Integrating Low-Level Visual Cues for Enhanced Unsupervised Semantic Segmentation

Yuhao Qing[1], Dan Zeng[2], Shaorong Xie[1], Kaer Huang[3], Yueying Wang[1]*

[1]School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China
[2]School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China
[3]Lenovo, Building 1, No.10 Xibeiwang East Road, Haidian District, Beijing, 100085, China
qingyuhao@shu.edu.cn, dzeng@shu.edu.cn, srxie@shu.edu.cn, huangke1@lenovo.com, wyy676@126.com

Abstract

Unsupervised semantic segmentation algorithms aim to identify meaningful semantic groups without annotations. Recent approaches leveraging self-supervised transformers as pre-training backbones have successfully obtained high-level dense features that effectively express semantic coherence. However, these methods often overlook local semantic coherence and low-level features such as color and texture. We propose integrating low-level visual cues to complement high-level visual cues derived from self-supervised pre-training branches. Our findings indicate that low-level visual cues provide a more coherent recognition of color-texture aspects, ensuring the continuity of spatial structures within classes. This insight led us to develop IL2Vseg, an unsupervised semantic segmentation method that leverages the complementation of low-level visual cues. The core of IL2Vseg is a spatially-constrained fuzzy clustering algorithm based on color affinities, which preserves the intra-class affinity of spatially-adjacent and similarly-colored pixels in low-level visual cues. Additionally, to effectively couple low-level and high-level visual cues, we introduce a feature similarity loss function to optimize the feature representation of fused visual cues. To further enhance consistent feature learning, we incorporate contrast loss functions based on color invariance and luminosity invariance, which improve the learning of features from different semantic categories. Extensive experiments on multiple datasets, including COCO-Stuff-27, Cityscapes, Potsdam, and MaSTr1325, demonstrate that IL2Vseg achieves state-of-the-art results.

Introduction

Semantic segmentation is a crucial task in computer vision, aiming to segment an image into different regions where each pixel is assigned a specific semantic label. Due to its pixel-level image segmentation, it has a wide range of applications in automated driving, medical imaging, agricultural monitoring, environmental mapping, and other fields.

Existing supervised methods have achieved significant results, but they rely heavily on a large number of annotated masks. Obtaining annotated data is both time-consuming and expensive, as pixel-level annotation of images requires substantial manpower and expertise. This reliance restricts the scalability and applicability of supervised methods, especially when annotated data is scarce or unavailable. Therefore, exploring unsupervised semantic segmentation methods offers a potential solution to these challenges.

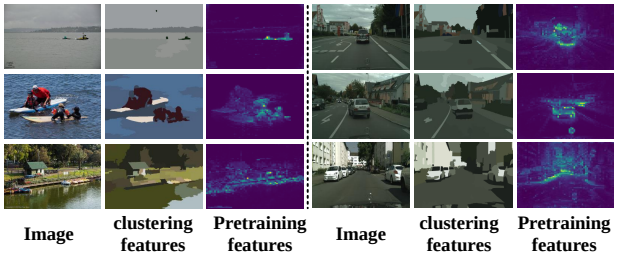

Figure 1: The visualization results of the high-level visual cues (clustered features) and low-level visual cues (pretrained features)

Traditional clustering-based methods, such as k-means(Na, Xumin, and Yong 2010) and Gaussian Mixture Models (GMM)(Reynolds et al. 2009), are relatively simple and easy to implement. Their principles are well understood, and the results can be interpreted. However, these traditional methods rely on manually created features, are sensitive to initial conditions and hyperparameters, and may fail to capture complex high-level semantic information in images, leading to suboptimal segmentation results. Additionally, clustering-based methods typically process pixels independently, with poor consideration of spatial relationships and contextual information, resulting in fragmented and incoherent segmentation outcomes.

Recently, unsupervised semantic segmentation methods based on self-supervised pre-trained visual backbones have garnered significant attention. TransFGU(Yin et al. 2022) was the first to acquire rich information about high-level structured semantic concepts from large-scale visual data in a self-supervised learning manner and use this information as a priori to discover potential semantic categories in the target dataset. STEGO(Hamilton et al. 2022) decomposed the problem into learning the representation and learning the segmentation header, using patch-level feature representations learned from a self-supervised pre-trained model, and made substantial progress in unsupervised segmentation results. Additionally, HP, Smooseg, and EAGLE(Kim et al.

*Corresponding author.

2024) further advanced the field by enhancing the semantic cues of patch-level features and learning object-level representations with smoothness priors and object feature consistency.

However, these methods perform semantic and structural analyses based on high-level feature information from self-supervised pre-trained models. Despite various efforts to enhance semantic context consistency, the saliency of basic constructive information, such as edges and textures, is significantly reduced after integration with multilayered network processing. This reduction leads to inconsistencies in segmentation results in local regions, and complex objects may be incorrectly segmented into multiple labels. Therefore, low-level structural features and local feature information are equally important cues for segmentation.

