Knowledge-Based Systems 299 (2024) 112097

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

CAA: Class-Aware Affinity calculation add-on for semantic segmentation

Huadong Tang [a,*], Youpeng Zhao [b], Chaofan Du [c], Min Xu [a], Qiang Wu [a]

[a] School of Electrical and Data Engineering Faculty of Engineering and IT, University of Technology Sydney, Sydeny, 2007, NSW, Australia
[b] Department of Computer Science, University of Central Florida, Orlando, FL, USA
[c] School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing, Hai Dian District, China

ARTICLE INFO

Keywords:
Semantic segmentation
Affinity
Contextual dependencies
Class association

ABSTRACT

Leveraging contextual dependencies is a commonly used technique to enhance the performance of image segmentation. However, existing solutions do not effectively catch the class-level association between the pixels along the boundary across the objects of the different classes but focus more on the local pixel-to-pixel relation. This work proposes a Class-Aware Affinity module (CAA) that considers both pixel-to-pixel relation and pixel-to-class association. We try to argue that the pixel-to-pixel relations still catch the relation (*e.g.* similarity, attention, or affiliation) on the local texture level. At the same time, it should also consider the association between the pixel and the class context produced by the given image. Pixel-to-class association can best reveal the co-occurrent dependency on the semantic level between the given pixels and their nearby context. Such pixel-to-class association combined with the pixel-to-pixel relations aggregating the local texture information will best mitigate the confusion caused in the boundary regions across the objects of the different classes. Moreover, the proposed framework can serve as a generic add-on to be integrated with the existing image segmentation solution to boost the current performance. Equipped with CAA, we achieve promising performance against the existing work with 54.59% mIoU on ADE20K, 49.96% mIoU on COCO-Stuff10k, and 64.38% mIoU on Pascal-Context.

1. Introduction

Semantic Segmentation is a challenging and fundamental problem in computer vision, which aims to assign each pixel a clear semantic label for a given image. This task can be applied to several real-world applications, such as medical image processing [1], autonomous driving [2], and precision agriculture [3].

Building upon the foundation laid by the fully convolutional networks (FCNs) [4], numerous methods have achieved remarkable advancements. However, the FCN has a problem of low efficiency and only provides insufficient contextual dependencies for the reason of structural weaknesses. The limitation of insufficient background information greatly affects its segmentation accuracy. Given this, researchers primarily focus on two aspects to enhance segmentation performance: (i) design a better encoder structure [5–7]; and (ii) model reliable contextual information [8–10].

Since the existence of co-occurrent visual patterns [11–13], a series of works focus on modeling context. Early study is mainly about multi-scale context for semantic segmentation, which exploits dilated convolutions or pyramid pooling to obtain feature maps by aggregating multi-scale contexts. Specifically, PSPNet [8] employs pyramid spatial pooling to aggregate context. However, they only can capture local features and bring limited contextual information. Deeplab [14–16] family introduces the atrous spatial pyramid pooling (ASPP) to capture local context from different scales of the image. Some other methods utilize dot self-attention to extract long-range dependencies. Non-local network [17] first proposes to utilize a self-attention module to learn global contextual dependencies. Inspired by this, some works [10,11,18,19] further utilize the attention mechanism for semantic segmentation. PSANet [20] introduces a pyramid spatial attention mechanism for capturing features at multiple scales. Meanwhile, DANet [10] introduces channel attention to cooperate with spatial attention to capture contextual information globally. [21] designs to enhance feature extraction and prediction stages by considering channel perspectives. However, these methods only focus on pixel-to-pixel dependencies. Multi-scale contexts are established within predefined regions, with the only pixel correlation being the overlap of receptive fields. They only focus on local pixel relationships, leading to the category confusion problem on a semantic level. Besides, attention-based contexts only catch the relation on the local texture level. We argue that they should consider the association between the pixel and the class context produced by the given image.

[*] Corresponding author.
E-mail addresses: huadong.tang@student.uts.edu.au (H. Tang), ypzhao@knights.ucf.edu (Y. Zhao), 18116005@bjtu.edu.cn (C. Du), min.xu@uts.edu.au (M. Xu), qiang.wu@uts.edu.au (Q. Wu).

https://doi.org/10.1016/j.knosys.2024.112097
Received 21 January 2024; Received in revised form 28 May 2024; Accepted 6 June 2024
Available online 8 June 2024

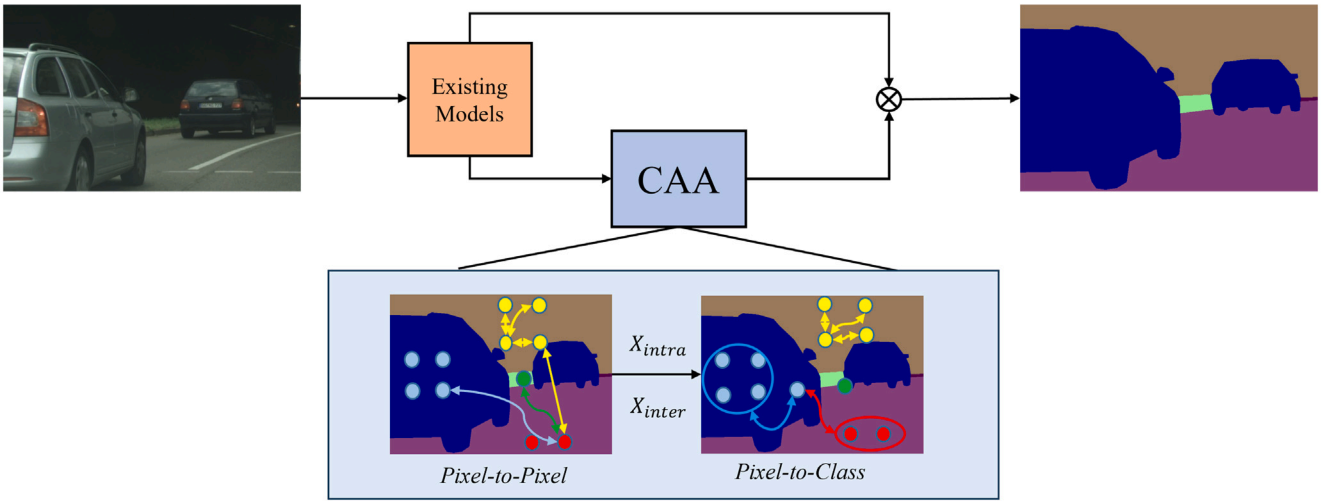


Fig. 1. The concept of our CAA. We first explore pixel-to-pixel relation: whether two pixels belong to the same class or not. Then, we calculate the pixel-to-class relation: the relation between the ith pixel and other pixels in the jth class region, *i.e.*, the blue point and the blue car region.

Apart from this, ACFNet [22] and OCRNet [23] first propose the idea of class-center, which calculates relations between pixels and class center (pixel-to-class). Then, ISNet [12] proposes a similar class-center by utilizing the class region representation, which shares the same class as the pixel representation instead of representations of every region. Besides, CPNet [24] proposes a context prior map to predict if two pixels belong to the same class. Those above-mentioned methods either focus on pixel-to-pixel or pixel-to-class relation. *So what is the context or reliable context?* The pixel-to-pixel (spatial) relation represents whether two pixels belong to the same class or not, while the pixel-to-class (class-level) association indicates the probability of a pixel belonging to the specific class. Therefore, we propose that reliable contexts can describe the pixel-to-pixel relation and pixel-to-class association.

