# Contextrast: Contextual Contrastive Learning for Semantic Segmentation

Changki Sung[1], Wanhee Kim[2]*, Jungho An[2]*, Wooju Lee[1], Hyungtae Lim[1]†, and Hyun Myung[1]†
[1]School of Electrical Engineering, KI-Robotics,
Korea Advanced Institute of Science and Technology, Republic of Korea
[2]Department of Automotive Engineering, Kookmin University, Republic of Korea
[1]{cs1032, dnwn24, shapelim, hmyung}@kaist.ac.kr [2]{gml78905, ajh427}@kookmin.ac.kr

## Abstract

*Despite great improvements in semantic segmentation, challenges persist because of the lack of local/global contexts and the relationship between them. In this paper, we propose Contextrast, a contrastive learning-based semantic segmentation method that allows to capture local/global contexts and comprehend their relationships. Our proposed method comprises two parts: a) contextual contrastive learning (CCL) and b) boundary-aware negative (BANE) sampling. Contextual contrastive learning obtains local/global context from multi-scale feature aggregation and inter/intra-relationship of features for better discrimination capabilities. Meanwhile, BANE sampling selects embedding features along the boundaries of incorrectly predicted regions to employ them as harder negative samples on our contrastive learning, resolving segmentation issues along the boundary region by exploiting fine-grained details. We demonstrate that our Contextrast substantially enhances the performance of semantic segmentation networks, outperforming state-of-the-art contrastive learning approaches on diverse public datasets, e.g. Cityscapes, CamVid, PASCAL-C, COCO-Stuff, and ADE20K, without an increase in computational cost during inference.*

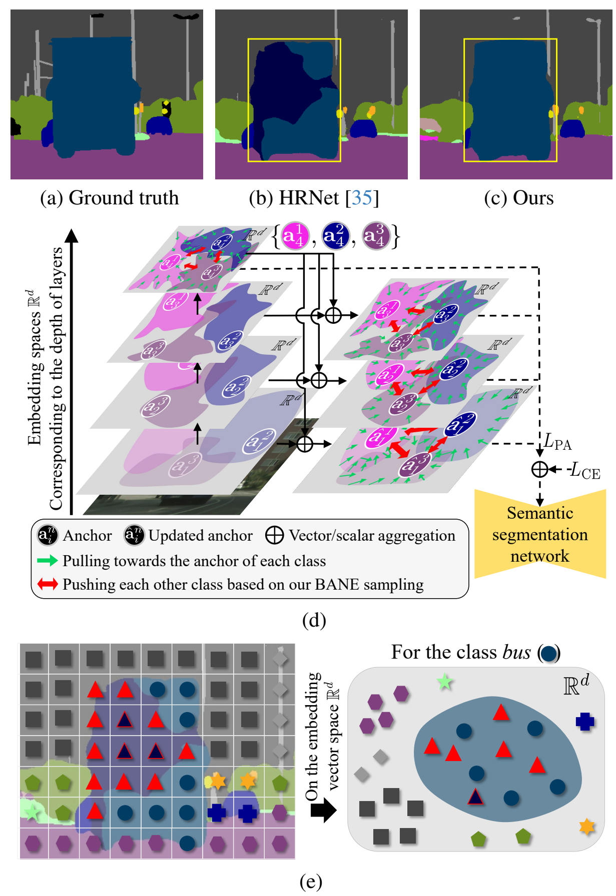


Figure 1. (a) Ground truth, (b) output of HRNet [35], (c) and that of ours. (d) Overview of our contextual contrastive learning (CCL): the representative anchors of the last layer, which are from the higher embedding space levels, are aggregated to representative anchors of the lower layer to encapsulate local and global context. By doing so, the anchor of the $n$-th class on the $i$-th layer $\mathbf{a}_i^n$ is updated as $\hat{\mathbf{a}}_i^n$ (on the right side, its position is shifted), enhancing the distinctiveness between anchors of each class. (e) Visual description of our boundary-aware negative (BANE) sampling (triangles with red color and red borders). Our sampling prioritizes selecting the features of incorrect predictions at the edges (red triangles) rather than those inside the regions (triangles with red borders) as negative samples. Each shape represents an embedding vector derived from the respective class (best viewed in color).

## 1. Introduction

Semantic segmentation is a fundamental technique utilized across diverse applications, including autonomous driving, medical imaging, and robotics [12, 13, 18, 32, 37, 40]. Recent empirical studies have achieved remarkable advancements in semantic segmentation, benefiting significantly from the availability of extensive datasets [1, 2, 9, 10, 29, 55]. To improve segmentation performance, researchers have proposed larger deep neural network (DNN) architectures [4–8, 11, 14, 16, 19, 21, 24, 25, 27, 34, 35, 39, 42–

*Work done during internship at KAIST
†Corresponding authors: Dr. Hyungtae Lim and Prof. Hyun Myung

| | Complexity | Scale | Boundary awareness |
|---|---|---|---|
| *Baseline* | - | - | - |
| *ICCV 21 [41]* | CL, MB | Single | - |
| *ICCV 21 [17]* | CL, MB, AL | Single | - |
| *ECCV 22 [31]* | CL | Multi | - |
| *Ours* | CL | Multi fusion | Aware |

Table 1. Comparison between SOTAs and ours (CL: contrastive learning, MB: memory bank, AL: additional loss)

44, 46–49, 51–54, 56] and novel loss functions [36, 38, 50].

Despite these achievements, semantic segmentation sometimes produces inaccurate segmentation, as illustrated in Fig. 1(b). It could be solved by increasing the complexity of networks [25, 27, 34, 42, 43]; however, these approaches require more memory and may potentially slow down the inference speed. Therefore, it is necessary to improve performance without any additional neural network modules for efficiency.

For these reasons, another noteworthy method, contrastive learning, has emerged as a valuable solution [17, 41] because contrastive learning aims to make the networks understand the semantic context better during the training process. This is achieved by attaching refinement modules during the training stage and detaching them on the inference so that the network models preserve the inference speed without increasing the complexity of architectures.

However, previous researches [17, 41] have overlooked the significance of multi-scale features, including both global and local contexts. To mitigate the problem, Pissas *et al*. [31] proposed a method to extract multi-scale embedding features from multiple encoder layers. However, the method could not consistently comprehend the relationships between different scales of features because multi-scale and cross-scale contrastive learning are considered independently. Consequently, it struggles to comprehend relationships between local and global contexts.