Taking these factors into account, we build on previous works(Hamilton et al. 2022),(Lan et al. 2024) by employing a fuzzy clustering branch based on color affinity and spatial constraints to ensure the underlying constructiveness and local consistency of the image. Specifically, our approach involves two steps: (1) performing fuzzy clustering with spatial constraints to capture the local consistency of neighboring pixel points and avoid noisy and isolated pixels, and (2) computing Euclidean distances of neighboring pixels to obtain the color affinity matrix using the Gaussian function to complement the results of the first step. The obtained low-level visual cues and high-level visual cues are then coupled and optimized using a feature similarity function, which provides the basis for obtaining a continuous and accurate semantic feature map. The visualization results of the high-level visual cues (clustered features) and low-level visual cues (pre-trained features) are shown in Figure 1. Additionally, we ensure that the object features within and between images remain consistent by introducing contrast loss in different branches.

Specifically, we make the following contributions:

- We propose using low-level visual cues to complement high-level pre-trained features to obtain more accurate and continuous intra-class relationships in images.
- We introduce a feature similarity loss function to further optimize the feature representation of fused visual cues and enhance the clustering effect of similar regions in images.
- Through extensive experimental validation on multiple datasets, our method achieves state-of-the-art performance.

Related Work

Clustered Image Segmentation. Learning meaningful visual features without human annotation is a long-term goal of computer vision. Clustering algorithms, such as fuzzy c-means (FCM), are widely used in image processing due to their simplicity and efficiency. However, classical FCM is sensitive to noise and brightness, posing challenges for complex image segmentation. To address these issues, FCM with spatial distance constraints has been proposed, enhancing segmentation by incorporating local spatial distances into the objective function.

Several methods have been developed to improve FCM's performance in image segmentation(Zhang et al. 2018). Krinidis et al.(Krinidis and Chatzis 2010) introduced fuzzy local spatial and gray level similarity metrics to mitigate noise sensitivity and preserve image details. Additionally, Zhang et al.(Zhang et al. 2017) tackled homogeneity segmentation and edge blurring in remote sensing images by introducing a fuzzy local similarity measure. Tang et al.(Tang, Ren, and Pedrycz 2020) combined weighted and structural similarity metrics with luminance dependency to overcome limitations in traditional FCM algorithms. Despite these advancements, these methods still struggle to extract high-level abstract features and capture effective semantic information in complex data.

Unsupervised Segmentation via Self-Supervised Pre-training. Recent approaches leverage self-supervised feature learning for unsupervised semantic segmentation(Cho et al. 2021; Yin et al. 2022). IIC(Ji and Vedaldi 2019) maximizes mutual information between related pairs for unsupervised clustering, while InfoSeg(Harb and Knöbelreiter 2021) enhances segmentation accuracy through a two-step learning process. STEGO(Hamilton et al. 2022) improves feature compactness and semantic consistency by separating feature learning from clustering and introducing a contrast loss function. SegSort(Hwang et al. 2019) maximizes intra-category similarity and minimizes inter-category similarity. Deng et al.(Deng and Luo 2023) propose a neural network-based spectral clustering method to enhance spectral decomposition efficiency and flexibility. Melas et al.(Melas-Kyriazi et al. 2022) address complex scenes using a self-supervised network for graph partitioning. HP(Seong et al. 2023) ensures local semantic consistency and enhances semantic correlation through contrast learning. SmooSeg(Lan et al. 2024) transforms segmentation into an energy minimization problem using a self-supervised approach. EAGLE(Kim et al. 2024) introduces the EiCue spectral technique and contrast loss to improve object-level semantic encoding in visual Transformers. In contrast, our proposed IL2Vseg complements high-level visual features with low-level visual cues to form continuous, accurate, and compact clusters, utilizing existing self-supervised pre-trained models.

Self-supervised learning. Self-supervised learning (SSL) is a paradigm that leverages large amounts of unlabeled data to learn useful representations, particularly valuable when labeled data is scarce. SSL typically involves constructing positive and negative sample pairs to maximize similarity within positive pairs and minimize it within negative pairs. MoCo(He et al. 2020) introduces a momentum encoder and a dynamic negative sample queue to enhance training efficiency. Clustering-based methods like SwAV(Caron et al. 2020) and DeepCluster(Caron et al. 2018) enhance similarity between samples of the same class by performing online clustering and generating pseudo-labels, respectively. Graph-based methods such as GraphCL(Hafidi et al. 2020) and MVGRL(Hassani and Khasahmadi 2020) combine contrast learning with graph neural networks to learn global and local graph representations. DINO(Caron et al. 2021) employs a momentum updating mechanism and a contrast

