Specifically, we evaluate the baseline framework alongside two key modules: Patch Collaboration and Patch Contrastive Learning. The quantitative results are summarized in Table 7.

Table 7: Two-factor ablation of Patch Collaboration and Patch Contrastive Learning on the PASCAL VOC 2012 validation set.

Original Framework	Patch Collaboration	Patch Contrastive Learning	mIoU (%)
✓			65.1
✓	✓		67.4
✓		✓	68.2
✓	✓	✓	71.1

The four configurations in Table 7 form a two-factor controlled ablation, allowing us to separate the individual effects of Patch Collaboration and Patch Contrastive Learning from their interaction. The subsequent experiments further isolate the spatial weighting term, the structure of contextual propagation, and the corresponding changes in patch-candidate quality.

Effectiveness of Patch Collaboration. As indicated in the second row of Table 7, incorporating the Patch Collaboration module into the original framework yields a noticeable performance enhancement, boosting the mIoU from 65.1% to 67.4% (+2.3 percentage points). This steady improvement demonstrates that enabling cooperative interactions among patches helps the model aggregate local context more effectively, thereby generating more coherent and consistent initial seed regions for semantic segmentation.

Effectiveness of Patch Contrastive Learning. Similarly, we investigate the impact of the Patch Contrastive Learning strategy. As shown in the third row, applying this contrastive constraint independently brings a substantial performance gain of 3.1 percentage points, elevating the mIoU to 68.2%. This result highlights the significance of pulling semantically similar patch features closer while pushing dissimilar or background patches apart, which drastically reduces false-positive activations in ambiguous regions.

Effectiveness of the Spatial Weighting Term. The most direct controlled alternative to the proposed spatial term is uniform positive-pair weighting. We therefore replace $\omega_{i,j}$ in the positive-pair term with $\omega_{i,j} = 1$ while keeping patch selection, pair construction, CF-BiLSTM features, the positive-negative term, and all other training settings unchanged. As shown in Table 8, uniform weighting achieves 70.6% mIoU, whereas the proposed spatial weighting achieves 71.1%. Because this comparison changes only $\omega_{i,j}$, the 0.5-point difference isolates the contribution of patch-grid-aware weighting beyond non-weighted patch contrast.

The performance contributions should therefore be interpreted hierarchically. In the two-factor ablation, Patch Collaboration alone improves the baseline by 2.3 points (65.1% to 67.4%), while Patch Contrastive Learning alone improves it by 3.1 points (65.1% to 68.2%). Combining them yields a 6.0-point total improvement (65.1% to 71.1%), including a positive interaction of 0.6 points. Within this complete framework, the uniform-vs-spatial comparison further attributes 0.5 points specifically to the spatial weighting term. Thus, the principal improvement arises from contextual fusion and patch contrastive regularization together, whereas spatial weighting provides a modest but controlled geometry-specific refinement. It introduces no trainable parameters and is absent during pseudo-label generation and downstream inference.

Table 8: Ablation study of the spatial weighting term on the PASCAL VOC 2012 validation set.

Contrastive Learning Configuration	mIoU (%)
PCCL with uniform positive-pair weighting, $\omega_{i,j} = 1$	70.6
PCCL with the spatial weighting term	71.1

Effect of CRF Post-processing. To quantify the contribution of CRF independently from the proposed representation-learning components, we compare PCCL with and without CRF while keeping the trained PCCL model, pre-CRF prediction maps, and downstream segmentation protocol otherwise unchanged. As shown in Table 9, the final segmentation mIoU increases from 68.3% without CRF to 71.1% with CRF, corresponding to a gain of 2.8 percentage points. This result quantifies the boundary-refinement contribution of the conventional CRF post-processing step. The CRF configuration is fixed to that of ViT-PCM [20] and is not treated as part of the methodological novelty of PCCL.

Table 9: Effect of CRF post-processing on final segmentation performance on the PASCAL VOC 2012 validation set.

Configuration	mIoU (%)
PCCL without CRF	68.3
PCCL with CRF	71.1

This ablation quantifies and separates the effect of conventional CRF post-processing from the proposed representation-learning components. Accordingly, the 68.3% result without CRF represents the performance before boundary refinement, while the additional 2.8 percentage points are attributed separately to CRF. CRF is not treated as part of the proposed PCCL contribution.