To address the aforementioned problems, we propose a supervised contrastive learning framework incorporating two novel methods for semantic segmentation, called *Contextrast*. First, contextual contrastive learning (CCL) is proposed to acquire embedded features from multiple encoder layers representing local and global contexts. Based on embedded features, we define the representative anchors in each layer, which act as the semantic centroids for each class. The anchors of the last layer represent more global context than the anchors in the lower layers. The anchors of the last layer are used to update anchors in each layer (green arrows in Fig. 1(d)). Thus, the anchors in the lower layers can have both global and local contexts. Consequently, it consistently understands relationships between local and global contexts using the updated representative anchors, which share the same global contexts. Second, boundary-aware negative (BANE) sampling, which is inspired by [50] and [38] that focus on the boundary regions, is proposed to sample negative examples along the boundaries of incorrectly predicted regions (Fig. 1(e)). It leverages the advantages of sampling harder negative examples and capturing fine-grained details, so the proposed method gets more informative gradients during the training process [20, 41]. We summarize several key properties of state-of-the-art methods and ours, as shown in Table 1.

In sum, this paper makes the following contributions:

- Our Contextrast enables a segmentation model to capture global/local context information from multi-scale features and consistently comprehend relationships between them through the representative anchors.
- Our BANE sampling enables the acquisition of informative negative samples for contrastive learning and fine-grained details. It guides the model to focus on confusion regions progressively during the training.
- To demonstrate the applicability of our Contextrast in semantic segmentation, we verify the state-of-the-art performance for contrastive learning-based semantic segmentation on various powerful CNN models [6, 35, 49] and public datasets: Cityscapes [9], CamVid [1], PASCAL-C [29], COCO-Stuff [2], and ADE20K [55], which were acquired in different domains.

## 2. Related works

### 2.1. Semantic segmentation

Semantic segmentation, a fundamental task in computer vision, entails pixel-wise object classification within an image. In recent years, remarkable advancements in deep learning have propelled the field of semantic segmentation to unprecedented levels of accuracy and efficiency. At first, fully connected networks (FCNs) [28] brought significant progress in semantic segmentation by introducing end-to-end dense feature learning. However, FCNs suffer from limited spatial and contextual information because of the narrow local receptive fields.

Thus, the following researchers focused on capturing better spatial and context information in the semantic segmentation. Atrous spatial pyramid pooling (ASPP) [6] captures a diverse range of contextual information. HRNet [35] maintains high-resolution representation throughout the network, ensuring the preservation of fine-grained details. For further improvements, OCRNet [49] architecture was introduced that integrates object-contextual representations, allowing the network to consider relationships between objects within a scene. As these advanced methods learn discrimination ability using contextual information within an individual image, there are limitations on the capability of global feature discrimination.

## 2.2. Contrastive learning for semantic segmentation

Contrastive learning is a feature learning criterion that aims to minimize the distance between intra-class features while maximizing the distance between inter-class features. Recent advancements in contrastive learning with semantic segmentation [17, 22, 31, 41] have demonstrated impressive performances.

Hu *et al*. [17] and Wang *et al*. [41] proposed novel methods for semantic segmentation in a fully supervised setting that explores global pixel relations, extracting features from multiple images to regularize segmentation embedding space globally. Wang *et al*. [41] introduced memory banks and segmentation-aware negative sampling methods, storing massive data to train distinctive representations and getting more gradient contributions during the training. Hu *et al*. [17] introduced class-wise weighted region centers that were generated with positive samples, to be utilized as the anchors in contrastive learning. However, solely focusing on positive samples for weighting could diminish the discrimination ability of the model. Additionally, [17, 41] have neglected multiple scales of features except for the features of the last layer, so they capture only limited local/global contexts and relationships between local and global contexts.

Finally, Pissas *et al*. [31] proposed a method leveraging the multiple scales of features for supervised contrastive learning. That is, the researchers applied contrastive learning to the multi-scale and cross-scale features. By doing so, the method [31] captures global/local context information from multi-scale features and the relationship between local and global contexts from cross-scale features. Nonetheless, it could not consistently grasp relationships across scales of features because multi-scale and cross-scale contrastive learning are operated separately. In some cases, features can be arranged differently in multi-scale and cross-scale contrastive learning. For example, a feature shifted by multi-scale contrastive learning can be shifted differently in cross-scale contrastive learning.

# 3. Contextrast: Contextual contrastive learning with BANE sampling

## 3.1. Overall framework

As shown in Fig. 2, we propose a supervised contrastive learning framework encompassing two novel methods for semantic segmentation.

First, we propose a concept of *representative anchors*, which are multi-scale-aware salient features implicitly representing the class by leveraging hierarchical design, as described in Sec. 3.2. Second, we deliberately sample the features corresponding to the boundaries within the regions that were incorrectly predicted as the negative samples, as described in Sec. 3.3.

## 3.2. Contextual contrastive learning (CCL)

Let us assume that an encoder consists of a total of $I$ layers. Then, we begin by expressing the representative anchors corresponding to the $i$-th encoder layer as $\mathbf{A}_i$, where $i \in \{1, \cdots, I\}$. $\mathbf{A}_i$ consists of $N$ class-wise representative anchors. Each anchor for the $n$-th class is denoted by $\mathbf{a}_i^n \in \mathbb{R}^d$, which is defined as the average of embedded feature vectors belonging to the ground truth semantic class within the batch images as follows:

$$\mathbf{a}_i^n = \frac{\sum\limits_{\mathbf{v}\in\mathbf{V}_i} \mathbf{v}\mathbb{1}[g(\mathbf{v})=n]}{\sum\limits_{\mathbf{v}\in\mathbf{V}_i} \mathbb{1}[g(\mathbf{v})=n]}, n = 1, 2, ..., N, \tag{1}$$

where $\mathbf{V}_i$ is an embedded feature vector set from the feature of the $i$-th encoder layer's feature map $\mathbf{f} \in \mathbf{F}_i$, as illustrated in Fig. 2, i.e. $\mathbf{v} = \pi(\mathbf{f})$; $g(\cdot)$ represents a function that returns the ground truth semantic label of each embedding feature vector; $\mathbb{1}[\cdot]$ is the Iverson bracket, which outputs one if the condition is satisfied and zero otherwise. By using Eq. (1), $\mathbf{A}_i$ is expressed as $\mathbf{A}_i = \{\mathbf{a}_i^1, \mathbf{a}_i^2, ..., \mathbf{a}_i^N\}$. For convenience, we interchangeably express $\mathbf{A}_i$ in a matrix form, i.e. $\mathbf{A}_i = [\mathbf{a}_i^1\ \mathbf{a}_i^2\ ...\ \mathbf{a}_i^N] \in \mathbb{R}^{d\times N}$.