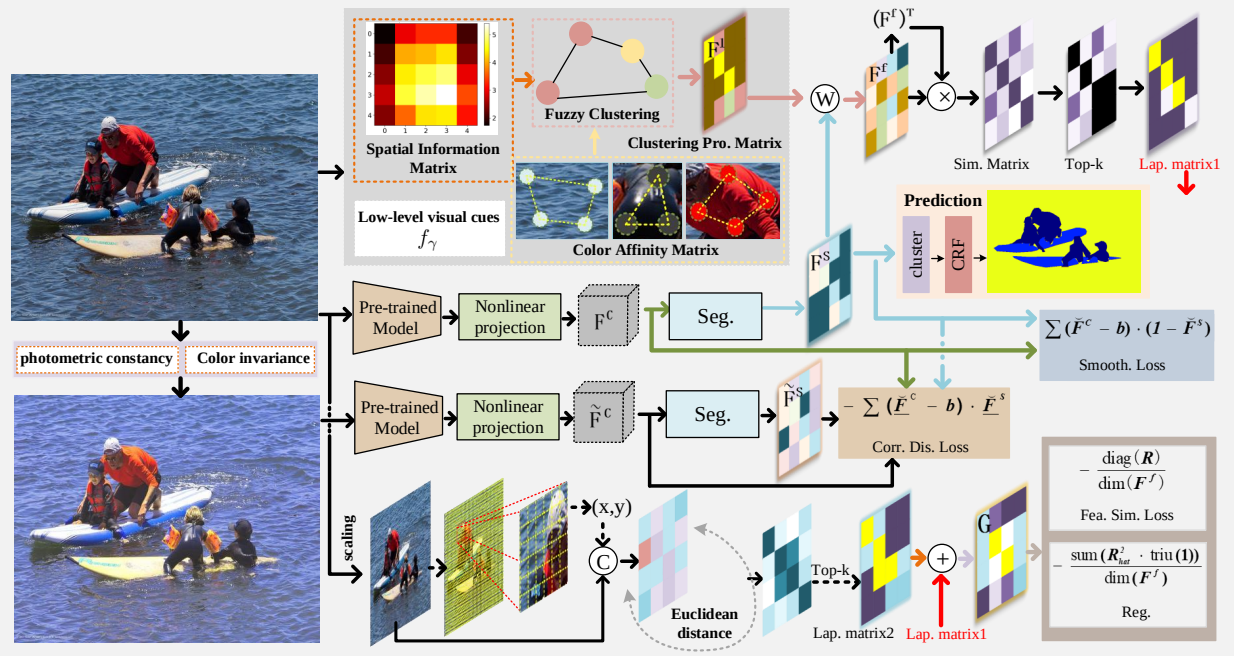

Figure 2: Overview of the proposed IL2Vseg. IL2Vseg has two components, high-level pre-trained features and low-level visual cues, where the low-level visual cues (f_γ) consists of spatially-constrained fuzzy clustering algorithm based on colour affinities, and the high-level visual cues consists of a self-supervised pre-trained model, where the two segmentation vectors are fused and optimized by a feature similarity function. We also perform contrast loss computation from within-image and image will to ensure semantic consistency..

learning strategy without negative samples, achieving strong semantic consistency in the extracted features. In this paper, we utilize a self-supervised pre-training model as an advanced feature extractor to further enhance semantic context consistency.

Methodology

Problem setting

Given a set of unannotated images $I_b = [I_1, \ldots, I_B] \in \mathbb{R}^{B\times 3\times H\times W}$, where B denotes the number of images, and 3, H, and W represent the channel, height, and width dimensions, respectively, and according to the principles of color invariance and luminosity invariance, we obtain $\tilde{I}_b = [\tilde{I}_1, \ldots, \tilde{I}_B] \in \mathbb{R}^{B\times 3\times H\times W}$. The goal of unsupervised semantic segmentation is to learn a labeling function $f \in F$ that predicts the semantic labels of each pixel in each image. We denote the predicted semantic mapping as $Y = [Y_1, \ldots, Y_B] \in \{1, \ldots, K\}^{B\times H\times W}$, where K is the number of predefined categories.

Preliminary

Pretrained Features. Firstly, for each image I_b, we use a self-supervised pre-trained backbone network as an encoder, focusing on the visual features of the last output layer, which can be expressed as $F_b^p = f_\theta(I_b) \in \mathbb{R}^{C_1\times H/8\times W/8}$. Similarly, the principles of color invariance and luminosity invariance yield $\tilde{F}_b^p = f_\theta(\tilde{I}_b) \in \mathbb{R}^{C_1\times H/8\times W/8}$, where f_θ denotes the frozen self-supervised pre-trained model.

low-level visual cue. Although F_i^p contains pre-trained high-level feature mappings, low-level visual cues cannot be effectively captured by pre-trained patterns alone. Therefore, to further enhance the semantic features at different levels, we propose a fuzzy clustering branch based on color affinity to improve the coherence of features such as color and texture. This can be expressed as:

$$F_b^l = f_\gamma(\text{bilinear}(I_b)) \in \mathbb{R}^{C_{cls}\times H/8\times W/8} \quad (1)$$

where f_γ denotes the color affinity-based fuzzy C-means clustering algorithm.

Unsupervised segmentation. We map the high-level features obtained from the frozen branch of the self-supervised pre-training model to a low-dimensional embedding space using a learnable linear projection structure, h_θ. This results in $F_b^c = h_\theta(F_b^p) \in \mathbb{R}^{C_2\times H/8\times W/8}$. Following the approach in the literature, we compute the cosine similarity, S_θ, between the dimensionality-reduced features and the global clustering center. The result with the highest similarity in the global clustering center is selected as the category index for the corresponding position prediction, which can be expressed as:

$$F_b^{seg} = S_\theta(F_b^c) \in \mathbb{R}^{C_{cls}\times H/8\times W/8} \quad (2)$$

The low-level features obtained from the color affinity-based fuzzy clustering branch are then combined with the high-level features from the self-supervised pre-training