To this end, we introduce a Class-Aware Affinity Module (CAA), which is regarded as an additional supplementary module to the existing mainstream segmentation framework [4,8,15] (see Fig. 1). The existing mainstream architectures provide feature extraction to aggregate the spatial information. Therefore, we further utilize those features to capture pixel-to-pixel relation to enhance pixel representations and get intra- and inter-class representations X_{intra} and X_{inter}. Intra-class representations refer to the augmentation of contextual dependencies among pixels belonging to the same class, whereas inter-class representations belong to the heightened relationships among pixels of distinct classes. Every colored dot represents a pixel along with its corresponding class. We just know if two pixels belong to the same class, however, we still do not explicitly build the relationship between pixel and class region *i.e.*, the relation between blue dots (‘*car*’), yellow dots (‘*tunnel*’), green dots (‘*terrain*’) and the red dots (‘*road*’). We further model the pixel-to-class associations *e.g.*, the probability of blue dots belonging to the ‘car’ region.

Specifically, we design an affinity map to generate two types of representations for each pixel. One only considers the dependency between pixels of the same class, which is intra-class representation. Another considers the relation of pixels across different classes, which is inter-class representations. Besides, an affinity loss is developed to enforce the network to better generate pixel-to-pixel relations. With the affinity map, we can better learn whether two pixels belong to the same class or not. Due to the fact that the label assigned to a pixel signifies the category of the object it is associated with, we further calculate the relation between pixels and the class center. The class center is determined through the aggregation of feature vectors from all pixels within the same class label.

We show examples of segmentation results produced by the attention-based (pixel-to-pixel) method (EncNet [9]) and class-level (pixel-to-class) method (OCRNet [23]) in Fig. 2. In the first row and the last row of Fig. 2, both EncNet [9] and OCRNet [23] recognize ‘earth’ as ‘sand’ and ‘pole’ as ‘fence’, while our CAA (joining pixel-to-pixel relation and pixel-to-class association together in a unified framework) successfully recognizes the class ‘earth’ and ‘pole’. In the second row, OCRNet [23] performs better than EncNet [9] but still makes some mistakes (recognizes ‘windowpane’ as ‘painting’), while our CAA recognizes the right class.

CAA contributes to the existing image segmentation solution regardless of the approaches of the supervised, unsupervised, or semi-supervised pipeline. It focuses on enhancing the contextual information to improve the quality of the semantic segmentation. In terms of a pipeline of unsupervised or semi-supervised approaches, given the generated class information (identified by pseudo-label), CAA is to explore the pixel-to-pixel and pixel-to-class contextual information which is then integrated into the existing segmentation method to further improve the overall quality.

In a nutshell, this paper presents the following key contributions:

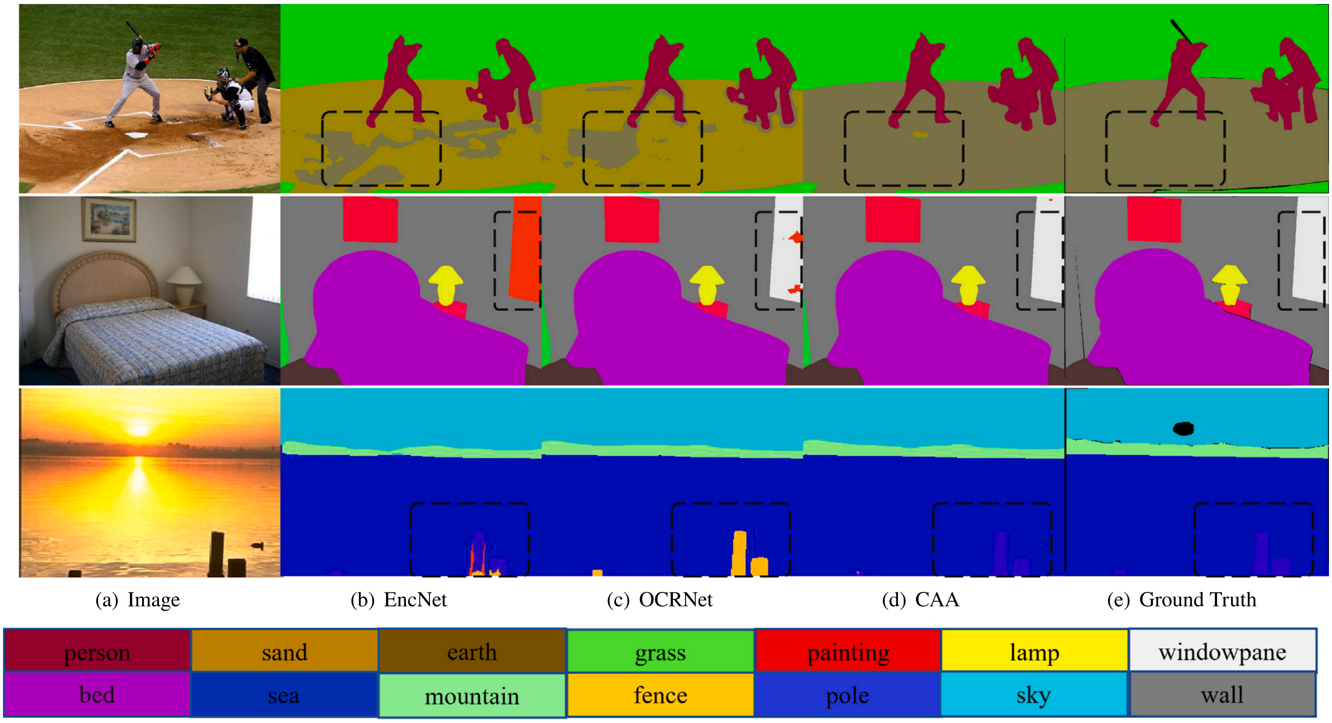
- We introduce a class-aware affinity module (CAA) to mitigate the issue of intra-class compactness and inter-class dispersion. Our CAA explores both pixel-to-pixel relations and pixel-to-class associations. This module can be effortlessly incorporated into existing segmentation frameworks and improve the performance of the corresponding model in which CAA is added.
- An affinity map is proposed to learn the pixel-to-pixel relation and generate intra- and inter-class representations.
- Class center is proposed to explore the pixel-to-class associations for further corresponding context calculation.

The remainder of this paper is organized as follows. A brief review on the related work of multi-scale context, attention-based context and class-level context for semantic segmentation in Section 2. We present the proposed class center and semantic affinity in Section 3. Experimental results are reported and analyzed in Section 4. Finally, Section 5 is a concluding summary, followed by future work in Section 6.

2. Related work

2.1. Multi-scale context for semantic segmentation

The Fully Convolutional Networks (FCNs) [4] is an epoch-making work to promote the advancement of semantic segmentation. Based



(a) Image (b) EncNet (c) OCRNet (d) CAA (e) Ground Truth

Fig. 2. Examples of segmentation results for ADE20K. The results of EncNet and OCRNet are shown in (b) and (c), which explore the pixel-to-pixel relations and pixel-to-class relations, respectively. Our proposed CAA module combines the relation of pixels-pixels and pixels-class association for the final prediction. Obviously, our method achieves better prediction than the methods mentioned above, as shown in (d).