Then, lower-level anchors $\mathbf{A}_i$ are updated with the representative anchor of the last layer $\mathbf{A}_I$ to encapsulate both high-level and low-level context, accounting for multi-scale. Consequently, the updated representative anchor $\hat{\mathbf{A}}_i$ is defined as $\hat{\mathbf{A}}_i = w_l\mathbf{A}_i + w_h\mathbf{A}_I = \{\hat{\mathbf{a}}_i^1, \hat{\mathbf{a}}_i^2, ..., \hat{\mathbf{a}}_i^N\}$, where $w_l$ and $w_h$ are weight hyperparameters for anchor update (see Fig. 2). By updating the lower-level anchors, $\hat{\mathbf{A}}_i$ can act as a criterion to capture relationships across different scales. For $i = I$, $\hat{\mathbf{A}}_I$ is defined as $\hat{\mathbf{A}}_I = \mathbf{A}_I$.

$$L_{\text{NCE}} = \frac{-1}{|\mathbf{V}_+|}\sum_{\mathbf{v}_+\in\mathbf{V}_+}\log\frac{\exp(\mathbf{v}\cdot\mathbf{v}_+/\tau)}{\exp(\mathbf{v}\cdot\mathbf{v}_+/\tau) + \sum\limits_{\mathbf{v}_-}\exp(\mathbf{v}\cdot\mathbf{v}_-/\tau)}, \tag{2}$$

Next, we incorporate InfoNCE [15, 30] loss in Eq. (2) with $\hat{\mathbf{A}}_i$, which is referred to as the *pixel-anchor (PA) loss*, as follows:

$$L_{\text{PA}} = \sum_{i=1}^{I}\lambda_i\left[\frac{1}{N}\sum_{\hat{\mathbf{a}}_i^n\in\hat{\mathbf{A}}_i}\frac{-1}{|\mathbf{V}_+|}\sum_{\mathbf{v}_+\in\mathbf{V}_+}L_a\right] \tag{3}$$

$$L_a = \log\frac{\exp(\hat{\mathbf{a}}_i^n\cdot\mathbf{v}_+/\tau)}{\exp(\hat{\mathbf{a}}_i^n\cdot\mathbf{v}_+/\tau) + \sum\limits_{\mathbf{v}_-\in\mathbf{V}_-}\exp(\hat{\mathbf{a}}_i^n\cdot\mathbf{v}_-/\tau)} \tag{4}$$

where $\mathbf{v}_{+/-}$ represents positive and negative samples, respectively, and $\lambda_i$ represents the weight hyperparameters

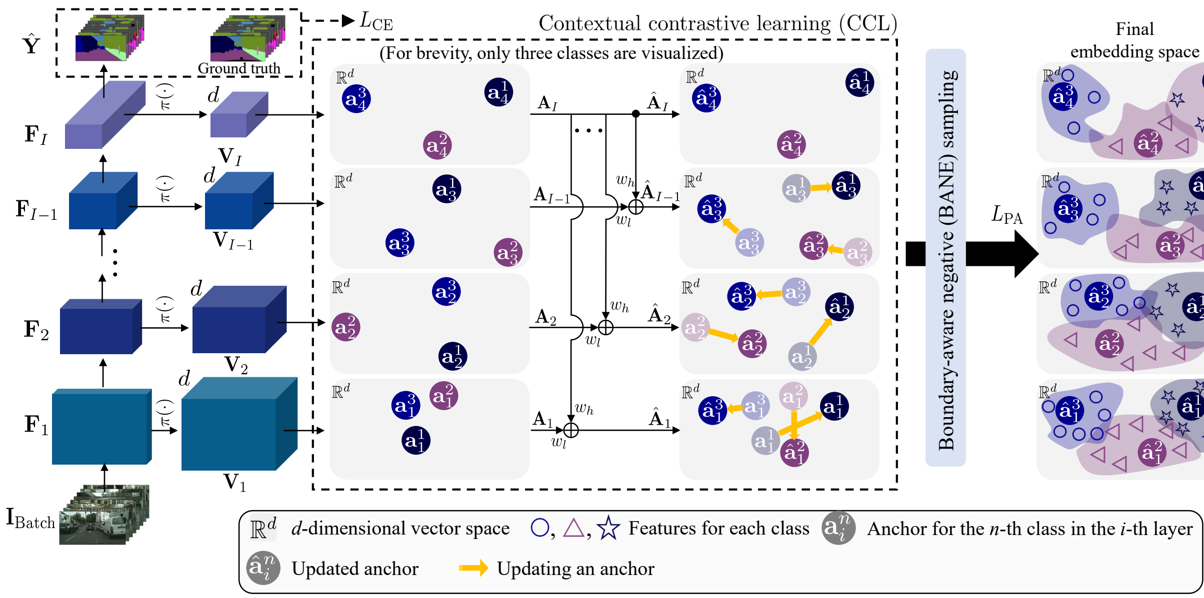


Figure 2. Overall Contextrast framework. Contextrast utilizes the representative anchors updated by the semantically rich representative anchor vector set $\mathbf{A}_I$. Thus, it integrates local/global contexts and their relationships. Then, BANE sampling samples examples that exist along the boundaries of prediction error regions. It samples more informative negative samples and captures fine-grained details for contrastive learning. $\mathbf{I}_{\text{Batch}}$ is the batch images. $\hat{\mathbf{Y}}$ is the prediction outcome from the model. $\mathbf{F}_i$ is the feature map of the $i$-th encoder layer. $\mathbf{V}_i$ is the $i$-th set of the embedded feature vector by the encoding function $\pi(\cdot)$. $\mathbf{A}_i$ denotes the representative anchors of the $i$-th embedded feature vector. The updated representative anchor $\hat{\mathbf{A}}_i$ results from adding low-level and highest-level anchors. $w_h$ and $w_l$ are weight hyperparameters for updating representative anchors. The $L_{\text{PA}}$ is the proposed pixel-anchor loss function. $L_{\text{CE}}$ represents the cross-entropy loss function. Features of each semantic class are illustrated in different shapes and colors (best viewed in color).

assigned to the pixel-anchor contrastive loss for the $i$-th encoder layer. While the anchor is set with individual features in Eq. (2), the anchor is set with a representative anchor $\hat{\mathbf{A}}_i$ in Eq. (4). The pixel-anchor loss aims to optimize embedding features by minimizing the distance between intra-class features and their corresponding representative anchors while maximizing the separation between inter-class features and their corresponding representative anchors. Thus, the network captures global context and intricate details from multi-scale features and their connection using the representative anchors as the criterion.