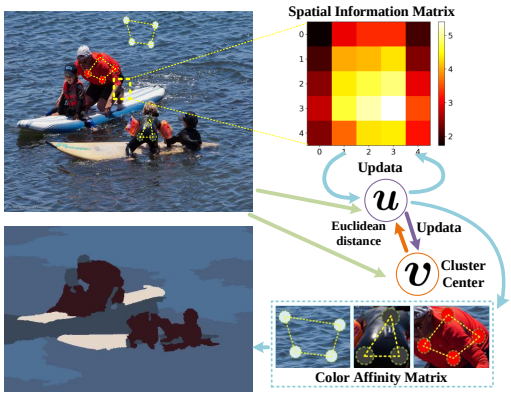

Figure 3: Spatially constrained fuzzy clustering algorithm based on colour affinity, where U denotes the affiliation matrix and V denotes the clustering centre, and U and V are continuously updated throughout the clustering process, and finally the low-level visual segmentation vector is obtained..

branch for model optimization. An overview of the proposed method is provided in Figure 2.

Low-level visual features

High-level pre-trained features provide semantic understanding but struggle with color and texture variations. Low-level visual cues, on the other hand, capture these aspects more coherently. Combining both can improve unsupervised semantic segmentation. We propose using fuzzy clustering to obtain low-level cues that complement high-level features from self-supervised pre-trained models.

Intra-class variations make it difficult to divide low-level cues into distinct regions, often resulting in noisy and isolated pixels. To address this, we introduce a spatially constrained fuzzy clustering algorithm based on color affinity, enhancing intra-class affinity of spatially adjacent and color-similar pixels. The process is illustrated in Figure 3. Our method involves: (1) Spatially Constrained Fuzzy Clustering, and (2) Colour Affinity for Feature Boosting.

Spatially Constrained Fuzzy Clustering. We apply bilinear interpolation to the original input image to obtain the downsampled result: $x = \text{bilinear}(I_b) \in \mathbb{R}^{C_3 \times H/8 \times W/8}$. FCM clustering is performed by minimizing an objective function, which is expressed as:

$$J(\boldsymbol{U}, \boldsymbol{V}) = \sum_{i=1}^{N} \sum_{j=1}^{C} (U_{ij})^m \cdot d_{ij}^2 \tag{3}$$

where $d_{ij} = \|x_i - v_j\|$ is the Euclidean distance between the i-th pixel and the j-th cluster center. The parameter m is the fuzzification index, v denotes the cluster center, and x denotes the pixel data. Our affiliation matrix $\boldsymbol{U}$ is calculated as follows:

$$U_{ij} = \left(\frac{1}{d_{ij}}\right)^{\frac{1}{m-1}} \Bigg/ \sum_{k=1}^{n} \left(\frac{1}{d_{ik}}\right)^{\frac{1}{m-1}} \tag{4}$$

To further capture the local consistency of neighboring pixel points in the image, we incorporate the spatial information matrix into the fuzzy clustering algorithm. The spatial information matrix is defined as follows:

$$S_{ij} = \sum_{k=-1}^{1} \sum_{l=-1}^{1} U_{i+k,j+l} \tag{5}$$

where S_{ij} is the spatial information value of the (i, j) pixel, and $U_{i+k,j+l}$ is the affiliation value of the $(i+k, j+l)$ pixel. The updated affiliation matrix U' is expressed as:

$$U'_{ij} = \frac{U_{ij} \cdot S_{ij}}{\sum_{k=1}^{n} U_{ik} \cdot S_{ik}} \tag{6}$$

where U'_{ij} is the updated affiliation value for the (i, j) pixels, and S_{ij} is the spatial information value of the (i, j) pixels. The update of the clustering center v is expressed as:

$$v_j = \frac{\sum_{i=1}^{N} (U'_{ij})^m \cdot x_i}{\sum_{i=1}^{N} (U'_{ij})^m} \tag{7}$$

where v_j is the j-th clustering center, and x_i is the feature vector of the i-th pixel. The clustering probability matrix is expressed as:

$$P_{ij} = \frac{U'_{ij}}{\sum_{k=1}^{n} U'_{ik}} \tag{8}$$

At this point, the clustering output effectively utilizes the domain space information, resulting in smoother and more coherent segmentation results while avoiding isolated pockets. The influence of noise is reduced by the common features of neighboring pixels, which further enhances boundary detection.

Colour Affinity for Feature Boosting. To enhance the correlation of similar features, we utilize the color affinity matrix to analyze the relationships between different feature vectors. First, the Euclidean distance between each pair of normalized pixel values $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$ is computed as follows:

$$d_{jk} = \|\boldsymbol{x}_j - \boldsymbol{x}_k\|_2 \tag{9}$$

Here, d_{jk} represents the Euclidean distance between pixel j and pixel k. Next, the color affinity is calculated using the Gaussian function:

$$a_{jk} = \exp\left(-\frac{d_{jk}^2}{2\sigma^2}\right) \tag{10}$$

where σ is the standard deviation of the Gaussian function, and a_{jk} denotes the color affinity between pixel j and pixel k. All elements of the resulting affinity matrix are then concatenated to form the final color affinity matrix:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \tag{11}$$