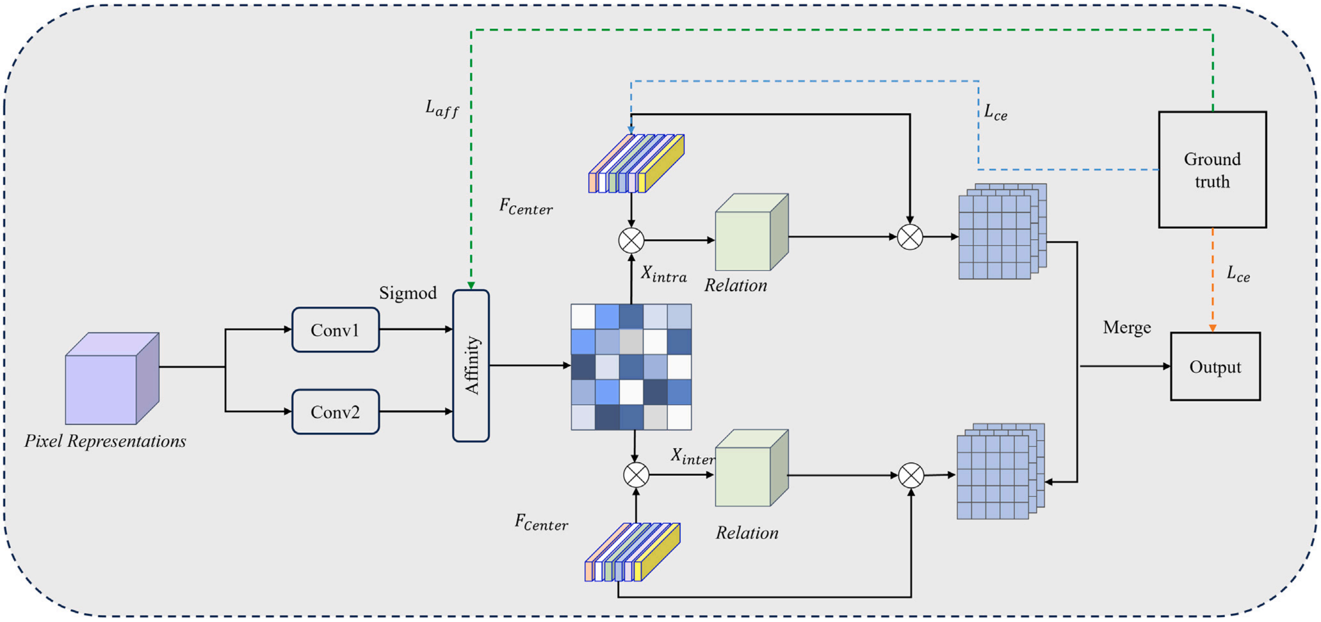


Fig. 3. Illustrating the pipeline of CAA. CAA explores the pixel-to-pixel relation and pixel-to-class association by levering the semantic affinity and class center. Semantic affinity explores the pixel dependencies to learn pixel-to-pixel relations. We utilize the class center to further calculate the pixel-class dependencies by considering category representations.

on this method, aggregating contextual information to enhance feature representation is a common practice. Many recent works [25–28] have been proposed to extract discriminative features by combining contextual features. Specifically, Deeplab methods [14–16] propose atrous spatial pyramid pooling (ASPP) to capture more contexts from multiple scales, while PSPNet [8] employs a pyramid parsing module for global context exploration, achieved through context aggregation across diverse regions. CCL [29] introduces a new approach for scene segmentation, which improves performance by incorporating contextual information and multi-scale features. Some works [9,29,30] aim to capture more comprehensive global context information by extending kernel size or proposing an efficient encoding layer. For instance, ACNet [30] captures pixel-aware context by integrating global and local contexts regarding different pixel requirements. EncNet [9] proposes to obtain a comprehensive global context and selectively emphasize context relevant to specific classes. These methods aggregate contextual information with multi-scale local features and cannot capture different dependencies for different pixels. Recently, attention-based models have shown considerable performance and emerged as a widely adopted approach for semantic segmentation.

2.2. Attention-based context semantic segmentation models

The attention mechanism is initially a technique used in Natural Language Processing (NLP) task [31,32] to get more useful information about the target and ignore extraneous or irrelevant data. For the semantic segmentation task, long-range dependencies are usually captured by stacking convolutional operations to obtain large receptive fields. However, these operations can only capture local contextual information and are computationally inefficient. The attention mechanism can capture the global context information.

Specifically, DANet [10] and RCANet [33] propose a two-attention module network to aggregate long-range spatial information. CCNet [13] and SPNet [11] come up with a criss-cross attention module and strip pooling module to capture long-range dependencies while reducing computational complexity. Besides, SANet [34] proposes a new squeeze-and-attention network, which takes into account dense predictions at multiple scales for individual pixels, as well as spatial attention for clusters of pixels. Although the attention methods could capture long-range contextual dependencies, they do not explicitly model the dependencies of pixels among classes.

2.3. Class-level context for semantic segmentation

ACFNet [22] introduces the idea of a class center, which captures the holistic context of each class. This enables pixels to discern the performance of distinct classes within the entire scene. Recently, OCRNet [23] and CTNet [35] propose to model the relation between object regions, exchanging regional context for enhancing pixel dependencies. All of these works [22,23,35] focus on extracting intra-class centers, but they ignore inter-class centers. Recently, CPNet [24] proposes to explore affinity-aware context information to model pixel-to-pixel relations. Albeit CPNet can indicate whether two pixels belong to the same class, they do not further differentiate current pixels with different classes.

The previous multi-scale context cannot capture different dependencies for different pixels, while attention-based context methods do not explicitly model the dependencies of pixels among classes. Besides, the class-level contexts only focus on pixel-to-class association. To this end, the challenge is to model a reliable context that includes the pixel-to-pixel relation and pixel-to-class association. Thus, a class-aware affinity module is proposed to help existing segmentation frameworks improve performance by integrating pixel-to-pixel relation and pixel-to-class association.

3. Methodology

3.1. Motivation

Analyzing the pixel-to-class association provides the most lucid depiction of the semantic interdependence between the specified pixels and their adjacent context. For instance, it is uncommon to find a bike in water. Therefore, classifying the presence of water can help minimize the likelihood of erroneously identifying an object in the water as a bike. Hence, it is necessary to enhance the pixel-to-class association.

Meanwhile, it is easily misclassified for the pixels on the margin of two objects in semantic segmentation. Therefore, we need to differentiate whether two adjacent pixels belong to the same class or not, that is, pixel-to-pixel relation. We advocate that pixel-to-class association combined with pixel-to-pixel relations will best reduce the ambiguity in the boundary areas between objects of distinct classes. To this end, we devise a Class-Aware Affinity Module (see Fig. 1), which can help existing models increase intra-class compactness and reduce inter-class dependency.

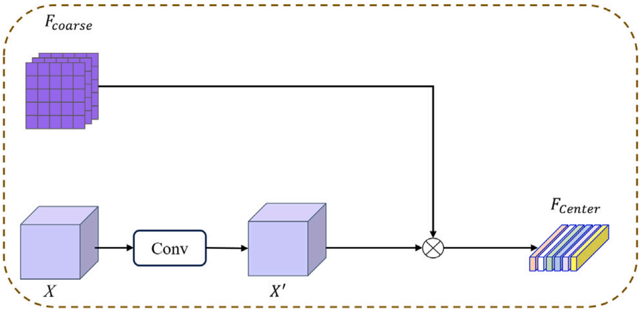


Fig. 4. Illustration of Class Center. We partition the initial image I into distinct class regions, facilitating subsequent pixel-to-class association calculations.

3.2. Class-aware affinity

3.2.1. Class center

The feature generator G receives the input image I and projects it into high-dimensional feature $F \in \mathbb{R}^{D\times H\times W}$ and coarse segmentation result $F_{coarse} \in \mathbb{R}^{N_{class}\times H\times W}$. To minimize the computational expenses, we apply 1×1 $conv \rightarrow BN \rightarrow ReLU$ operations to decrease the channel dimension to D'. Then, we reshape F_{coarse} to $R^{N_{class}\times HW}$ and F to $R^{D'\times HW}$, as illustrated in Fig. 4. After that, the class center, denoting a pixel-level probability output for each class, is computed by:

$$F_{center} = Softmax(F_{coarse}) \otimes F^{T} \tag{1}$$

where $F_{center} \in \mathbb{R}^{N_{class}\times D'}$. We learn the class center with the supervision of ground truth using cross-entropy loss during the training phase. The class center aids the model in comprehensively learning representations for all classes from a global perspective. Moreover, we can calculate the consistency between a pixel and each class center to improve the segmentation performance.