Furthermore, the pixel-anchor loss operates in conjunction with the conventional pixel-wise cross-entropy loss $L_{\text{CE}}$ [35], providing a complementary approach to enhance segmentation performance. This synergy is purposeful: while pixel-wise cross-entropy loss aims to predict the correct label for each sample, pixel-anchor loss aims to learn good data representations by considering the relationships between different samples.

As a result, the primary objective of the entire framework is to optimize the following loss:

$$L = L_{\text{CE}} + \alpha L_{\text{PA}}, \tag{5}$$

where $\alpha$ represents the weight for our pixel-anchor loss.

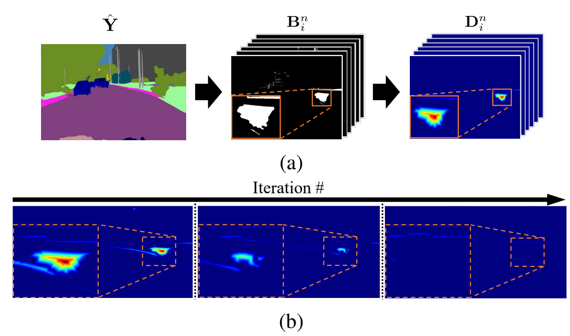


Figure 3. Visual description of boundary-aware negative sampling and how the under/over-segmentation problems are addressed during the training. (a) The prediction outcome $\hat{\mathbf{Y}}$ is decomposed into class-wise binary maps $\mathbf{B}_i^n$. Then, class-wise distance maps $\mathbf{D}_i^n$ are generated with the Distance Transform [23]. (b) The evolution of the distance map over iterations. The wrongly predicted regions shrink during training (best viewed in color).

### 3.3. Boundary-aware negative (BANE) sampling

While enhancing the loss function, we also propose an effective negative sampling approach that considers the boundaries of the prediction error to increase the quality of $\mathbf{v}_-$ in Eq. (4). To do that, we incorporate a simple but effective boundary extraction method from $\hat{Y}$ [38, 50]. The method mainly consists of three steps as illustrated in Fig. 3: 1) decomposing prediction output to class-wise binary error maps, 2) distance transform based on the class-wise error maps, and 3) selecting negative samples.

To extract negative samples, the class-wise binary error map $\mathbf{B}_i^n$ for each pixel $(u, v)$ is defined as:

$$\mathbf{B}_i^n(u, v) = \mathbb{1}[(\hat{y}_i \neq n) \wedge (g(\hat{y}_i) = n)], \tag{6}$$

where $\hat{y}_i$ denotes the predicted class for the $i$-th layer down-sampled from the predicted class in the final layer. $g(\cdot)$ denotes a labeling function in the same way as in Sec. 3.2. The $\mathbf{B}_i^n$ has a value of one for the wrongly predicted pixel, i.e. a negative sample, and zero otherwise, as shown in Fig. 3(a).

Next, $\mathbf{B}_i^n$ is converted to a class-wise distance map $\mathbf{D}_i^n$ by the distance transform [23]. The pixel value of $\mathbf{D}_i^n$ is the minimal distance between the pixel $(u, v)$ and the edge pixels $\mathbf{E}_i^n$, which is from the corresponding class-wise error map and is defined as follows:

$$\mathbf{D}_i^n(u, v) = \min_{(x,y)\in\mathbf{E}_i^n} \sqrt{(u-x)^2 + (v-y)^2}. \tag{7}$$

This implies that within the incorrectly predicted regions, where the pixel value of $\mathbf{B}_i^n$ is one (white regions in Fig. 3(a)), a lower value indicates a higher probability of the pixel being on the boundary.

Finally, among the regions whose values in $\mathbf{B}_i^n$ are one, we select embedding vectors corresponding to the lower $K$ percentage of the smallest distances in $\mathbf{D}_i^n$ as negative samples for each $n$-th representative anchor in Eq. (4). By incorporating these vectors into Eq. (3), Contextrast allows these vectors to be close to the anchor of the true class and far from the features of incorrectly predicted classes during training. Thus, these boundary-aware negative samples help the network learn the inter-spatial relationship between the segmentation classes better.

# 4. Experiments

## 4.1. Experimental setup

**Datasets.** We conduct our experiments using five public datasets: Cityscapes [9], ADE20K [55], PASCAL-C [29], COCO-Stuff [2], and CamVid [1] datasets. For a fair comparison, we follow the existing training and validation settings of the datasets (details are explained in the supplementary materials). Among them, because the Cityscapes dataset additionally provides the public benchmark by using test data, we differentiate the validation and test sets using the suffix `test`, i.e. Cityscapes-`test`.

**Training settings.** To demonstrate the efficacy of our proposed approach, we employ three networks: a) DeepLabV3 [6], b) HRNet [35], and c) OCRNet [49]. D-ResNet-101 backbone is utilized in DeepLabV3. HRNetV2-W48 backbone is employed in HRNet and OCRNet networks. We used same hyperparameters and initialized the network using pre-trained weights on ImageNet [10] while the remaining layers were randomly initialized. We utilized color jittering, horizontal flipping, and random scaling for data augmentation. Stochastic gradient descent (SGD) is applied as an optimizer for CNN backbones with a momentum of 0.9. In addition, polynomial annealing policy [6] is applied to schedule the learning rate, which is multiplied by $(1 - \frac{\text{Iteration \#}}{\text{Total iterations}})^{0.9}$. $\lambda_{4\rightarrow 1}$ is set to 1.0, 0.7, 0.4, and 0.1. $\alpha$ is set to 0.1. On the Cityscapes dataset, we have set a batch size of 8 for 40K iterations and cropped from 1024×2048 to 512×1024. The model is trained on the CamVid dataset with a batch size of 16 for 6K iterations. On ADE20K, the models are trained with a crop size of 512×512 and a batch size of 12 for 80K iterations. On COCO-Stuff and PASCAL-C, the models are trained with a crop size of 512×512 and a batch size of 16 for 60K iterations. Note that we do not use any extra training data.

