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

1. Detailed explanation of experimental setups and datasets

The details of the experimental setups are described in Table 1. The details of the datasets are described as follows.

- **Cityscapes** [3] includes images from 50 cities across Germany, captured in both rural and urban environments. It contains 5,000 images with 2,975 train, 500 validation, and 1,525 test images with 19 semantic classes.
- **ADE20K** [10] has images depicting various scenes, including indoor and outdoor environments. Unlike datasets focusing on specific domains such as autonomous driving, ADE20K contains diverse scenes such as bedrooms, offices, parks, and more. It comprises 20,210 train and 2,000 validation images with 150 semantic classes.
- **PASCAL-C** [5] contains 4,998 train and 5,105 test images with 59 semantic classes. It also includes both indoor and outdoor environments.
- **COCO-Stuff** [2] has 9,000 train and 1,000 test images. It provides 80 object classes and 91 stuff classes.
- **CamVid** [1] contains 367 train, 101 validation, and 233 test images with 11 semantic classes.

2. Detailed explanation of metrics for feature-level analyses

We adopted evaluation metrics from [4], denoted as alignment, uniformity, and neighborhood uniformity. The intra-class alignment, denoted as A, indicates how well the intra-class features are converged and is defined as follows:

$$\mathrm{A} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|V_i|^2}\sum_{v_j,v_k\in V_i}||v_j - v_k||_2, \tag{1}$$

where N, i, and V_i represent the number of semantic classes, the i-th semantic class, and the feature set of the i-th semantic class, respectively. By doing so, Eq. (1) represents how closely intra-class features are clustered just before reaching the segmentation head. Effective clustering of intra-class features signifies improved discrimination capabilities.

The inter-class uniformity, denoted as U, represents how well the centers of inter-class features are separated in the feature space and is defined as follows:

$$\mathrm{U} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}||\mu_i - \mu_j||_2, \tag{2}$$

where N and μ_i represent the number of semantic classes and center of i-th semantic class, respectively.

Finally, the neighborhood uniformity, denoted as U_l, measures the separation of l closest center of inter-class features. Neighborhood uniformity is defined as follows:

$$\mathrm{U}_l = \frac{1}{Nl}\sum_{i=1}^{N}\min_{j_1,\cdots,j_l}\left(\sum_{j=1,j\neq i}^{l}||\mu_i - \mu_j||_2\right). \tag{3}$$

Both uniformity and neighborhood uniformity imply how well the model defines the decision boundaries between inter-class features. As a result, alignment A, uniformity U, and neighborhood uniformity U_l represent the model's ability to distinguish intra-class and inter-class features.

3. Gradients of the loss function

In this section, we prove that harder negative samples in contrastive learning bring more gradient contribution during the training procedure. The proposed loss function is as follows:

$$L_i = \frac{1}{N}\sum_{\hat{\mathbf{a}}_i^n\in\hat{\mathbf{A}}_i}\frac{1}{|\mathbf{V}_+|}\sum_{\mathbf{v}_+\in\mathbf{V}_+}L_{\mathbf{a}}, \tag{4}$$

$$L_{\mathbf{a}} = -\log\frac{\exp(\hat{\mathbf{a}}_i^n\cdot\mathbf{v}_+/\tau)}{\exp(\hat{\mathbf{a}}_i^n\cdot\mathbf{v}_+/\tau) + \sum_{\mathbf{v}_-\in\mathbf{V}_-}\exp(\hat{\mathbf{a}}_i^n\cdot\mathbf{v}_-/\tau)}. \tag{5}$$

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

	Method		Training Settings						
	Model	Backbone	Crop size	Learning rate (Lr)	Weight decay	Optimizer	Lr scheduler	Batch size	Training steps
Cityscapes	DeepLabV3	D-ResNet-101	512 × 1024	10^{-2}	5×10^{-4}	SGD	Poly	8	40K
	HRNet	HRNetV2-W48	512 × 1024	10^{-2}	5×10^{-4}	SGD	Poly	8	40K
	OCRNet	HRNetV2-W48	512 × 1024	10^{-2}	5×10^{-4}	SGD	Poly	8	40K
	UPerNet	Swin-T	512 × 1024	6×10^{-5}	10^{-2}	ADAMW	Linear	6	40K
CamVid	DeepLabV3	D-ResNet-101	360 × 480	2×10^{-2}	5×10^{-4}	SGD	Poly	16	6K
	HRNet	HRNetV2-W48	360 × 480	2×10^{-2}	5×10^{-4}	SGD	Poly	16	6K
	OCRNet	HRNetV2-W48	360 × 480	2×10^{-2}	5×10^{-4}	SGD	Poly	16	6K
	UPerNet	Swin-T	360 × 480	6×10^{-5}	10^{-2}	ADAMW	Linear	16	6K
ADE20K	DeepLabV3	D-ResNet-101	512 × 512	10^{-2}	5×10^{-4}	SGD	Poly	12	80K
	HRNet	HRNetV2-W48	512 × 512	10^{-2}	5×10^{-4}	SGD	Poly	12	80K
	OCRNet	HRNetV2-W48	512 × 512	10^{-2}	5×10^{-4}	SGD	Poly	12	80K
COCO-Stuff	DeepLabV3	D-ResNet-101	512 × 512	10^{-3}	5×10^{-4}	SGD	Poly	16	60K
	HRNet	HRNetV2-W48	512 × 512	10^{-3}	5×10^{-4}	SGD	Poly	16	60K
	OCRNet	HRNetV2-W48	512 × 512	10^{-3}	5×10^{-4}	SGD	Poly	16	60K
PASCAL-C	DeepLabV3	D-ResNet-101	512 × 512	10^{-3}	10^{-4}	SGD	Poly	16	60K
	HRNet	HRNetV2-W48	512 × 512	10^{-3}	10^{-4}	SGD	Poly	16	60K
	OCRNet	HRNetV2-W48	512 × 512	10^{-3}	10^{-4}	SGD	Poly	16	60K