Comparison with Alternative Contextual Modeling Modules. To further investigate the effectiveness of the proposed bidirectional contextual modeling strategy, we compare the Patch LSTM Context Fusion module with simpler contextual alternatives under the same training and evaluation settings. As shown in Table 10, the ViT-only baseline achieves an mIoU of 65.1%. Introducing a single-axis BiLSTM improves the performance to 66.3%, demonstrating that directional contextual propagation is beneficial for enhancing patch representations. The proposed horizontal-and-vertical BiLSTM further increases the mIoU to 67.4%, outperforming the single-axis variant by 1.1 percentage points. These results indicate that explicitly modeling contextual dependencies along both spatial dimensions provides more comprehensive information propagation and is more effective for recovering weakly activated and incomplete object regions.

Table 10: Comparison with alternative contextual modeling modules on the PASCAL VOC 2012 validation set.

Contextual Modeling Module	mIoU (%)
ViT-only baseline	65.1
Single-axis BiLSTM	66.3
Horizontal and vertical BiLSTM	67.4

Synergistic Effect of Both Components. Finally, when both the Patch Collaboration and Patch Contrastive modules are simultaneously integrated into our framework (the fourth row), our model achieves the peak performance of 71.1% mIoU. This cumulative improvement (+6.0 percentage points over the baseline)

indicates that the two components are complementary. Patch Collaboration promotes broader and more coherent contextual aggregation, while Patch Contrastive Learning improves feature discrimination and semantic consistency.

More specifically, the factorial interaction between the two components is $\Delta_{\text{int}} = 71.1 - 67.4 - 68.2 + 65.1 = 0.6$ percentage points. Equivalently, Patch Collaboration improves mIoU by 2.3 points without contrastive learning ($65.1 \rightarrow 67.4$), but by 2.9 points when contrastive learning is enabled ($68.2 \rightarrow 71.1$). This positive interaction supports the claimed semantic-spatial collaboration mechanism: contrastive regularization benefits from operating on context-enhanced patch embeddings, beyond the separate main effects of the two components. Together with the uniform-vs-spatial weighting comparison in Table 8 and the patch-quality diagnostics in Table 11, the revised analysis isolates the mechanism at the component, objective, and intermediate-candidate levels.

Table 11: Patch candidate quality and residual confirmation-bias statistics on the PASCAL VOC 2012 validation set. The upward and downward arrows indicate that higher and lower values are preferable, respectively.

Configuration	ϵ	Positive Precision $\uparrow$	Positive Coverage $\uparrow$	Negative Error Rate $\downarrow$
PCCL w/o CF-BiLSTM	0.8	84.8%	27.7%	6.7%
PCCL w/o image-label constraint	0.8	82.4%	35.7%	4.9%
Full PCCL	0.7	86.0%	46.1%	6.6%
Full PCCL	0.8	90.9%	37.5%	4.2%
Full PCCL	0.9	94.8%	21.3%	3.0%

Sensitivity Analysis of Contrastive Learning. To further investigate the influence of contrastive learning, we report a post-hoc sensitivity analysis on the PASCAL VOC 2012 validation set for two key hyperparameters: the contrastive loss weight λ_{con} and the confidence threshold ϵ . The final values $\lambda_{\text{con}} = 0.01$ and $\epsilon = 0.8$ had already been selected through the training-set grid search and were not changed based on this validation-set analysis. As illustrated in Fig. 5, this configuration achieves 71.1% mIoU on the VOC validation set. When λ_{con} is too small, the contrastive constraint becomes insufficient to effectively enhance patch-level semantic discrimination. Conversely, an excessively large λ_{con} may overemphasize feature separation and disturb the image-level classification objective, leading to degraded segmentation performance. Similarly, a lower confidence threshold may introduce unreliable patch candidates into contrastive learning, while a higher threshold may discard useful foreground patches. These results illustrate the sensitivity of PCCL to contrastive regularization strength and reliable patch selection, rather than serving as the basis for selecting the final configuration.

Patch Candidate Quality and Confirmation-Bias Analysis. To directly assess the reliability of prediction-driven sampling, we conduct a post-hoc diagnostic analysis on the PASCAL VOC 2012 validation set. The validation-set pixel-level annotations are used only to evaluate the mined candidates: they are not used to train or retrain the model, construct candidate sets, or alter the previously selected threshold. Each patch is assigned a reference class from the majority of its non-void ground-truth pixels. Positive precision is the fraction of selected positive candidates whose reference class matches the target class; positive coverage is the fraction of ground-truth target-class patches that are correctly selected as positives; and negative error rate is the fraction of selected negative candidates whose reference class is actually the target class.