**Testing settings.** We follow the general setup [6, 35, 49], averaging the segmentation results over multiple scales with flipping for CamVid, COCO-Stuff, ADE20K, and PASCAL-C datasets. The scaling factor is set from 0.75 to 2.0 with intervals of 0.25. We employed single-scale evaluation for Cityscapes to follow the experimental setup of multi/cross-scale contrastive learning [31].

**Evaluation metric.** In the experiments, we quantitatively analyze the performance with respect to a) semantic segmentation results and b) the distinctiveness of the features.

To evaluate the semantic segmentation performance, the mean of class-wise intersection over union (mIoU) [17, 41] is used as an evaluation metric. For the Cityscapes-`test`, an instance-level intersection-over-union metric (iIoU) [9] is also used to evaluate how the individual instances are well-segmented. This is because mIoU can be biased toward object instances that cover a large image area in the street scenes. The iIoU is defined as follows:

$$\text{iIoU} = \frac{\text{iTP}}{(\text{iTP} + \text{FP} + \text{iFN})}, \tag{8}$$

where iTP, FP, and iFN denote the numbers of true positive, false positive, and the number of false negative pixels, respectively. Note that iTP and iFN are calculated with weighted pixel contributions based on the ratio of each class's average instance size to the corresponding ground truth instance size.

Next, for the feature-level analysis, we adopt the following three metrics: intra-class alignment (A) to evaluate how well the intra-class features are closely clustered, inter-class uniformity (U) to evaluate how far the centroids of features originating from different classes are separated in the embedding space, and inter-class neighborhood uniformity ($\text{U}_l$) to measure the separation of the

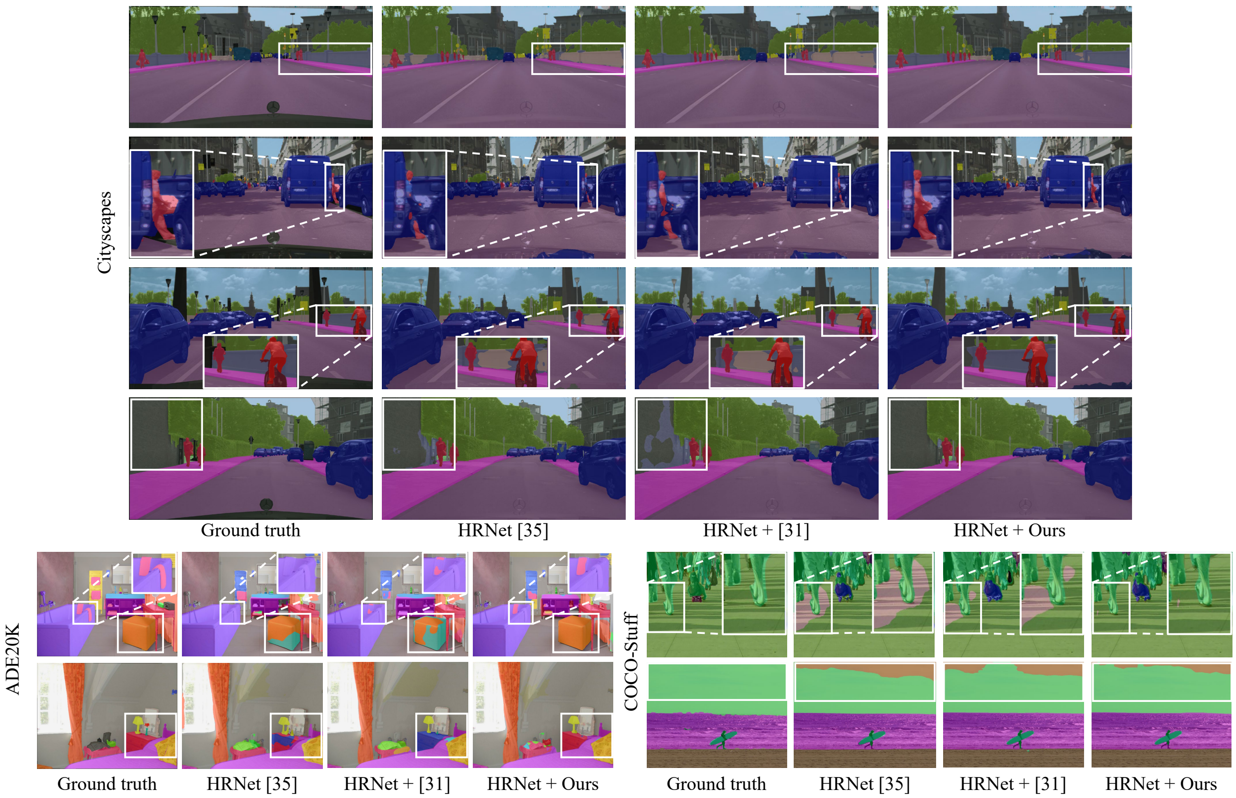


Figure 4. Qualitative results from HRNet [35], HRNet + [31], and HRNet + Ours on the Cityscapes, ADE20K, and COCO-Stuff datasets, respectively (best viewed on color).

| Method | Description | | Dataset [mIoU (%)] | | | | |
|---|---|---|---|---|---|---|---|
| | Loss | Sampling | Cityscapes | CamVid | COCO-Stuff | ADE20K | PASCAL-C |
| DeepLabV3 | $L_{\mathrm{CE}}$ | None | 77.12 | 78.80 | 37.92 | 42.85 | 52.01 |
| DeepLabV3 + [31] | $L_{\mathrm{CE}} + L_{\mathrm{cms}} + L_{\mathrm{ccs}}$ | Random | 78.94 (+1.82) | 79.67 (+0.87) | 37.39 (-0.53) | 43.86 (+1.01) | 51.52 (-0.49) |
| DeepLabV3 + Ours | $L_{\mathrm{CE}} + L_{\mathrm{PA}}$ (Ours) | Boundary-aware (Ours) | **79.35 (+2.23)** | **79.98 (+1.18)** | **38.12 (+0.20)** | **44.12 (+1.27)** | **52.62 (+0.61)** |
| HRNet | $L_{\mathrm{CE}}$ | None | 78.48 | 82.17 | 36.04 | 41.86 | 51.86 |
| HRNet + [41] | $L_{\mathrm{CE}} + L_{\mathrm{NCE}}$ | Semi-hard | 81.00 (+2.52) | N/A | N/A | N/A | N/A |
| HRNet + [17] | $L_{\mathrm{CE}} + L_{\mathrm{NCE}} + L_{\mathrm{Aux}}$ | Random | 81.90 (+3.42) | N/A | N/A | N/A | N/A |
| HRNet + [31] | $L_{\mathrm{CE}} + L_{\mathrm{cms}} + L_{\mathrm{ccs}}$ | Random | 81.50 (+3.02) | 83.14 (+0.97) | **36.35 (+0.31)** | 43.27 (+1.41) | 52.11 (+0.25) |
| HRNet + Ours | $L_{\mathrm{CE}} + L_{\mathrm{PA}}$ (Ours) | Boundary-aware (Ours) | **82.20 (+3.72)** | **84.33 (+2.16)** | 36.34 (+0.30) | **43.42 (+1.56)** | **52.17 (+0.31)** |
| OCRNet | $L_{\mathrm{CE}}$ | None | 79.95 | 82.69 | 39.00 | 41.51 | 54.35 |
| OCRNet + [31] | $L_{\mathrm{CE}} + L_{\mathrm{cms}} + L_{\mathrm{ccs}}$ | Random | 81.51 (+1.56) | 83.82 (+1.13) | 38.55 (-0.45) | 43.28 (+1.77) | 54.48 (+0.13) |
| OCRNet + Ours | $L_{\mathrm{CE}} + L_{\mathrm{PA}}$ (Ours) | Boundary-aware (Ours) | **81.94 (+1.99)** | **84.10 (+1.41)** | **39.08 (+0.08)** | **43.84 (+2.33)** | **54.64 (+0.29)** |