3.2.2. Semantic affinity

Affinity map. Intra-class compactness and inter-class dispersion will determine segmentation performance to some extent. However, some works [8,16] propose to aggregate local features as a mixture, which can result in the misclassification of distinct categories. To simulate extensive contextual information, we incorporate affinity maps to represent refined local features by distinguishing pixels of the same category from those of distinct categories. As depicted in Fig. 3, $F \in \mathbb{R}^{D\times H\times W}$ represents the local feature, where $H \times W$ represents the resolution and D represents the dimension of channels.

Then, to reduce the channel dimension, we apply two 1×1 convolutional layers on feature F to produce Affinity Map, denoted as M, with dimensions $M \in \mathbb{R}^{(H\times W)\times(H\times W)}$. We learn a direct representation of the correlation between pixels of intra- and inter-class from the affinity map. That is, whether the ith pixel and jth belong to the same category. We further extract pixel representations of intra- and inter-class in the following manner:

$$F_{intra} = M \otimes F \tag{2}$$

$$F_{inter} = (1 - M) \otimes F \tag{3}$$

the feature size of F_{intra} and F_{inter} are $H \times W$, and $\otimes$ represents matrix multiplication.

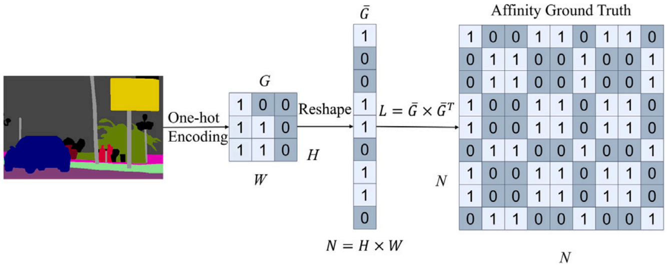


Fig. 5. The process of generating the Affinity Ground Truth. We down-sample the ground truth and apply one-hot encoding to obtain $\bar{G}$. Then, matrix multiplication is conducted to generate the affinity ground truth.

We employ binary cross-entropy loss to supervise the affinity map, treating it as a binary classification task for each pixel, distinguishing between the same class or different classes. In particular, we employ a down-sample operation to reshape ground truth G and match the size of feature F. Next, we perform one-hot encoding to obtain the transformed ground truth denoted as $\bar{G}$. Ultimately, a multiplication operation is conducted between $\bar{G}$ and its transpose to derive the affinity ground truth: $L = \bar{G} \times \bar{G}^T$. The process of constructing L is illustrated in Fig. 5. From the affinity ground truth, we can learn that $L_{ij} = 1$ means the ith pixel and the jth pixel from the original image are in the same class, while $L_{ij} = 0$ means the ith pixel and the jth from the original image are in different classes. We formulate the affinity loss as:

$$L_{aff} = -\frac{1}{N^2}\sum_{n=1}^{N^2}(P_g \log P_a + (1 - P_g)\log(1 - P_a)) \tag{4}$$

P_g represents the probability assigned to the target class, derived from the ground truth. P_a represents the predicted probabilities across different categories within the images.

Intra-class association. These pixel representations only contain the spatial relation between any two pixels. That is, we can only know whether two pixels belong to the same category or not, but the category information has not been extracted. Thus, we compute the association between intra-class pixel representations F_{intra} and class center representation F_{center} in the following manner:

$$r_{ij} = \frac{exp(F_{intra_i} \cdot F_{center}j)}{\sum_{j=1}^{N} exp(F_{intra_i} \cdot F_{center}j)} \tag{5}$$

where r_{ij} stands for the correlation between the ith pixel and other pixels in the jth class center. Furthermore, it consolidates the correlation between the ith pixel and other pixels within the same category.

We enhance the feature representations by conducting a matrix multiplication between the relation s_{ij} and the class context representation F_{center} to yield the output $F^{\approx}_{intra}$ as outlined below:

$$F^{\sim}_{intra} = \sum_{j=1}^{N}(r_{ij}F_{center}j) \tag{6}$$

$F^{\sim}_{intra}$ refers to the augmented intra-class representations.

Inter-class association. To mitigate the issue of inter-class dispersion, we calculate the correlation between the inter-class pixel representations F_{inter} and the class context representation $F_{center}j$. Much like the intra-class association, we also have:

$$t_{ij} = \frac{exp(F_{inter_i} \cdot F_{center}j)}{\sum_{j=1}^{N} exp(F_{inter_i} \cdot F_{center}j)} \tag{7}$$

t_{ij} denotes the relation between the ith pixel and other pixels in the jth class center. We enhance the feature representations by conducting a matrix multiplication between the relation t_{ij} and the class center representation $F_{center}j$ to obtain the output $F^{\approx}_{inter}$ as outlined below:

$$F^{\sim}_{inter} = \sum_{j=1}^{N}(t_{ij}F_{center}j) \tag{8}$$

$F^{\sim}_{inter}$ represents the inter-class representations. We further combine F, $F^{\approx}_{intra}$ and $F^{\approx}_{inter}$ for final prediction:

$$F_{final} = \Delta(Concat(F, F^{\sim}_{intra}, F^{\sim}_{inter})) \tag{9}$$

where Δ denotes a convolutional layer utilizing a 1×1 filter to decrease the dimension of output channels.

3.2.3. Loss function

We employ the binary cross-entropy loss to supervise the training of the affinity map as specified in Section 3.2.2, considering the binary classification nature of each pixel. Additionally, we introduce an auxiliary loss L_{au} to guide the training of the class center, which utilizes pixel-wise cross-entropy. Ultimately, we use another pixel-wise cross-entropy loss L_{seg} to predict the final segmentation. Thus, the overall loss is specified as outlined below.

$$L = \lambda_1 L_{seg} + \lambda_2 L_{aff} + \lambda_3 L_{au} \tag{10}$$

where λ_1, λ_2, and λ_3 represent the weights used to balance the loss of L_{seg}, L_{aff}, and L_{au}. We set these as $\lambda_1 = 1$, $\lambda_2 = 1$, and $\lambda_3 = 0.4$.

3.3. Integrating with other methods

Our CAA module can be effortlessly incorporated into the existing segmentation frameworks such as FCN [4], PSPNet [8], DeeplabV3 [15], UperNet [36], and ISNet [12]), yielding notable performance enhancements. These works either focus on pixel-to-pixel (spatial) relation or pixel-to-class (class-level) association, while our CAA can enhance both two correlations.

To construct the affinity map, we need to aggregate more spatial information. Specifically, we utilize existing mainstream frameworks to provide feature extraction *i.e.*, dilated convolutions [14,16] or pyramid pooling [8,37,38]. We further create an affinity map to differentiate if two pixels belong to the same class (intra-semantic dependency) or different classes (inter-semantic dependency). We build the class center with coarse segmentation results to calculate consistency between pixels and each class center.

4. Experiment

We evaluate the effectiveness of our approach on three distinct semantic segmentation benchmarks: ADE20K [39], COCO-Stuff dataset [40], and Pascal-Context dataset [41].

4.1. Datasets

- ADE20K. ADE20K is a challenging semantic segmentation dataset, composed of more than 20,000 images. The dataset contains 150 semantic categories, *i.e.*, sky, road, grass as well as discrete objects such as people, cars, and beds. It is partitioned into 20k/2k/3k images for training, validation, and testing respectively.
- COCO-Stuff. The COCO-Stuff10k dataset consists of 171 semantic classes, comprising 81 thing classes and 91 stuff classes. The training set comprises 9,000 images, while the testing set comprises 1,000 images.
- PASCAL-Context. The PASCAL-Context dataset encompasses 59 semantic pixel-level categories in all its training images. This includes 4,998 images designated for training and 5,105 images allocated for testing.

Table 1
Ablation study conducted on the ADE20K validation set. "CC" denotes the use of only Class Center, while "SA" denotes the use of only Semantic Affinity. All methods employ a single scale for testing.