Finally, the color affinity matrix and clustering results are combined to enhance the coherence and accuracy of clustering by leveraging the color information of the image. This approach increases the likelihood that pixels with similar colors will be assigned to the same cluster:

$$F_b^l = A \cdot P \tag{12}$$

Visual characteristics of the fused. As previously demonstrated, the high-level feature vector obtained from the self-supervised pre-training branch is mapped to a low-dimensional space. After passing through a lightweight segmentation header, it is represented as $F_b^{seg} \in \mathbb{R}^{C_{cls} \times H/8 \times W/8}$. The low-level visual features, derived from the spatially-constrained fuzzy clustering algorithm based on color affinity, are given by $F_b^l \in \mathbb{R}^{C_{cls} \times H/8 \times W/8}$. Finally, the two feature vectors are weighted to obtain the fused segmented feature vector:

$$F_b^f = (1-\alpha) \cdot F_b^{seg} + \alpha \cdot F_b^l \in \mathbb{R}^{C_{cls} \times H/8 \times W/8} \tag{13}$$

Loss function

To effectively couple low-level and high-level visual cues, we propose a feature similarity loss function to optimize the feature representation of fused visual cues. Specifically, this function enhances the spatial structure and feature continuity of the fused feature vectors by combining feature similarity and spatial continuity, thereby improving the clustering effect of similar regions in the image.

$$\mathcal{L}_{fs}, \mathcal{L}_{reg} = f(F^c, F_b^f, I) \tag{14}$$

The feature similarity matrix captures feature similarity, while the pixel similarity matrix represents spatial relationships. By integrating these matrices with the Gram matrix projection, we compute $\mathcal{L}_{fs}$ and the regularization term $\mathcal{L}_{reg}$.

To promote smoothness within segments and preserve discontinuities, we introduce the Smoothing Loss(Lan et al. 2024) Function $\mathcal{L}_{smo}$. Additionally, the Correspondence Distillation Loss(Hamilton et al. 2022) Function $\mathcal{L}_{corr}$ enhances training accuracy and semantic categories.

The final total loss function is:

$$\mathcal{L}_{total} = \beta \cdot \text{reg} + \gamma \cdot \mathcal{L}_{fs} + \mathcal{L}_{smo} + \mathcal{L}_{corr} \tag{15}$$

where β and γ are hyperparameters ranging between [0,1].

Experiments

Experimental Settings

Datasets. We use the COCOStuff(Caesar, Uijlings, and Ferrari 2018), Cityscapes(Cordts et al. 2016), Potsdam-3(Ji, Henriques, and Vedaldi 2019), and Mastr1325(Bovcon et al. 2019) datasets. COCOStuff, derived from COCO, includes 91 object categories and is widely used for semantic segmentation due to its rich scenes and detailed annotations. Cityscapes focuses on urban environments, emphasizing vehicle and street object segmentation. We merge the categories of COCOStuff and Cityscapes separately, resulting in 27 evaluation categories. Potsdam-3 consists of high-resolution remote sensing images with urban annotations, while MaSTr1325 is a maritime dataset aimed at enhancing obstacle segmentation for small coastal USVs, both evaluated using three categories.

Methods	backbone	Acc.	mIoU
ResNet50 (He et al. 2016)	ResNet50	24.6	8.9
IIC (Ji and Vedaldi 2019)	R18+FPN	21.8	6.7
MDC (Cho et al. 2021)	R18+FPN	32.2	9.8
PiCIE (Cho et al. 2021)	R18+FPN	48.1	13.8
PiCIE+H (Cho et al. 2021)	R18+FPN	50.0	14.4
SlotCon (Wen et al. 2022)	ResNet50	42.4	18.3
MoCoV2 (Chen et al. 2020)	ResNet50	25.2	10.4
+ STE(Hamilton et al. 2022)	ResNet50	43.1	19.6
+ SmooSeg(Lan et al. 2024)	ResNet50	52.4	18.8
+ ours	ResNet50	**54.2**	**20.1**
DINO (Caron et al. 2021)	ViT-S/8	29.6	10.8
+ TransFGU (Yin et al. 2022)	ViT-S/8	52.7	17.5
+ STE(Hamilton et al. 2022)	ViT-S/8	48.3	24.5
+ HP(Seong et al. 2023)	ViT-S/8	57.2	24.6
+ SmooSeg(Lan et al. 2024)	ViT-S/8	63.2	26.7
+ EAGLE(Kim et al. 2024)	ViT-S/8	64.2	27.2
+ ours	ViT-S/8	**65.1**	**27.6**

Table 1: Quantitative results on the COCOStuff dataset.

Evaluation metrics. Consistent with existing methods, we use minibatch K-means based on cosine similarity for cluster segmentation. Since no real labels are used, we optimize the matching relationship between the predictions and the real semantic graph using the Hungarian matching algorithm. Additionally, we apply Conditional Random Fields (CRF)(Krähenbühl and Koltun 2011) to post-process the predicted results, further refining the semantic mapping. For quantitative evaluation, we use accuracy (Acc) and mean intersection over union (mIoU) to measure the performance of different methods.