Table 2. Quantitative results on public datasets compared with the state-of-the-art contrastive learning-based semantic segmentation approaches. We employed DeepLabV3 [6], HRNet [35], and OCRNet [49] as segmentation models.

$l$-closest centroids of inter-class features, which indicates how clearly the decision boundaries are discriminated between the $l$-closest centroids. More details can be found in [26] (see Sec. 4.3).

### 4.2. Semantic segmentation performance

The first experiment compares the performance of our proposed approach with that of the existing contrastive learning-based methods, to support the claim that our approach enables networks to output more precise semantic segmentation, particularly resolving under- and over-segmentation issues. For our comparison, we used the following existing contrastive learning-based approaches: ContrastiveSeg [41] that incorporates $L_{\mathrm{CE}}$ with $L_{\mathrm{NCE}}$; semi-hard negative sampling that just randomly selects the negative samples corresponding to the wrongly predicted regions; region-aware contrastive learning [17] that additionally employs auxiliary loss $L_{\mathrm{Aux}}$; multi/cross-scale con-

| Method | Classes | | Categories | |
|---|---|---|---|---|
| | mIOU (%) | iIOU (%) | mIOU (%) | iIOU (%) |
| HRNet | 79.51 | 57.96 | 91.33 | 80.29 |
| HRNet + [31] | 80.12 (+0.61) | 59.04 (+1.08) | 91.42 (+0.09) | 81.64 (+1.35) |
| HRNet + Ours | **80.39 (+0.88)** | **61.06 (+3.10)** | **91.59 (+0.26)** | **82.14 (+1.85)** |
| OCRNet | 80.64 | 58.72 | 91.41 | 80.77 |
| OCRNet + Ours | **81.94 (+1.30)** | **62.11 (+3.39)** | **91.60 (+0.19)** | **81.83 (+1.06)** |
| DeeplabV3 | 77.01 | 55.56 | 89.64 | 77.42 |
| DeeplabV3 + Ours | **78.23 (+1.22)** | **56.83 (1.27)** | **89.86 (+0.22)** | **77.93 (+0.51)** |

Table 3. Quantitative segmentation results on Cityscapes-`test`.

| | B-mIoU (5px) | B-mIoU (7px) | B-mIoU (10px) |
|---|---|---|---|
| HRNet | 59.93 | 65.82 | 69.25 |
| HRNet + [31] | 60.44 (+0.51) | 66.29 (+0.47) | 69.65 (+0.4) |
| Ours | **61.76 (+1.83)** | **67.58 (+1.76)** | **70.93 (+1.68)** |

Table 4. B-mIoU performance comparison with Cityscapes dataset. B-mIoU represents the mIoU of the boundary region which is within pixels from the boundary.

trastive learning [31] that exploits multi-scale and cross-scale contrastive loss terms, i.e. $L_{\text{cms}}$ and $L_{\text{ccs}}$.

As shown in Fig. 4, Table 2, and Table 3, the state-of-the-art methods showed precise semantic segmentation results, mostly improving the mIoU compared with the baseline segmentation networks. Our proposed method exhibits noticeable improvements on the public datasets, mostly achieving the highest mIoU. In particular, our proposed method even resolved the under- and over-segmentation more clearly (see Fig. 4).

In addition to the substantial performance improvement, we specifically focus on the differences in the loss terms. As presented in Table 2, the performance after the application of the auxiliary loss $L_{\text{Aux}}$ [17] showed higher mIoU compared with ContrastiveSeg [41], which only incorporates $L_{\text{CE}}$ with $L_{\text{NCE}}$. The multi-scale and cross-scale contrastive losses [31], which are improved versions of $L_{\text{NCE}}$ in a different scale level, also showed a substantial increase in mIoU. However, our method showed a large performance increase.

In particular, it is noticeable that both multi/cross-scale contrastive learning [31] and ours considered multi-scale, yet our approach showed more stable performance increase. Occasionally, multi/cross-scale contrastive learning [31] showed the degraded performance owing to the conflict of the influences of two disentangled loss terms. That is, considering multi-scale and cross-scale separately in the learning process can sometimes result in the direction in which the moved embedding vector becomes undesirable, leading to a situation where the distinctiveness of vectors in the embedded space may not significantly increase. Furthermore, the proposed method significantly improved boundary mIoU (B-mIoU) by incorporating BANE sampling into the contrastive learning, as demonstrated in Table 4.

Therefore, we conclude that our Contextrast is more effective for accurate semantic segmentation than the existing methods.

### 4.3. Feature-level in-depth analyses

Furthermore, we conducted two experiments to demonstrate that our Contextrast enhances the distinctiveness of the vectors on the embedded feature space.

First, we assessed how the embedding vectors are aligned in the last layer just before reaching the segmentation head by using feature-level metrics, which were explained in Sec. 4.1. As presented in Table 5, we demonstrate that our proposed method aligns intra-class features and pushes away inter-class features, improving all the metrics. Thus, it implies that our method makes the model have distinctive decision boundaries because intra-class features are well-organized and inter-class features are well-discriminated in the latent space.