Baseline	CC	SA	Backbone	mIoU (%)
✓			ResNet50	36.10
✓	✓		ResNet50	40.39
✓		✓	ResNet50	42.43
✓	✓	✓	ResNet50	43.12

4.2. Implementation details

We utilize Pytorch and MMSegmentation [42] toolbox to conduct the experiment. Then, we utilize the ImageNet [43] pre-trained ResNet [44] and transformer *e.g.*, ViT [45] and Swin-Transformer [46] as the backbone. We trained our model on two NVIDIA A40 GPUs, each with 48 GB of memory.

For network optimization, we employ the stochastic gradient descent (SGD) algorithm with a momentum value of 0.9 and the "poly" learning rate policy with a factor of $(1 - \frac{iter}{iter_{max}})^{0.9}$. Additionally, we incorporate Synchronized batch normalization (SyncBN) during the model training process. Following the approach of prior studies [10, 23,47], we utilize a multi-scale ratio ranging from 0.5 to 1.75 and apply data augmentation techniques like flipping and random cropping. In addition, we adopt the mean intersection of union (mIoU) as the evaluation metrics. More specific settings are introduced for different benchmarks (see Table 1).

- ADE20K: For ADE20K, the initial learning rate is 0.02, crop size is 512 × 512, and weight decay is 0.0005. If not specified, we set 160k training iterations with batch size 16.
- COCO-Stuff: For COCO-Stuff, the initial learning rate is 0.001, crop size is 512 × 512, and weight decay is 0.0001. If not specified, we set 60k training iterations with batch size 16.
- PASCAL-Context: For PASCAL-Context, the initial learning rate is 0.001, crop size is 512 × 512, and weight decay is 0.0001. If not specified, we set 60k training iterations with batch size 16.

4.3. Ablation study

4.3.1. For class center and semantic affinity

We conduct ablation experiments on the ADE20K val set to verify the efficacy of our CAA. Then, we choose FCNs [4] as our baseline, selecting ResNet-50 as the backbone due to its relatively lower computational burden compared to models like ResNet-101, ViT, and Swin-Transformer. To better show the importance of the class center and semantic affinity, we separate our CAA into Class Center (CC) and Semantic Affinity (SA), respectively.

Class Center. To mitigate the problem of category confusion, class-level context is very important. A car barely appears in the sky, so we need pixel-to-class association to predict a clear class for an object. As indicated in Table 1, the class center improves the performance significantly. In comparison to the baseline FCN (ResNet-50), it achieves a 40.39% mIoU on the ADE20K validation set, marking a notable improvement of 4.29%. Besides, we show the visualization segmentation results of intra-class representations in Fig. 6. Likewise, it is justifiable to investigate the impact of pixel-to-pixel relationships.

Semantic Affinity. The objective of semantic segmentation is to allocate a semantic label to each individual pixel. Hence, pixel-level context plays a critical role in achieving superior performance. As indicated in Table 1, it is evident that the pixel-to-pixel relation yields a noteworthy enhancement of 6.33% mIoU on the ADE20K dataset in contrast to the baseline method, demonstrating that different pixels need different contextual dependencies. We also show the visualization segmentation results of inter-class representations in Fig. 7.

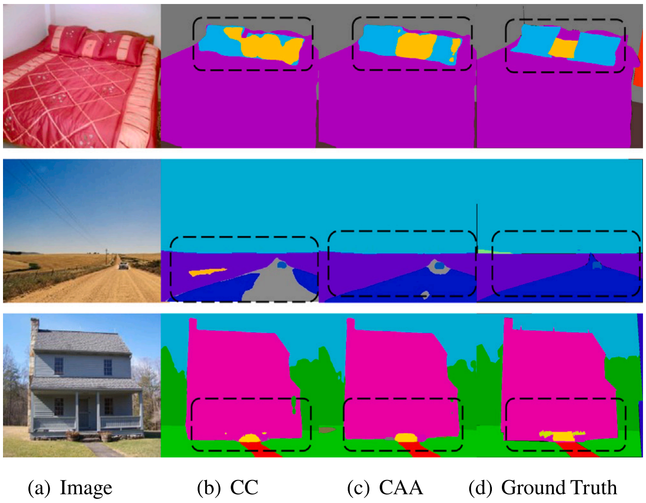


Fig. 6. Illustration of segmented examples of CAA and CC.

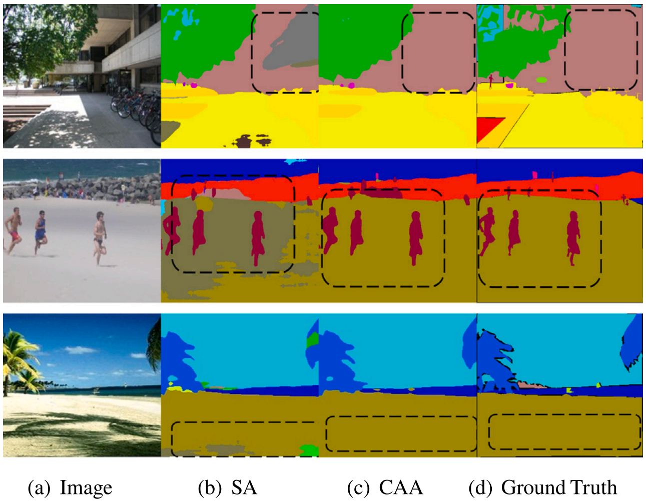


Fig. 7. Illustration of segmented examples of CAA and SA.

Table 2
Ablation study conducted on the ADE20K validation set. All methods employ a single scale for testing.

Baseline	Intra-class	Inter-class	Backbone	mIoU(%)
✓			ResNet50	36.10
✓	✓		ResNet50	42.76
✓		✓	ResNet50	42.39
✓	✓	✓	ResNet50	43.12

Class-Aware Affinity. Existing methods do not explicitly model the pixel-to-pixel relation and pixel-to-class association. We propose to combine these two dependencies to aggregate the contextual information. As illustrated in Table 1, the CAA outperforms the baseline method by 7.02% mIoU on the ADE20K dataset. To further substantiate the effectiveness, we show the visualization of qualitative results in Figs. 6 and 7. From the results, we can indicate that pixel-to-pixel and pixel-to-class relations are complementary.

4.3.2. For single class association

We can extract intra- and inter-class pixel representations from our semantic affinity. Experiments show the effectiveness of the two-pixel representations.

Table 3
Ablation study of loss function of the ADE20K validation set. Aux represents only using auxiliary loss and the final cross-entropy loss. Aff represents only using affinity loss and the final cross-entropy loss.

CAA	Aux	Aff	Backbone	mIoU(%)
✓			ResNet50	40.45
✓	✓		ResNet50	41.44
✓		✓	ResNet50	42.36
✓	✓	✓	ResNet50	43.12

Table 4
Comparison of computational complexity and accuracy on ADE20K.

Method	Params(M)	FLOPs(G)	mIoU (%)
PSPNet [8]	23.14	77.71	42.48
CCNet [13]	24.00	99.65	42.08
Deeplabv3 [15]	42.28	168.91	42.66
UperNet [36]	40.58	179.75	42.05
PSPNet [8] + CAA	26.75	90.76	44.12
CCNet [13] + CAA	27.60	112.69	43.98
Deeplabv3 [15] + CAA	45.88	181.96	44.23
UperNet [36] + CAA	44.18	192.48	43.81

Intra-class contextual information. Since each pixel has a semantic label, the intra-class contextual information is vital for semantic image segmentation. As shown in Table 2, the intra-class representations lead to a substantial performance boost. In comparison to the baseline FCN (ResNet-50), the intra-class representations achieve a mIoU of 42.76% on the ADE20K validation set, marking a notable improvement of 6.66%. Similarly, it is reasonable to explore the effect of inter-class representations.