Methods	backbone	Acc.	mIoU
IIC (Ji and Vedaldi 2019)	R18+FPN	47.9	6.4
MDC (Cho et al. 2021)	R18+FPN	40.7	7.1
PiCIE (Cho et al. 2021)	R18+FPN	65.5	12.3
DINO(Caron et al. 2021)	ViT-S/8	40.5	13.7
+ TransFGU(Yin et al. 2022)	ViT-S/8	77.9	16.8
+ STE(Hamilton et al. 2022)	ViT-S/8	69.8	17.6
+ HP(Seong et al. 2023)	ViT-S/8	80.1	18.4
+ SmooSeg(Lan et al. 2024)	ViT-S/8	81.8	19.7
+ EAGLE(Kim et al. 2024)	ViT-S/8	82.8	18.4
+ ours	ViT-S/8	**83.0**	**20.6**

Table 2: Quantitative results on the Cityscapes dataset.

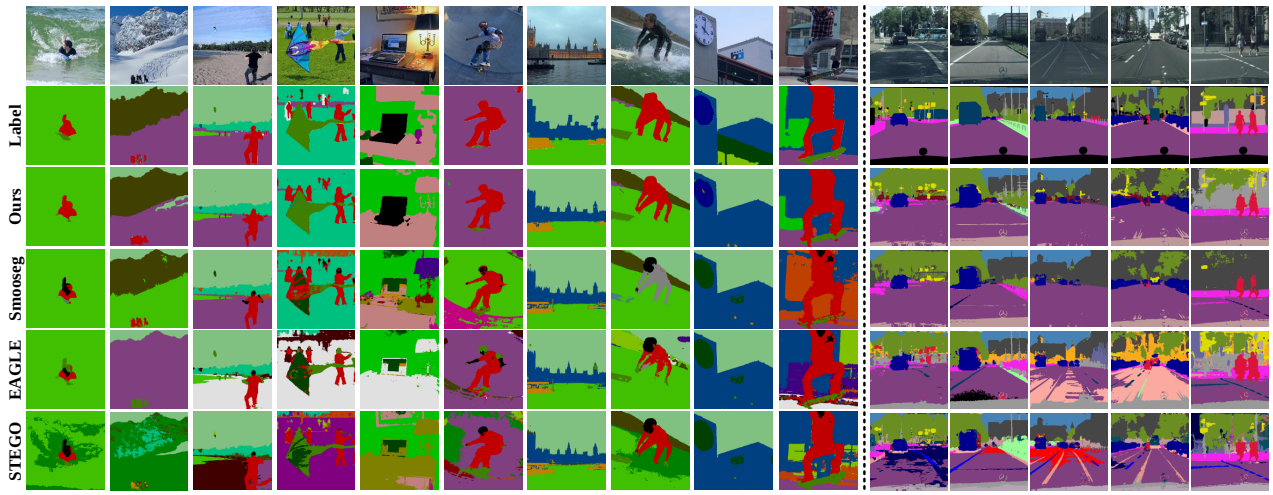

Figure 4: Quantitative comparison of the results of the proposed IL2Vseg and other state-of-the-art methods on the COCOStuff (left) and Cityscapes (right) datasets.

Implementation Details. We conducted our experiments using the PyTorch 1.13 framework, running on an RTX 4090 GPU. For a fair comparison with previous works, we used the DINO pre-trained model as our self-supervised feature extractor f_θ, which was kept frozen during training. The nonlinear projection h_θ consists of a linear convolution and two MLP nonlinear layers with SiLUs to obtain more compact correlated features. The cosine similarity of these compact correlation features is computed with the local and global clustering centers, respectively. The result with the highest similarity at the global clustering center is chosen as the category index for the corresponding positional projection. Similar to Smooseg, an exponential moving average (EMA)(Haynes, Corns, and Venayagamoorthy 2012) is used to update the global clustering center. We used the Adam optimizer to optimize the nonlinear projections and local clustering centers, with learning rates set to 1×10^{-4} and 5×10^{-4}, respectively, to maximize the similarity of the input features of the same category to the corresponding clustering center. The weights β and γ are taken as 0.05 and 0.1 respectively.

Comparison with State-of-the-art Methods

We compared our proposed method with several recent unsupervised semantic segmentation methods, both quantitatively and qualitatively.

Methods	backbone	Acc.	mIoU
DINO (Caron et al. 2021)	ViT-S/8	55.1	37.3
+ STE(Hamilton et al. 2022)	ViT-S/8	70.8	54.4
+ HP (Seong et al. 2023)	ViT-S/8	82.7	71.2
+ SmooSeg(Lan et al. 2024)	ViT-S/8	88.3	74.1
+ EAGLE(Kim et al. 2024)	ViT-S/8	84.0	70.5
+ ours	ViT-S/8	**93.6**	**82.5**

Table 3: Quantitative results on the MaSTr1325 Dataset.

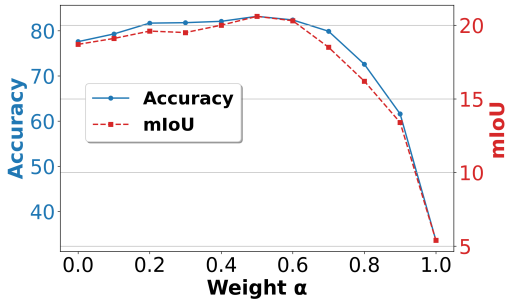

Figure 5: Segmentation results of IL2Vseg after using low-level visual cues with different weights(training).