Second, we examined the cosine similarity between representative anchors and all the negative samples by considering their distances in the distance map, i.e. $\mathbf{D}_i^n$ in Fig. 3(a), for each layer. The lower distance implies that the negative samples are more likely to be from the edge regions of the wrongly predicted segments. A lower cosine similarity means that the vector that is supposed to be close to the anchor is far apart, implying that the negative samples are more challenging to discriminate in the feature space. As presented in Fig. 5, the features existing along the boundaries of the incorrect prediction regions, which have lower distance values, are harder to discriminate well in all encoder layers. Thus, it corroborates that our BANE sampling successfully prioritizes harder-negative samples, triggering more desirable gradient contributions for our contextual contrastive learning.

As a result, these analyses support our key claim that our multi-scale-aware representative anchors align features well on the embedding vector space and our BANE sampling successfully chooses informative negative examples.

### 4.4. Ablation study

**The impact of individual component.** To further examine the effectiveness of each module more closely, we conducted an ablation study, as shown in Table 6. Applying our contextual contrastive learning led to enhancements of mIoU. In particular, compared with ContrastiveSeg [41], the combination of our methods exhibits a larger gap in performance increase, which supports our key claim that the negative samples chosen by our BANE sampling on the

| | Method | A ↓ | U ↑ | $U_3$ ↑ | $U_5$ ↑ |
|---|---|---|---|---|---|
| Cityscapes | HRNet | 0.70 | 0.98 | 0.47 | 0.55 |
| | HRNet + [31] | 0.53 (-0.17) | 1.00 (+0.02) | **0.50 (+0.03)** | 0.58 (+0.03) |
| | HRNet + Ours | **0.42 (-0.28)** | **1.01 (+0.03)** | **0.50 (+0.03)** | **0.59 (+0.04)** |
| CamVid | OCRNet | 0.60 | 1.09 | 0.64 | 0.72 |
| | OCRNet + [31] | 0.55 (-0.05) | **1.10 (+0.01)** | 0.67 (+0.03) | 0.74 (+0.02) |
| | OCRNet + Ours | **0.52 (-0.08)** | **1.10 (+0.01)** | **0.68 (+0.04)** | **0.75 (+0.03)** |

Table 5. Ablation analysis of Alignment (A), Uniformity (U), and the $l$-closest Neighborhood Uniformity ($U_l$) on the Cityscapes and CamVid datasets with HRNet [35] and OCRNet [6] as segmentation models.

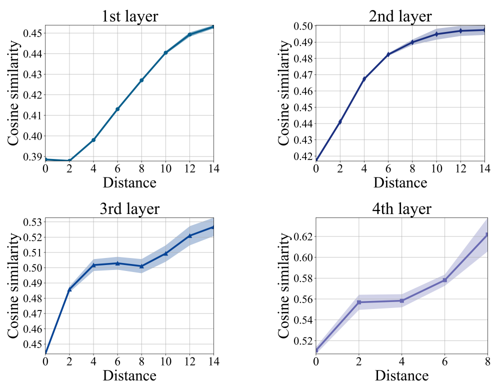


Figure 5. Average cosine similarity between error pixels and representative anchors in each layer was computed based on the distance from incorrect prediction boundaries. The results demonstrate that samples located along the incorrect prediction boundaries are harder-negative samples compared with features in the inner region.

contrastive learning is helpful by making vectors on the embedded space more distinct. These results highlight the effectiveness of our approach, demonstrating a substantial increase in mIoU, making it a promising solution for semantic segmentation tasks across diverse datasets.

**Analysis of anchor fusion weight and ratio of boundary-aware negative sampling.** Table 7 demonstrates the proposed method's resilience to weight selection variations. The sum of $w_h$ and $w_l$ is 1. Except when $w_h$ is set to 0 or 1, the proposed method consistently improves the performance. Thus, Table 7 shows that our proposed method is stable with regard to hyperparameter tuning. In addition, we examine the impact of BANE sampling using different sampling ratios $K$, which is presented in Sec. 3.3. As shown in Table 8, our analysis reveals that the sampling method mostly enhances performance with the ratio of 50%. However, the performance declines when excessive negative sampling is applied because it leads to local minima [3, 33, 45]. More ablation studies on hyper-parameters are presented in the supplementary material.

| | CCL (Ours) | Sampling | | mIOU (%) |
|---|---|---|---|---|
| | | Semi-hard [41] | BANE (Ours) | |
| Cityscapes | | | | 78.48 |
| | ✓ | | | 81.88 (+3.40) |
| | ✓ | ✓ | | 82.01 (+3.53) |
| | ✓ | | ✓ | **82.20 (+3.72)** |
| CamVid | | | | 82.17 |
| | ✓ | | | 83.14 (+0.97) |
| | ✓ | ✓ | | 83.38 (+1.21) |
| | ✓ | | ✓ | **84.33 (+2.16)** |

Table 6. Ablation study: performance according to the presence or absence of each component of our proposed method on the Cityscapes and CamVid datasets with HRNet [35] (CCL: contextual contrastive learning).

| $w_h$ | 0.0 | 0.3 | 0.5 | 0.7 | 1.0 |
|---|---|---|---|---|---|
| mIoU (%) | 81.15 | 81.27 | 81.80 | **81.88** | 81.31 |

Table 7. Comparison of different weights for the representative anchor fusion on Cityscapes-`val` with HRNet [35]. The sum of $w_h$ and $w_l$ is equal to 1.

| Ratio $K$ (%) | 0 | 10 | 30 | 50 | 70 | 100 |
|---|---|---|---|---|---|---|
| mIoU (%) | 81.88 | 81.89 | 81.58 | **82.20** | 81.58 | 81.10 |

Table 8. Comparison of different sampling ratios for boundary-aware negative sampling on Cityscapes-`val` with HRNet [35].

## 4.5. Conclusions

In this paper, we have proposed a novel boundary-aware contrastive learning for semantic segmentation, called *Contextrast*. By leveraging multi-scale contextual contrastive learning, we enable the network capture local/global context information and consistently understand their relationship. In particular, we demonstrate our BANE sampling substantially increases mIoU by providing more harder negative samples on the contrastive learning stage. Consequently, our approach achieved promising results compared with other contrastive learning approaches on public datasets.