Inter-class contextual information. The inter-class contextual information could differentiate the pixels in different classes. As shown in Table 2, we can see that inter-class representations bring an improvement of 6.29% mIoU on ADE20K, demonstrating that different pixels need different contextual dependencies. The performance of the inter-class is inferior to the intra-class so we can conclude that intra-class representations are more important.

Intra+Inter-class contextual information. We further combine the intra- and inter-class representations. As depicted in Table 2, CAA outperforms the baseline method by 7.02% mIoU on the ADE20K dataset. From the results, we can indicate that the intra- and inter-class representations are complementary and promoted by each other. In a word, it is necessary to synchronously capture intra- and inter-class contexts to make them promote each other.

4.3.3. For loss function

The CAA module leverages both an affinity loss and an auxiliary loss to boost segmentation performance. We explore the significance of these two losses in this section. The results are presented in Table 3. Only using auxiliary loss and the final cross-entropy loss, our CAA module attains a mIoU of 41.44%. On the contrary, only utilizing the affinity loss and the ultimate cross-entropy loss leads to a mIoU of 42.36%. When leveraging both affinity loss and auxiliary loss, we achieve a mIoU of 43.12%, indicating that both the affinity map and class context need to be supervised.

4.3.4. Computational complexity and training time

We report the computational complexity of our CAA integrated with existing segmentation frameworks in Table 4, including the increased parameters, computation complexity, and accuracy. The results are calculated based on input size 512 × 64 × 64 (8 times down-sampled from 512 × 512). Since CAA aims to aggregate the contextual information by modeling pixel-to-pixel relation and pixel-to-class association, CAA is complementary to the existing segmentation frameworks. After integrating CAA with existing segmentation schemes, the whole model complexity is still acceptable. As an illustration, when incorporating CAA with PSPNet, the parameters and FLOPs experience a mere increase of 3.61M and 23.05G, respectively. However, the segmentation performance improves from 42.48% to 44.12% on ADE20k.

Besides, we also report the time of training 6k iterations and inference time per image on Pascal-Context testing set in Table 6. There is no doubt that the training and inference time is increased because of the additional processing carried out by the proposed CAA. However, the clear value is that additional CAA provides clear performance boosting in terms of mIoU.

4.3.5. Integrated with various semantic segmentation frameworks

Our CAA module seamlessly integrates with pre-existing segmentation frameworks. To affirm the effectiveness and robustness of our approach, we incorporate the proposed CAA into five existing segmentation frameworks, namely FCN, PSPNet, UperNet, DeepLabV3, and ISNet. The performance comparison across three benchmark datasets: ADE20k, COCO-Stuff, and Pascal-Context, is presented in Table 5. We conduct single-scale testing to ensure a fair comparison. In the case of the vanilla segmentation framework (i.e., FCN), integrating CAA leads to a remarkable improvement of 7.02% in mIoU on ADE20K, 4.37% on COCO-Stuff, and 5.06% on Pascal-Context. With a stronger baseline like DeeplabV3, CAA elevates the mIoU performance from 42.66% to 44.23% on ADE20K, from 35.73% to 37.39% on COCO-Stuff, and from 50.73% to 52.37% on Pascal-Context. This performance comparison highlights the flexibility and applicability of the CAA module.

4.4. Comparison with other methods

We perform experiments on diverse benchmarks and compare the performance with other methods in this section.

4.4.1. Results of ADE20K

In Table 7, we present a comparison of our results with those of other competitive methods on the ADE20K dataset. ADE20K is a comprehensive scene-parsing dataset, comprising a total of 150 semantic classes. We maintain consistent training and testing configurations for this dataset. To validate the efficacy of CAA, we conduct experiments and compare the results with various methods. The previous best method ISNet [12] utilizes semantic-level and image-level context for semantic segmentation, achieving 47.31% mIoU. Our UperNet+CAA achieves 47.45% mIoU with ResNet-101 that is 2.17%, 1.18% and 0.14% higher than OCRNet [23], CPNet [24], and ISNet [12] respectively. Owing to the effectiveness of the proposed CAA, our UperNet+CAA achieves 51.86% mIoU by adopting InternImage-B. Furthermore, employing the Swin-Large backbone, our UperNet+CAA achieves a new state-of-the-art performance with a mIoU of 54.59%.

Moreover, to validate the efficacy of CAA, we provide visualizations of qualitative results obtained from the val set of ADE20K (refer to Fig. 8). The figure illustrates that our UperNet+Swin+CAA effectively handles issues related to intra-class compactness and inter-class dispersion, resulting in superior segmentation outcomes compared to the baseline UperNet+Swin.

4.4.2. Results of COCO-Stuff

COCO-Stuff is a challenging benchmark, including 81 thing classes (objects with distinct shapes, *e.g.*, car and person) and 91 stuff classes (amorphous background regions, *e.g.*, grass and sky). Table 7 compares the results with CCL [29], DANet [10], OCRNet [23], ISNet [12], etc. When utilizing ResNet-101 as a pre-trained network, our UperNet+CAA achieves 41.00% mIoU, exceeding OCRNet [47] and DANet [10] with a mIoU of 1.5% and 1.3%, respectively. When leveraging a more robust backbone InternImage-B, our method achieves mIoU of 46.64% mIoU. We have also integrated CAA into the widely used Swin-Large backbone and UperNet+CAA obtains 49.96% mIoU. These results prove the effectiveness of our CAA.

Table 5
We evaluate the performance of integrating CAA into various mainstream frameworks across different benchmarks.

Method	Backbone	ADE20K	COCO-Stuff	Pascal-Context
FCN	ResNet50	36.10	30.86	45.76
FCN+CAA	ResNet50	43.12 (+7.02)	35.23 (+4.37)	50.82 (+5.06)
CCNet	ResNet50	42.08	35.43	47.93
CCNet+CAA	ResNet50	43.98 (+1.9)	36.27(+0.84)	49.79 (+1.86)
UperNet	ResNet50	42.05	35.80	48.90
UperNet+CAA	ResNet50	43.81 (+1.76)	37.08 (+1.28)	51.57 (+2.67)
PSPNet	ResNet50	42.48	36.33	50.21
PSPNet+CAA	ResNet50	44.12 (+1.64)	37.59 (+1.26)	52.05 (+1.84)
Deeplabv3	ResNet50	42.66	35.73	50.73
Deeplabv3+CAA	ResNet50	44.23 (+1.57)	37.39 (+1.66)	52.37 (+1.64)
ISNet	ResNet50	43.77	38.06	51.74
ISNet+CAA	ResNet50	43.98 (+0.21)	38.35 (+0.29)	52.08 (+0.34)