Quantitative Evaluation. We report the results for the COCO-Stuff and Cityscapes datasets in Table 1 and Table 2. On the COCO-Stuff dataset, using ViT-S/8 as the backbone, both TransFGU and STEGO(STE) significantly outperform previous methods due to the advanced features provided by the self-supervised pre-trained model. For example, compared to PiCIE, TransFGU improves accuracy (Acc) by 4.6 and mean Intersection over Union (mIoU) by 3.1. Smooseg obtains coherent semantic segments from a smoothness prior, significantly improving the performance of unsupervised semantic segmentation, with an improvement of 14.9 in Acc and 2.1 in mIoU compared to STEGO. Our proposed IL2Vseg outperforms all state-of-the-art methods in terms of pixel accuracy and mIoU. Compared to the baseline STEGO, our method achieves a significant improvement (+16.8 Acc, +3.3 mIoU) and also yields better results compared to Smooseg (+1.9 Acc, +0.9 mIoU). This is due to IL2Vseg's use of low-level visual cues to complement high-level visual cues based on self-supervised pre-trained branches, ensuring coherent recognition of color-texture aspects and continuity of the spatial structure within the class. The same trend is observed in the Cityscapes dataset. Com-

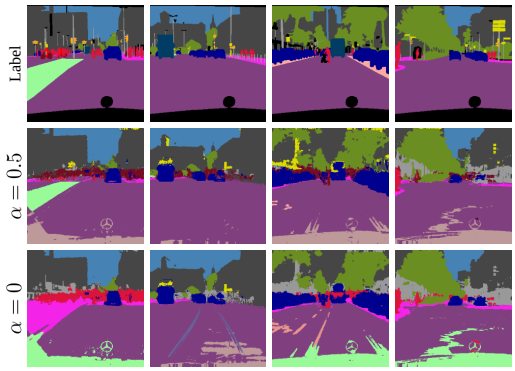

Figure 6: Segmentation results of IL2Vseg after using low-level visual cues with different weights.

pared with Smooseg and EAGLE, our IL2Vseg achieves better segmentation performance. It is worth noting that all subsequent work based on STEGO maintains a low mIoU value because these methods achieve better class-balanced segmentation results with small batch k-means, without much attention to pixel-wise accuracy.

The quantitative results for the Mastr1325 dataset are presented in Table 3. Our method significantly outperforms others, achieving better results than EAGLE (+9.6 ACC, +12 mIoU) and Smooseg (+5.3 ACC, +8.4 mIoU). This is because the simpler scene distribution of the Mastr1325 dataset, with large continuous areas and distinct decompositions of different regions, allows low-level visual cues to work more effectively, resulting in more accurate and continuous semantics.

Quantitative Evaluation. Figure 4 presents a qualitative comparison of our proposed method with other state-of-the-art techniques. Our method demonstrates superior segmentation details compared to others. Although STEGO tends to form compact correlation features, it still produces discontinuous regions during segmentation, as observed in columns 1 and 2 of the left COCO dataset and columns 1 and 2 of the right Cityscapes dataset. Smooseg maintains differences between segments by smoothing the prior, but it requires further improvement in semantic mapping accuracy, as shown in columns 2, 5, 6, and 8. While Smooseg achieves more continuous and complete semantic segments in columns 6 and 8, the segmented categories are often incorrect. Additionally, Smooseg sometimes results in overly complete segments, such as in column 5 of the right Cityscapes dataset, where tree semantics are largely mistaken for buildings. In contrast, our IL2Vseg method achieves better results in both semantic continuity and category accuracy.

Ablation Study

low-level visual cue. To further explore the proposed IL2Vseg, we conducted a series of ablation experiments. To assess the importance of using low-level visual cues for supplementation, we evaluated the performance of different α values, as shown in Figure 5. An α value of 0 indicates that only segmentation feature vectors obtained from the self-supervised pre-training model are used, while an α value of 1 indicates that only low-level visual cues are used as segmentation feature vectors. The progression of α from 0 to 1 signifies an increasing weight of low-level visual cues. The experimental results demonstrate that the best segmentation performance is achieved when α is 0.5 during the training phase, indicating that low-level visual cues effectively complement the self-supervised pre-trained features. When only low-level visual cues are used, the results are significantly inferior to other experiments, suggesting that the deep features extracted by self-supervised pre-training play a crucial role in semantic coherence. Finally, we present the segmentation results obtained with $\alpha = 0.5$ (indicating the use of low-level visual cues) and $\alpha = 0$ (indicating no use of low-level visual cues), as shown in Figure 6.

	L_{smo}	L_{fs}	L_{corr}	Acc.	mIoU
(1)	✓			75.7	17.4
(2)	✓	✓		81.1	19.6
(3)	✓		✓	80.3	18.3
(4)	✓	✓	✓	83.0	20.6

Table 4: Analysis of Loss function on the Cityscapes dataset.

Loss function. Table 4 illustrates the effect of different loss functions on IL2Vseg. The model using all loss functions outperforms the other methods. The smoothing loss produces more coherent and semantically meaningful segmentation maps, as evidenced by the Group 1 experiments, which also achieve good performance using only the smoothing loss. The feature similarity loss function effectively couples low-level visual cues with high-level visual cues. As shown in the Group 3 and Group 4 experiments, the feature similarity loss function combines feature similarity and spatial continuity to enhance the clustering results of similar regions in an image by optimizing the feature representation. Additionally, to further improve the accuracy of semantic categories, we use Correspondence Distillation Loss for the original image and the image after invariant enhancement.