**Acknowledgements.** This research was supported in part by the KAIST Convergence Research Institute Operation Program and in part by Korea Evaluation Institute Of Industrial Technology (KEIT) grant funded by the Korea government(MOTIE) (No.20023455, Development of Cooperate Mapping, Environment Recognition and Autonomous Driving Technology for Multi Mobile Robots Operating in Large-scale Indoor Workspace). We thank KI Cloud of Division of National Supercomputing Center, Korea Institute of Science and Technology Information(KISTI).

## References


[1] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. 2, 3, 6

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 2, 3, 6

[3] Tiffany Tianhui Cai, Jonathan Frankle, David J Schwab, and Ari S Morcos. Are all negatives created equal in contrastive instance discrimination? *arXiv preprint arXiv:2010.06682*, 2020. 9

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. *arXiv preprint arXiv:1412.7062*, 2014. 2

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3, 6, 7, 9

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 801–818, 2018.

[8] Sungha Choi, Joanne T Kim, and Jaegul Choo. Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitionn*, pages 9373–9383, 2020. 2

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 2, 3, 6

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2, 6

[11] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Semantic correlation promoted shape-variant context for segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8885–8894, 2019. 2

[12] Nguyen Thanh Duc, Nguyen Thi Oanh, Nguyen Thi Thuy, Tran Minh Triet, and Viet Sang Dinh. ColonFormer: An efficient transformer-based method for colon polyp segmentation. *IEEE Access*, 10:80575–80586, 2022. 2

[13] Razvan-Gabriel Dumitru, Darius Peteleaza, and Catalin Craciun. Using DUCK-Net for polyp image segmentation. *Scientific Reports*, 13(1):9803, 2023. 2

[14] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019. 2

[15] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 297–304. JMLR, 2010. 4

[16] Yuanduo Hong, Huihui Pan, Weichao Sun, and Yisong Jia. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *arXiv preprint arXiv:2101.06085*, 2021. 2

[17] Hanzhe Hu, Jinshi Cui, and Liwei Wang. Region-aware contrastive learning for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16291–16301, 2021. 3, 4, 6, 7, 8

[18] Juana Valeria Hurtado and Abhinav Valada. Semantic scene segmentation for robotics. In *Deep Learning for Robot Perception and Cognition*, pages 279–311. Elsevier, 2022. 2

[19] Chuong Huynh, Anh Tuan Tran, Khoa Luu, and Minh Hoai. Progressive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16755–16764, 2021. 2

[20] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809, 2020. 3

[21] Tsung-Wei Ke, Jyh-Jing Hwang, Ziwei Liu, and Stella X Yu. Adaptive affinity fields for semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 587–602, 2018. 2

[22] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 4

[23] Ron Kimmel, Nahum Kiryati, and Alfred M Bruckstein. Sub-pixel distance maps and weighted distance transforms. *Journal of Mathematical Imaging and Vision*, 6:223–233, 1996. 5, 6

[24] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018. 2

[25] Liulei Li, Tianfei Zhou, Wenguan Wang, Jianwu Li, and Yi Yang. Deep hierarchical semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1246–1257, 2022. 2, 3

[26] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6918–6928, 2022. 7

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 3

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 3

[29] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014. 2, 3, 6

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[31] Theodoros Pissas, Claudio S Ravasio, Lyndon Da Cruz, and Christos Bergeles. Multi-scale and cross-scale contrastive learning for semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 413–429. Springer, 2022. 3, 4, 6, 7, 8, 9

[32] Edward Sanderson and Bogdan J Matuszewski. FCN-transformer feature fusion for polyp segmentation. In *Proceedings of the Annual Conference on Medical Image Understanding and Analysis*, pages 892–907. Springer, 2022. 2

[33] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 9

[34] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021. 2, 3

[35] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019. 2, 3, 5, 6, 7, 9

[36] Haoru Tan, Sitong Wu, and Jimin Pi. Semantic diffusion network for semantic segmentation. *Advances in Neural Information Processing Systems*, 35:8702–8716, 2022. 3

[37] Maria Tzelepi and Anastasios Tefas. Semantic scene segmentation for robotics applications. In *Proceedings of the International Conference on Information, Intelligence, Systems & Applications*, pages 1–4, 2021. 2

[38] Chi Wang, Yunke Zhang, Miaomiao Cui, Peiran Ren, Yin Yang, Xuansong Xie, Xian-Sheng Hua, Hujun Bao, and Weiwei Xu. Active boundary loss for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2397–2405, 2022. 3, 6

[39] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2020. 2

[40] J Wang, Q Huang, F Tang, J Meng, J Su, and S Song. Stepwise feature fusion: Local guides global. *arXiv preprint arXiv:2203.03635*, 2023. 2

[41] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313, 2021. 3, 4, 6, 7, 8, 9

[42] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng 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*, pages 14408–14419, 2023. 2, 3

[43] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNext v2: Co-designing and scaling ConvNets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023. 3

[44] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision*, pages 418–434, 2018. 3

[45] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Delving into inter-image invariance for unsupervised visual representations. *International Journal of Computer Vision*, 130(12):2994–3013, 2022. 9

[46] Jiacong Xu, Zixiang Xiong, and Shankar P Bhattacharyya. PIDNet: A real-time semantic segmentation network inspired by PID controllers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19529–19539, 2023. 3

[47] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1857–1866, 2018.

[48] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12416–12425, 2020.

[49] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 173–190. Springer, 2020. 3, 6, 7

[50] Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. SegFix: Model-agnostic boundary refinement for segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 489–506. Springer, 2020. 3, 6

[51] Salih Can Yurtkulu, Yusuf Hüseyin Şahin, and Gozde Unal. Semantic segmentation with extended deeplabv3 architecture. In *Proceedings of the Signal Processing and Commu-

*nications Applications Conference*, pages 1–4. IEEE, 2019. 3

[52] Li Zhang, Xiangtai Li, Anurag Arnab, Kuiyuan Yang, Yunhai Tong, and Philip HS Torr. Dual graph convolutional network for semantic segmentation. *arXiv preprint arXiv:1909.06121*, 2019.

[53] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.

[54] Zhisheng Zhong, Jiequan Cui, Yibo Yang, Xiaoyang Wu, Xiaojuan Qi, Xiangyu Zhang, and Jiaya Jia. Understanding imbalanced semantic segmentation through neural collapse. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19550–19560, 2023. 3

[55] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017. 2, 3, 6

[56] Jingchun Zhou, Mingliang Hao, Dehuan Zhang, Peiyu Zou, and Weishi Zhang. Fusion PSPNet image segmentation based method for multi-focus image fusion. *IEEE Photonics Journal*, 11(6):1–12, 2019. 3