At the fixed threshold $\epsilon = 0.8$, adding CF-BiLSTM increases positive precision from 84.8% to 90.9% and positive coverage from 27.7% to 37.5%, while reducing the negative error rate from 6.7% to 4.2%. These

changes indicate that bidirectional contextual propagation both suppresses isolated false positive responses and recovers additional true object patches before contrastive pairs are formed. The image-level constraint produces a complementary effect. Relative to the model without this constraint, Full PCCL improves positive precision from 82.4% to 90.9%, positive coverage from 35.7% to 37.5%, and negative error from 4.9% to 4.2%. The dominant gain in positive precision is consistent with the positive-candidate definition, in which the image-level presence label directly prevents absent categories from contributing positive candidates. The smaller changes in coverage and negative error arise indirectly because this gate changes which class-specific contrastive terms participate during training, although the negative threshold itself remains unchanged.

The threshold analysis further shows the expected reliability–coverage trade-off. Increasing ε from 0.7 to 0.9 raises positive precision from 86.0% to 94.8% and reduces negative error from 6.6% to 3.0%, but decreases positive coverage from 46.1% to 21.3%. The adopted $\varepsilon = 0.8$ provides an intermediate operating point, retaining 37.5% coverage while achieving 90.9% positive precision and a 4.2% negative error rate. These results provide direct evidence that contextual enhancement, image-level category filtering, and confidence-margin selection reduce the amount of erroneous self-supervision. Nevertheless, the remaining nonzero negative error confirms that confirmation bias is alleviated rather than fully removed.

At the adopted operating point $\varepsilon = 0.8$, these results quantify the residual severity of confirmation bias more directly: 9.1% of the selected positive candidates disagree with the reference class, and 4.2% of the selected negative candidates actually belong to the target class. Positive coverage is 37.5%, meaning that 62.5% of the ground-truth target-class patches are not admitted to the positive set. This last value should not be interpreted as a 62.5% false-negative rate, because patches within the confidence margin are ignored rather than automatically selected as negatives. Overall, the candidate sets retain measurable label noise, but the high positive precision and low negative error show that the adopted mining strategy prioritizes candidate reliability at the cost of conservative coverage; confirmation bias is reduced but not eliminated.

5 Conclusion

In this paper, we propose Patch Collaboration with Contrastive Learning (PCCL) for weakly supervised semantic segmentation using only image-level labels. Rather than targeting a specific application domain, PCCL focuses on the fundamental challenge of improving semantic localization and pseudo-label generation under weak supervision. PCCL introduces a geometry-aware semantic–spatial collaboration mechanism that applies distance-dependent positive-pair contrast to recurrently context-enhanced embeddings. The formulation gives stronger compactness constraints to nearby positive patches and weaker constraints to distant pairs, while a separate positive–negative term maintains semantic discrimination. This explicit use of patch-grid geometry distinguishes PCCL from related non-weighted patch contrast designs and enables contextual propagation and semantic regularization to be optimized in a coordinated manner.

Extensive experiments on PASCAL VOC 2012 and MS COCO 2014 demonstrate the effectiveness of the proposed method. PCCL achieves competitive pseudo-label quality and final segmentation performance compared with existing WSSS approaches. In particular, our method obtains 71.7% pseudo-label mIoU and 71.1% final segmentation mIoU on PASCAL VOC 2012, as well as 45.9% mIoU on MS COCO 2014. The ablation study further verifies that Patch Collaboration and Patch Contrastive Learning are complementary, and their combination leads to the best performance. These results confirm that jointly exploiting contextual collaboration and contrastive semantic consistency is effective for generating more complete and reliable segmentation masks under weak supervision.

A limitation of the current study is that downstream segmentation performance is evaluated only with DeepLabv2. PCCL produces standard pixel-wise class masks and does not encode DeepLabv2-specific features or parameters; its pseudo-label output is therefore structurally compatible with other

fully supervised segmentation architectures. In addition, the reported pseudo-label mIoU is measured before downstream segmentation training and directly evaluates the generated supervision independently of DeepLabv2. However, structural compatibility and intrinsic pseudo-label quality do not by themselves establish performance generalization across segmentation architectures. A controlled cross-architecture study would require complete architecture-specific retraining, matched baseline and PCCL pseudo-label runs, appropriate optimization protocols, and repeated trials. Such an evaluation was beyond the experimental scope and computational budget of the current study. We therefore retain DeepLabv2 as a fixed standard downstream protocol to isolate pseudo-label quality and explicitly limit our empirical conclusions to this setting, rather than reporting an isolated under-tuned pilot result.