Table 1. Details of the experimental setup for each dataset and semantic segmentation model.

Method	Description		Dataset [mIOU(%)]	
	Loss	Sampling	Cityscapes	CamVid
UPerNet	L_{CE}	None	78.99	80.85
UPerNet + [6]	$L_{\text{CE}} + L_{\text{cms}} + L_{\text{ccs}}$	Random	78.90 (-0.09)	80.69 (-0.16)
UPerNet + Ours	$L_{\text{CE}} + L_{\text{PA}}$ (Ours)	Boundary-aware (Ours)	**79.98** (+0.99)	**80.88** (+0.03)

Table 2. Quantitative results on CamVid and Cityscapes compared with baseline model and with multi/cross-scale contrastive learning.

Then, the derivative of L_i with respect to the anchor $\hat{\mathbf{a}}_i^n$ is obtained as follows:

$$\frac{\partial L_i}{\partial \hat{\mathbf{a}}_i^n} = \frac{-1}{\tau N |\mathbf{V}_+|} \sum_{\hat{\mathbf{a}}_i^n \in \hat{\mathbf{A}}_i} \sum_{\mathbf{v}_+ \in \mathbf{V}_+} \Big((1-p_+) \cdot \mathbf{v}_+ - \sum_{\mathbf{v}_- \in \mathbf{V}_-} p_- \cdot \mathbf{v}_- \Big), \tag{6}$$

where $p_{+/-} = \frac{\exp(\hat{\mathbf{a}}_i^n \cdot \mathbf{v}_{+/-}/\tau)}{\sum_{\mathbf{v} \in \mathbf{V}} \exp(\hat{\mathbf{a}}_i^n \cdot \mathbf{v}/\tau)}$ denotes a matching probability between anchor and samples. Thus, once we sample harder negative samples by our BANE sampling, the dot product between anchor $\mathbf{a}_i^n$ and negative sample $\mathbf{v}_-$ is close to 1. Thus, the matching probability of negative p_- is increased. As a result, the gradient of the loss function is increased when the negative samples are harder.

4. Additional quantitative results

In this section, we demonstrate more quantitative results with transformer-based semantic segmentation model on Cityscapes and CamVid datasets. As shown in Tables 2 and 3, our Contextrast also improves segmentation performance compared with the baseline model [8] and with multi/cross-scale contrastive learning [6]. Contextrast aligns intra-class features and separates inter-class features better than baseline model and multi/cross-scale contrastive learning, as shown in Table 4.

Method	Classes		Categories	
	mIOU (%)	iIOU (%)	mIOU (%)	iIOU (%)
UPerNet	78.71	56.82	90.54	79.24
UPerNet + [6]	79.00 (+0.29)	57.57 (+0.75)	**90.79** **(+0.25)**	79.45 (+0.21)
UPerNet + Ours	**79.51** **(+0.80)**	**58.12** **(+1.30)**	90.66 (+0.12)	**79.48** **(+0.24)**

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

	Method	A ↓	U ↑	U_3 ↑	U_5 ↑
Cityscapes	UPerNet	0.83	1.39	0.73	0.82
	UPerNet + [6]	0.70 (-0.13)	1.48 (+0.09)	0.77 (+0.04)	0.87 (+0.05)
	UPerNet + Ours	**0.65** **(-0.18)**	**1.49** **(+0.10)**	**0.79** **(+0.06)**	**0.89** **(+0.07)**
CamVid	UPerNet	0.78	2.24	1.17	1.37
	UPerNet + [6]	0.69 (-0.09)	2.30 (+0.06)	1.23 (+0.06)	1.43 (+0.06)
	UPerNet + Ours	**0.65** **(-0.13)**	**2.34** **(+0.10)**	**1.27** **(+0.10)**	**1.47** **(+0.10)**

Table 4. Feature-level quantitative analysis of intra-class alignment (A), inter-class uniformity (U), and the l-closest neighborhood uniformity (U_l) on Cityscapes and CamVid datasets with UPerNet.