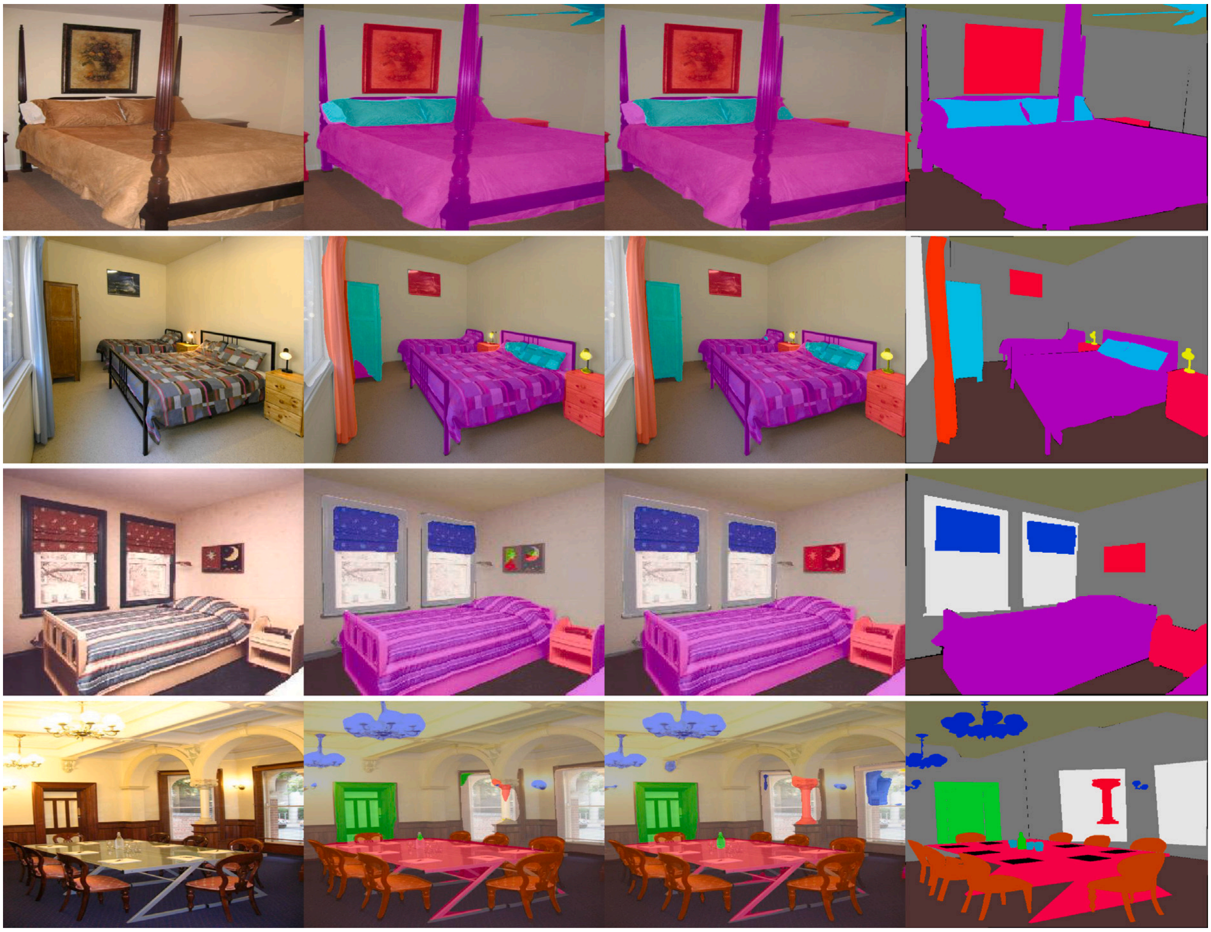

(a) Image (b) UperNet+Swin (c) UperNet+Swin+CAA (d) Ground Truth

Fig. 8. Visualizations on the ADE20K validation set. We compare the qualitative results of UperNet+Swin and our UperNet+Swin+CAA.

4.4.3. Results of PASCAL-Context

We also perform experiments on the PASCAL-Context to conduct a comparative analysis with existing methods.

Table 7 showcases the segmentation outcomes. With ResNet-101 as the pre-trained network, our UperNet+CAA attains a mIoU of 55.57%. We exceed PSPNet [8], DANet [10] and CPNet [24] with a mIoU of 2.07%, 2.97%, and 1.67%. By employing InternImage-B, UperNet+CAA achieves an impressive mIoU of 62.08%. Furthermore, our approach, UperNet+CAA with Swin-Large, establishes a mIoU of 64.38%.

In general, our UperNet+CAA attains the best result on ADE20K and Pascal-Context. Although we do not reach the best result with ResNet101 on the COCO-Stuff dataset compared with ISNet [12] and PCAA [52], we still rank within the top-3. This achievement showcases the competitiveness and effectiveness of our approach across multiple benchmarks. By achieving top rankings across various datasets, our model demonstrates its robustness and adaptability in addressing diverse visual understanding tasks.

5. Conclusion

This paper presents a new approach to tackle context aggregation in semantic segmentation. We propose to extract the pixel-to-pixel relation and pixel-to-class association. A class-aware affinity module (CAA) is set to learn class center and semantic affinity. We use the

Table 6
Comparison of training time and inference time on Pascal-Context. ms/p represents the time of inferencing a picture.

Method	Training(min)	Inference(ms/p)	mIoU
FCN [4]	61	59.6	45.76
PSPNet [8]	69	57.6	50.21
Deeplabv3 [15]	91	67.0	50.73
UperNet [36]	44	66.3	48.90
FCN [4] +CAA	75(↑23%)	68.6(↑15%)	50.82
PSPNet [8] + CAA	91(↑32%)	72.6(↑26%)	52.05
Deeplabv3 [15] + CAA	114(↑25%)	79.3(↑18%)	52.37
UperNet [36] + CAA	71(↑61%)	78.2(↑18%)	51.57

Table 7
Comparisons with the state-of-the-art methods. We employ a multi-scale and flipped testing approach to compare segmentation performance on the validation set of ADE20K, testing set of COCO-Stuff, and testing set of Pascal-Context. We utilize mIoU as the evaluation metric and the best performance is in bold.

Method	Backbone	ADE20K	COCO-Stuff	Pascal-Context
CCL [29]	ResNet101	–	35.70	51.60
PSPNet [8]	ResNet101	43.29	38.86	53.50
PSANet [20]	ResNet101	43.77	–	–
EncNet [9]	ResNet101	44.65	–	51.70
CFNet [48]	ResNet101	44.89	–	54.00
DANet [10]	ResNet101	45.22	39.70	52.60
OCRNet [23]	ResNet101	45.28	39.50	54.37
OCNet [47]	ResNet101	45.40	39.10	54.00
SpyGR [49]	ResNet101	–	39.90	52.80
CTNet [35]	ResNet101	45.94	–	55.50
RANet [50]	ResNet101	–	40.70	54.90
SPNet [11]	ResNet101	45.60	–	54.50
ACNet [30]	ResNet101	45.90	40.10	54.10
CPNet [24]	ResNet101	46.27	–	53.90
FLANet [51]	ResNet101	46.68	–	–
PCAA [52]	ResNet101	46.74	–	55.60
ISNet [12]	ResNet101	47.31	41.60	–
UperNet [36]	Swin-Large	52.10	48.60	60.30
UperNet+CAR [53]	Swin-Large	–	44.88	58.97
GSS-FT-W [54]	Swin-Large	48.54	–	–
SegDeformer [55]	Swin-Large	53.90	49.51	–
TSG [56]	Swin-Large	54.2	–	63.3
UperNet+CAC [57]	Swin-Large	54.43	–	–
UperNet [58]	InternImage-B	51.30	45.30	61.34
UperNet+CAA (*ours*)	ResNet101	47.45	41.00	55.57
UperNet+CAA (*ours*)	ViT-Large	50.35	45.64	61.11
UperNet+CAA (ours)	InternImage-B	51.86	46.64	62.08
UperNet+CAA (*ours*)	Swin-Large	**54.59**	**49.96**	**64.38**

semantic affinity map to discern whether two pixels belong to the same class, subsequently generating intra-class and inter-class pixel representations. Furthermore, we incorporate the pixel's class association into the calculation of the respective class center and aggregate both intra- and inter-class pixel representations. At last, CAA can be an additional supplementary module to the existing mainstream segmentation framework. Comprehensive experiments verify the effectiveness of our approach. Our CAA achieves mIoU of 54.59% on ADE20k, 49.96% on COCO-Stuff, and 64.38% on Pascal-Context.

6. Future work

This study has addressed pixel-to-pixel relation and pixel-to-class association for semantic segmentation within individual images. Recent works [59,60] propose to mine cross-image contextual information for improving the performance of semantic segmentation. Inspired by this, we plan to extend the idea of CAA to explore potential cross-image pixel-to-pixel relations and pixel-to-class association to further enhance contextual information.