In future work, we will further investigate more adaptive patch relationship modeling strategies and explore how to extend PCCL to larger-scale vision-language segmentation scenarios. We also plan to improve the efficiency of patch collaboration modules for high-resolution images while maintaining strong segmentation accuracy. Furthermore, we will evaluate PCCL on more specialized application domains, such as medical image segmentation and remote sensing scenarios, to investigate its generalization ability under different visual characteristics and domain-specific challenges. We will also conduct controlled cross-architecture experiments with stronger downstream segmentation models, including DeepLabv3+ and Mask2Former, to determine whether the gains from PCCL-generated pseudo-labels remain consistent across different segmentation architectures.

Acknowledgement: None.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Zhenhong Chen and Wangyu Wu; methodology, Zhenhong Chen and Wangyu Wu; software, Zhenhong Chen; validation, Zhenhong Chen and Wangyu Wu; investigation, Zhenhong Chen; data curation, Zhenhong Chen; formal analysis, Zhenhong Chen and Wangyu Wu; visualization, Zhenhong Chen; writing—original draft preparation, Zhenhong Chen; writing—review and editing, Zhenhong Chen and Wangyu Wu; supervision, Wangyu Wu; project administration, Wangyu Wu. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study are publicly available. The PASCAL VOC 2012 dataset is available at <https://www.robots.ox.ac.uk/~vgg/projects/pascal/VOC/voc2012/>, and the MS COCO 2014 dataset is available at <https://cocodataset.org/#download>. An anonymized repository has been created at <https://anonymous.4open.science/r/PCCL-D632>. The source code, configuration files, and reproducibility instructions are being prepared and are expected to be made available through this repository following publication.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A Implementation Details of the SpatiallyWeighted Contrastive Loss

Algorithm A1 gives implementation-oriented pseudocode corresponding to Eqs. (14)–(17). The inputs are the context-enhanced patch embeddings, patch-level class probabilities, image-level foreground labels, the confidence threshold ε , and the spatial decay parameter τ_s . Background is excluded from the class loop.

Algorithm A1: Implementation of the spatially weighted collaborative patch contrastive loss**Require:** Embeddings $\mathbf{F} \in \mathbb{R}^{B \times H \times W \times d}$, probabilities $\mathbf{Z} \in \mathbb{R}^{B \times H \times W \times C}$, image labels $\mathbf{y}$, ε , and τ_s **Ensure:** Batch contrastive loss $\mathcal{L}_{\text{con}}$

```

1: Flatten the  $H \times W$  grid and L2-normalize every embedding.
2: Construct row-column coordinates  $\mathbf{r}_i$  for all patch indices  $i$ .
3: Set  $\mathcal{L}_{\text{batch}} \leftarrow 0$ .
4: for each image  $b = 1, \dots, B$  do
5:   Set  $\mathcal{L}_b \leftarrow 0$  and  $n_b \leftarrow 0$ .
6:   for each foreground class  $c = 1, \dots, C - 1$  do
7:      $\mathcal{P}_{b,c}^+ \leftarrow \{i \mid y_{b,c} = 1 \wedge Z_{b,i,c} > \varepsilon\}$ .
8:      $\mathcal{P}_{b,c}^- \leftarrow \{k \mid Z_{b,k,c} < 1 - \varepsilon\}$ .
9:     if  $|\mathcal{P}_{b,c}^+| < 2$  or  $|\mathcal{P}_{b,c}^-| < 1$  then
10:      Skip class  $c$ .
11:   else
12:     For each unordered pair  $i < j$  in  $\mathcal{P}_{b,c}^+$ , compute  $\omega_{i,j} = \exp(-\|\mathbf{r}_i - \mathbf{r}_j\|_2^2 / \tau_s)$ .
13:      $M_{b,c}^+ \leftarrow |\mathcal{P}_{b,c}^+|(|\mathcal{P}_{b,c}^+| - 1)/2$ .
14:      $\mathcal{L}_{b,c}^+ \leftarrow \frac{1}{M_{b,c}^+} \sum_{i < j} \omega_{i,j} (1 - \hat{\mathbf{f}}_{b,i}^\top \hat{\mathbf{f}}_{b,j})$ .
15:      $\mathcal{L}_{b,c}^- \leftarrow \frac{1}{|\mathcal{P}_{b,c}^+| |\mathcal{P}_{b,c}^-|} \sum_{i \in \mathcal{P}_{b,c}^+} \sum_{k \in \mathcal{P}_{b,c}^-} \hat{\mathbf{f}}_{b,i}^\top \hat{\mathbf{f}}_{b,k}$ .
16:      $\mathcal{L}_b \leftarrow \mathcal{L}_b + \mathcal{L}_{b,c}^+ + \mathcal{L}_{b,c}^-$ ;  $n_b \leftarrow n_b + 1$ .
17:   end if
18: end for
19: if  $n_b > 0$  then
20:    $\mathcal{L}_b \leftarrow \mathcal{L}_b / n_b$ .
21: else
22:    $\mathcal{L}_b \leftarrow 0$ .
23: end if
24:  $\mathcal{L}_{\text{batch}} \leftarrow \mathcal{L}_{\text{batch}} + \mathcal{L}_b$ .
25: end for
26: Return  $\mathcal{L}_{\text{con}} \leftarrow \mathcal{L}_{\text{batch}} / B$ .