Conclusions

We found that low-level visual cues in unsupervised semantic segmentation can complement features extracted from a self-supervised pre-trained visual backbone. These low-level visual cues are obtained through a spatially constrained fuzzy clustering algorithm based on color affinity. Additionally, we propose a feature similarity loss function to integrate the fused segmented feature cues. To enhance the accuracy of semantic category features, we introduce Correspondence Distillation Loss, which improves the learning of consistency across semantic categories using color and luminosity invariant transformations. Results on several public datasets, such as COCO-Stuff and Cityscapes, indicate that IL2Vseg achieves state-of-the-art performance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62122046, U24A20279, 62473243, supported by the Shanghai Commission of Science and Technology, China (Grant 23010500100), and supported by National Key RD Program of China under Grant 2023YFB4707000.

References

Bovcon, B.; Muhovič, J.; Perš, J.; and Kristan, M. 2019. The mastr1325 dataset for training deep usv obstacle detection models. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3431–3438. IEEE.

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

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, 132–149.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33: 9912–9924.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Cho, J. H.; Mall, U.; Bala, K.; and Hariharan, B. 2021. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16794–16804.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Deng, Z.; and Luo, Y. 2023. Learning neural eigenfunctions for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 551–561.

Hafidi, H.; Ghogho, M.; Ciblat, P.; and Swami, A. 2020. Graphcl: Contrastive self-supervised learning of graph representations. arXiv 2020. *arXiv preprint arXiv:2007.08025*.

Hamilton, M.; Zhang, Z.; Hariharan, B.; Snavely, N.; and Freeman, W. T. 2022. Unsupervised semantic segmentation by distilling feature correspondences. *International Conference on Learning Representations*.

Harb, R.; and Knöbelreiter, P. 2021. Infoseg: Unsupervised semantic image segmentation with mutual information maximization. In *DAGM German Conference on Pattern Recognition*, 18–32. Springer.

Hassani, K.; and Khasahmadi, A. H. 2020. Contrastive multi-view representation learning on graphs. In *International conference on machine learning*, 4116–4126. PMLR.

Haynes, D.; Corns, S.; and Venayagamoorthy, G. K. 2012. An exponential moving average algorithm. In *2012 IEEE Congress on Evolutionary Computation*, 1–8. IEEE.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hwang, J.-J.; Yu, S.; Shi, J.; Collins, M.; Yang, T.-J.; Zhang, X.; and Chen, L.-C. 2019. SegSort: Segmentation by Discriminative Sorting of Segments. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7333–7343.

Ji, H. J. F., Xu; and Vedaldi, A. 2019. Invariant Information Clustering for Unsupervised Image Classification and Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Ji, X.; Henriques, J. F.; and Vedaldi, A. 2019. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9865–9874.

Kim, C.; Han, W.; Ju, D.; and Hwang, S. J. 2024. EAGLE: Eigen Aggregation Learning for Object-Centric Unsupervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3523–3533.

Krähenbühl, P.; and Koltun, V. 2011. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24.

Krinidis, S.; and Chatzis, V. 2010. A robust fuzzy local information C-means clustering algorithm. *IEEE transactions on image processing*, 19(5): 1328–1337.

Lan, M.; Wang, X.; Ke, Y.; Xu, J.; Feng, L.; and Zhang, W. 2024. SmooSeg: smoothness prior for unsupervised semantic segmentation. *Advances in Neural Information Processing Systems*, 36.

Melas-Kyriazi, L.; Rupprecht, C.; Laina, I.; and Vedaldi, A. 2022. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8364–8375.

Na, S.; Xumin, L.; and Yong, G. 2010. Research on k-means clustering algorithm: An improved k-means clustering algorithm. In *2010 Third International Symposium on intelligent information technology and security informatics*, 63–67. Ieee.

Reynolds, D. A.; et al. 2009. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663).

Seong, H. S.; Moon, W.; Lee, S.; and Heo, J.-P. 2023. Leveraging hidden positives for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19540–19549.

Tang, Y.; Ren, F.; and Pedrycz, W. 2020. Fuzzy C-means clustering through SSIM and patch for image segmentation. *Applied Soft Computing*, 87: 105928.

Wen, X.; Zhao, B.; Zheng, A.; Zhang, X.; and Qi, X. 2022. Self-supervised visual representation learning with semantic grouping. *Advances in neural information processing systems*, 35: 16423–16438.

Yin, Z.; Wang, P.; Wang, F.; Xu, X.; Zhang, H.; Li, H.; and Jin, R. 2022. Transfgu: a top-down approach to fine-grained unsupervised semantic segmentation. In *European conference on computer vision*, 73–89. Springer.

Zhang, H.; Wang, Q.; Shi, W.; and Hao, M. 2017. A novel adaptive fuzzy local information C-means clustering algorithm for remotely sensed imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9): 5057–5068.

Zhang, Y.; Bai, X.; Fan, R.; and Wang, Z. 2018. Deviation-sparse fuzzy c-means with neighbor information constraint. *IEEE Transactions on Fuzzy systems*, 27(1): 185–199.