CRediT authorship contribution statement

Huadong Tang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization. **Youpeng Zhao:** Writing – review & editing. **Chaofan Du:** Writing – review & editing. **Min Xu:** Writing – review & editing, Supervision. **Qiang Wu:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[2] B.-k. Chen, C. Gong, J. Yang, Importance-aware semantic segmentation for autonomous driving system., in: IJCAI, 2017, pp. 1504–1510.
[3] M. Fawakherji, A. Youssef, D. Bloisi, A. Pretto, D. Nardi, Crop and weeds classification for precision agriculture using context-independent pixel-wise segmentation, in: 2019 Third IEEE International Conference on Robotic Computing, IRC, IEEE, 2019, pp. 146–152.
[4] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[5] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al., Deep high-resolution representation learning for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2020) 3349–3364.
[6] F. Yu, V. Koltun, T. Funkhouser, Dilated residual networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 472–480.
[7] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, et al., Resnest: Split-attention networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2736–2746.
[8] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.
[9] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, A. Agrawal, Context encoding for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7151–7160.
[10] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.
[11] Q. Hou, L. Zhang, M.-M. Cheng, J. Feng, Strip pooling: Rethinking spatial pooling for scene parsing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4003–4012.
[12] Z. Jin, B. Liu, Q. Chu, N. Yu, Isnet: Integrate image-level and semantic-level context for semantic segmentation, in: ICCV, 2021, pp. 7189–7198.
[13] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, Ccnet: Criss-cross attention for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 603–612.
[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.
[15] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, arXiv preprint arXiv:1706.05587.
[16] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 801–818.
[17] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[18] B. Zhan, E. Song, H. Liu, X. Xu, W. Li, C.-C. Hung, Segmenting medical images via explicit–implicit attention aggregation, Knowl.-Based Syst. 279 (2023) 110932.
[19] J. Chen, Y. Chen, W. Li, G. Ning, M. Tong, A. Hilton, Channel and spatial attention based deep object co-segmentation, Knowl.-Based Syst. 211 (2021) 106550.

[20] H. Zhao, Y. Zhang, S. Liu, J. Shi, C.C. Loy, D. Lin, J. Jia, Psanet: Point-wise spatial attention network for scene parsing, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 267–283.

[21] S. Bai, C. Wang, Information aggregation and fusion in deep neural networks for object interaction exploration for semantic segmentation, Knowl.-Based Syst. 218 (2021) 106843.

[22] F. Zhang, Y. Chen, Z. Li, Z. Hong, J. Liu, F. Ma, J. Han, E. Ding, Acfnet: Attentional class feature network for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6798–6807.

[23] Y. Yuan, X. Chen, J. Wang, Object-contextual representations for semantic segmentation, in: European Conference on Computer Vision, Springer, 2020, pp. 173–190.

[24] C. Yu, J. Wang, C. Gao, G. Yu, C. Shen, N. Sang, Context prior for scene segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12416–12425.

[25] Z. Liu, L. Meng, Y. Tan, J. Zhang, H. Zhang, Image compression based on octave convolution and semantic segmentation, Knowl.-Based Syst. 228 (2021) 107254.

[26] C. Gao, H. Ye, F. Cao, C. Wen, Q. Zhang, F. Zhang, Multiscale fused network with additive channel–spatial attention for image segmentation, Knowl.-Based Syst. 214 (2021) 106754.

[27] X. Liu, L. Jiao, L. Li, X. Tang, Y. Guo, Deep multi-level fusion network for multi-source image pixel-wise classification, Knowl.-Based Syst. 221 (2021) 106921.

[28] H. Tang, Y. Zhao, Y. Jiang, Z. Gan, Q. Wu, Class-aware contextual information for semantic segmentation, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2023, pp. 1–5.

[29] H. Ding, X. Jiang, B. Shuai, A.Q. Liu, G. Wang, Context contrasted feature and gated multi-scale aggregation for scene segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2393–2402.

[30] J. Fu, J. Liu, Y. Wang, Y. Li, Y. Bao, J. Tang, H. Lu, Adaptive context network for scene parsing, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6748–6757.

[31] Z. Lin, M. Feng, C.N.d. Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, A structured self-attentive sentence embedding, 2017, arXiv preprint arXiv:1703.03130.

[32] A. Galassi, M. Lippi, P. Torroni, Attention in natural language processing, IEEE Trans. Neural Netw. Learn. Syst. 32 (10) (2020) 4291–4308.

[33] B. Lu, Q. Hu, Y. Wang, G. Hu, RCANet: Row-column attention network for semantic segmentation, in: ICASSP, IEEE, 2022, pp. 2604–2608.

[34] Z. Zhong, Z.Q. Lin, R. Bidart, X. Hu, I.B. Daya, Z. Li, W.-S. Zheng, J. Li, A. Wong, Squeeze-and-attention networks for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13065–13074.

[35] Z. Li, Y. Sun, L. Zhang, J. Tang, Ctnet: Context-based tandem network for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. (2021).

[36] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, J. Sun, Unified perceptual parsing for scene understanding, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 418–434.

[37] J. He, Z. Deng, L. Zhou, Y. Wang, Y. Qiao, Adaptive pyramid context network for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7519–7528.

[38] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[39] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Scene parsing through ade20k dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 633–641.

[40] H. Caesar, J. Uijlings, V. Ferrari, Coco-stuff: Thing and stuff classes in context, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1209–1218.

[41] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, A. Yuille, The role of context for object detection and semantic segmentation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 891–898.

[42] M. Contributors, MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark, 2020, https://github.com/open-mmlab/mmsegmentation.

[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[44] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[45] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[46] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[47] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, J. Wang, Ocnet: Object context network for scene parsing, 2018, arXiv preprint arXiv:1809.00916.

[48] H. Zhang, H. Zhang, C. Wang, J. Xie, Co-occurrent features in semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 548–557.

[49] X. Li, Y. Yang, Q. Zhao, T. Shen, Z. Lin, H. Liu, Spatial pyramid based graph reasoning for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8950–8959.

[50] D. Shen, Y. Ji, P. Li, Y. Wang, D. Lin, Ranet: Region attention network for semantic segmentation, NIPS 33 (2020) 13927–13938.

[51] Q. Song, J. Li, C. Li, H. Guo, R. Huang, Fully attentional network for semantic segmentation, in: AAAI, vol. 36, (2) 2022, pp. 2280–2288.

[52] S.-A. Liu, H. Xie, H. Xu, Y. Zhang, Q. Tian, Partial class activation attention for semantic segmentation, in: CVPR, 2022, pp. 16836–16845.

[53] Y. Huang, D. Kang, L. Chen, X. Zhe, W. Jia, L. Bao, X. He, Car: Class-aware regularizations for semantic segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 518–534.

[54] J. Chen, J. Lu, X. Zhu, L. Zhang, Generative semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7111–7120.

[55] B. Shi, D. Jiang, X. Zhang, H. Li, W. Dai, J. Zou, H. Xiong, Q. Tian, A transformer-based decoder for semantic segmentation with multi-level context mining, in: European Conference on Computer Vision, Springer, 2022, pp. 624–639.

[56] H. Shi, M. Hayat, J. Cai, Transformer scale gate for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3051–3060.

[57] Z. Tian, J. Cui, L. Jiang, X. Qi, X. Lai, Y. Chen, S. Liu, J. Jia, Learning context-aware classifier for semantic segmentation, 2023, arXiv preprint arXiv:2303.11633.

[58] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, et al., Internimage: Exploring large-scale vision foundation models with deformable convolutions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14408–14419.

[59] Z. Jin, T. Gong, D. Yu, Q. Chu, J. Wang, C. Wang, J. Shao, Mining contextual information beyond image for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7231–7241.

[60] Z. Jin, D. Yu, Z. Yuan, L. Yu, Mcibi++: Soft mining contextual information beyond image for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 45 (5) (2022) 5988–6005.