```

The implementation uses only the upper-triangular positive-positive similarity entries and excludes the diagonal, so each unordered positive pair contributes once. Because both cosine similarity and $\omega_{i,j}$ are symmetric, averaging the unique pairs $i < j$ is algebraically equivalent to averaging the two ordered entries (i, j) and (j, i) in Eq. (14). The positive term is normalized by the number of valid pairs rather than by the sum of spatial weights, and the positive-negative term averages all corresponding similarities. Losses are then averaged over valid foreground classes within each image and over images in the batch. If an image has no valid foreground class, its contrastive contribution is set to zero. In all reported experiments, $\varepsilon = 0.8$ and $\tau_s = 1.0$. The returned $\mathcal{L}_{\text{con}}$ is combined with $\mathcal{L}_{\text{cls}}$ using $\lambda_{\text{con}} = 0.01$ in the final objective in Section 3.4. The contrastive loss creates no trainable variables and is evaluated only during first-stage training.

References

1. Kolesnikov A, Lampert CH. Seed, expand and constrain: three principles for weakly-supervised image segmentation. Eur Conf Comput Vis. 2016;9908(7):695–711. doi:10.1007/978-3-319-46493-0_42.

2. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA.
3. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Semantic image segmentation with deep convolutional nets and fully connected CRFs. arXiv:1412.7062. 2014.
4. Ahn J, Kwak S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–22; Salt Lake City, UT, USA. p. 4981–90.
5. Lee J, Kim E, Mok J, Yoon S. Anti-adversarially manipulated attributions for weakly supervised semantic segmentation and object localization. *IEEE Trans Pattern Anal Mach Intell.* 2024;46(3):1618–34. doi:10.1109/tpami.2022.3166916.
6. Wu W, Dai T, Huang X, Ma F, Xiao J. Image augmentation with controlled diffusion for weakly-supervised semantic segmentation. In: Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2024 Apr 14–19; Seoul, Republic of Korea. p. 6175–9.
7. Yao Y, Chen T, Xie GS, Zhang C, Shen F, Wu Q, et al. Non-salient region object mining for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 2623–32.
8. Lee S, Lee M, Lee J, Shim H. Railroad is not a train: saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 19–25; Virtual. p. 5495–505.
9. Chang YT, Wang Q, Hung WC, Piramuthu R, Tsai YH, Yang MH. Weakly-supervised semantic segmentation via sub-category exploration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 8991–9000.
10. Chen Q, Yang L, Lai JH, Xie X. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 4288–98.
11. Du Y, Fu Z, Liu Q, Wang Y. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 4320–9.
12. Zhang F, Gu C, Zhang C, Dai Y. Complementary patch for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 10–17; Montreal, QC, Canada. p. 7242–51.
13. Wang Y, Zhang J, Kan M, Shan S, Chen X. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 12275–84.
14. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16×16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
15. Gao W, Wan F, Pan X, Peng Z, Tian Q, Han Z, et al. TS-CAM: token semantic coupled attention map for weakly supervised object localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 10–17; Montreal, QC, Canada. p. 2886–95.
16. Xu L, Ouyang W, Bennamoun M, Boussaid F, Xu D. Multi-class token transformer for weakly supervised semantic segmentation. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA.
17. Xu L, Bennamoun M, Boussaid F, Laga H, Ouyang W, Xu D. Mctformer+: multi-class token transformer for weakly supervised semantic segmentation. *IEEE Trans Pattern Anal Mach Intell.* 2024;46(12):8380–95. doi:10.1109/cvpr52688.2022.00427.
18. Ru L, Zhan Y, Yu B, Du B. 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; 2022 Jun 18–24; New Orleans, LA, USA. p. 16846–55.

19. Ru L, Zheng H, Zhan Y, Du B. Token contrast for weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–24; Vancouver, BC, Canada. p. 3093–102.
20. Rossetti S, Zappia D, Sanzari M, Schaerf M, Pirri F. Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation. In: European Conference on Computer Vision. Berlin/Heidelberg, Germany: Springer; 2022. p. 446–63.
21. Fu Y, Li H, Wu W. Context patch fusion with class token enhancement for weakly supervised semantic segmentation. *Comput Model Eng Sci.* 2026;146(1):1–10. doi:10.32604/cmcs.2025.074467.
22. Wu W, Dai T, Huang X, Ma F, Xiao J. Top-k pooling with patch contrastive learning for weakly-supervised semantic segmentation. In: Proceedings of the 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2024 Oct 6–10; Kuching, Malaysia. p. 5270–5.
23. Wu W, Dai T, Chen Z, Huang X, Xiao J, Ma F, et al. Adaptive patch contrast for weakly supervised semantic segmentation. *Eng Appl Artif Intell.* 2025;139(4):109626. doi:10.1016/j.engappai.2024.109626.
24. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans Pattern Anal Mach Intell.* 2017;40(4):834–48. doi:10.1109/tpami.2017.2699184.
25. Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with Gaussian edge potentials. *Adv Neural Inf Process Syst.* 2011;24.
26. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A. The pascal visual object classes (VOC) challenge. *Int J Comput Vis.* 2010;88(2):303–38. doi:10.1007/s11263-009-0275-4.
27. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: common objects in context. *Eur Conf Comput Vis.* 2014;8693(2):740–55. doi:10.1007/978-3-319-10602-1_48.
28. Hariharan B, Arbeláez P, Bourdev L, Maji S, Malik J. Semantic contours from inverse detectors. In: Proceedings of the IEEE International Conference on Computer Vision; 2011 Nov 6–13; Barcelona, Spain. p. 991–8.
29. Su Y, Sun R, Lin G, Wu Q. Context decoupling augmentation for weakly supervised semantic segmentation. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 7004–14.
30. Xu L, Ouyang W, Bennamoun M, Boussaid F, Sohel F, Xu D. Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 6984–93.
31. Chen L, Lei C, Li R, Li S, Zhang Z, Zhang L. FPR: false positive rectification for weakly supervised semantic segmentation. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. pp. 1108–18.
32. Hao X, Jiang X, Ni W, Tan W, Yan B. Prompt-guided semantic-aware distillation for weakly supervised incremental semantic segmentation. *IEEE Trans Circuits Syst Video Technol.* 2024;34(11):10632–45. doi:10.1109/tcsvt.2024.3412996.
33. Zhang B, Gao X, Yu S, Liu W. Enhanced online CAM: single-stage weakly supervised semantic segmentation via collaborative guidance. *Pattern Recognit.* 2024;156(4):110787. doi:10.1016/j.patcog.2024.110787.
34. Xu R, Wang C, Sun J, Xu S, Meng W, Zhang X. Self correspondence distillation for end-to-end weakly-supervised semantic segmentation. *AAAI Conf Artif Intell.* 2023;37(3):3045–53. doi:10.1609/aaai.v37i3.25408.
35. Shao X, Han J, Li L, Zhao X, Yan J. CPEWS: contextual prototype-based end-to-end weakly supervised semantic segmentation. *Comput Mater Contin.* 2025;83(1):595–617. doi:10.32604/cmc.2025.060295.
36. Kim S, Park D, Shim B. Semantic-aware superpixel for weakly supervised semantic segmentation. *AAAI Conf Artif Intell.* 2023;37(1):1142–50. doi:10.1609/aaai.v37i1.25196.
37. Zhao X, Tang F, Wang X, Xiao J. SFC: shared feature calibration in weakly supervised semantic segmentation. *AAAI Conf Artif Intell.* 2024;38(7):7525–33. doi:10.1609/aaai.v38i7.28584.
38. Lin M, Li G, Xu S, Hao Y, Zhang S. Decoupling foreground and background with Siamese ViT networks for weakly-supervised semantic segmentation. *Neurocomputing.* 2024;610(11):128540. doi:10.1016/j.neucom.2024.128540.