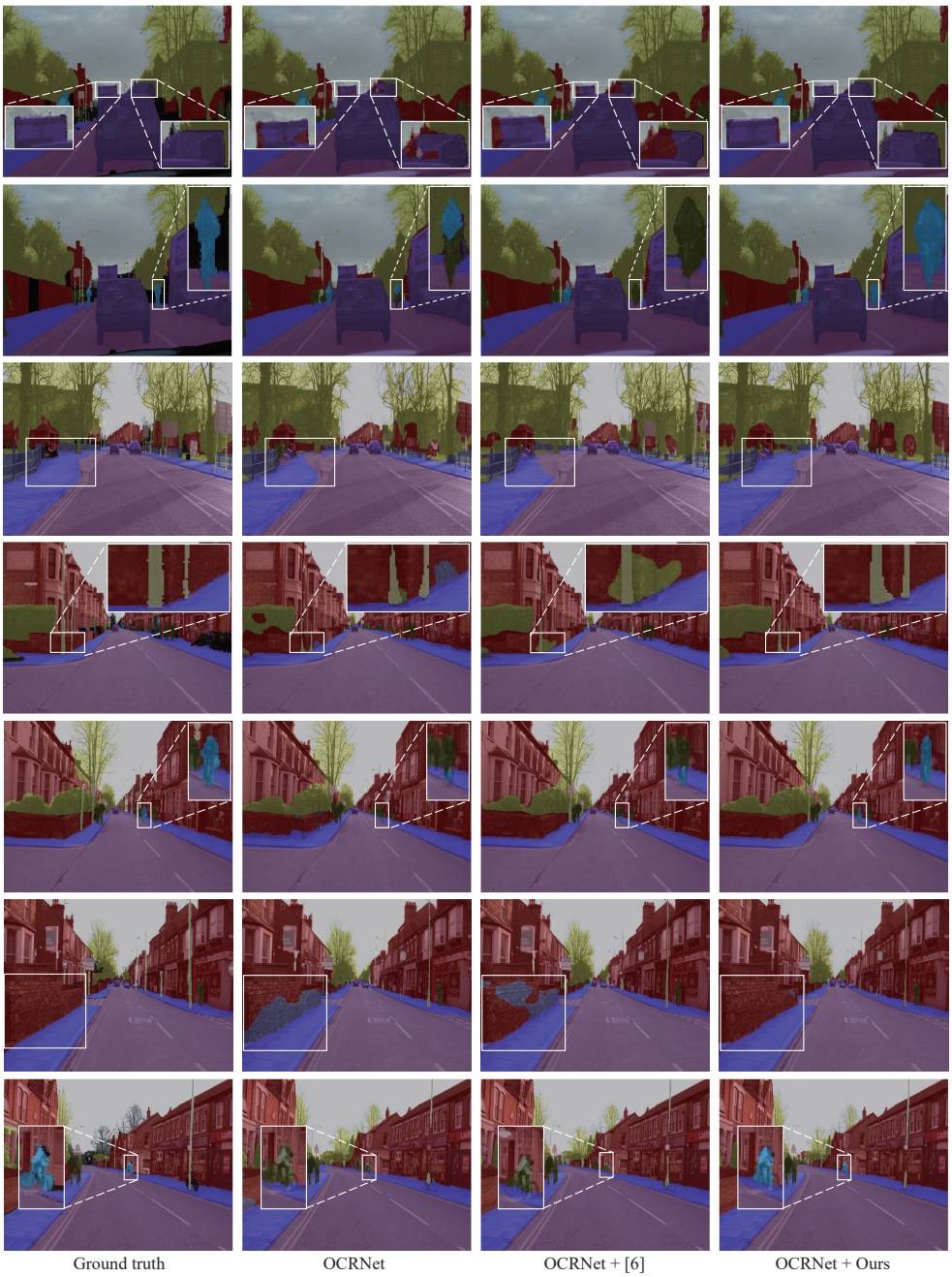


Figure 1. Qualitative results from OCRNet, OCRNet + [6], and OCRNet + Ours on CamVid (best viewed on color).

5. Additional qualitative results

This section demonstrates more qualitative comparisons between the baseline model, multi/cross-scale contrastive learning [6], and Contextrast. Fig. 1 demonstrates qualitative results with OCRNet [9] on CamVid. In addition, Fig. 2 shows more qualitative comparisons with OCRNet on Cityscapes, ADE20K, and COCO-Stuff. Finally, qualitative results for transformer-based semantic segmentation are shown in Fig. 3. We observed that Contextrast shows better semantic segmentation results with OCRNet and even better results with the transformer.

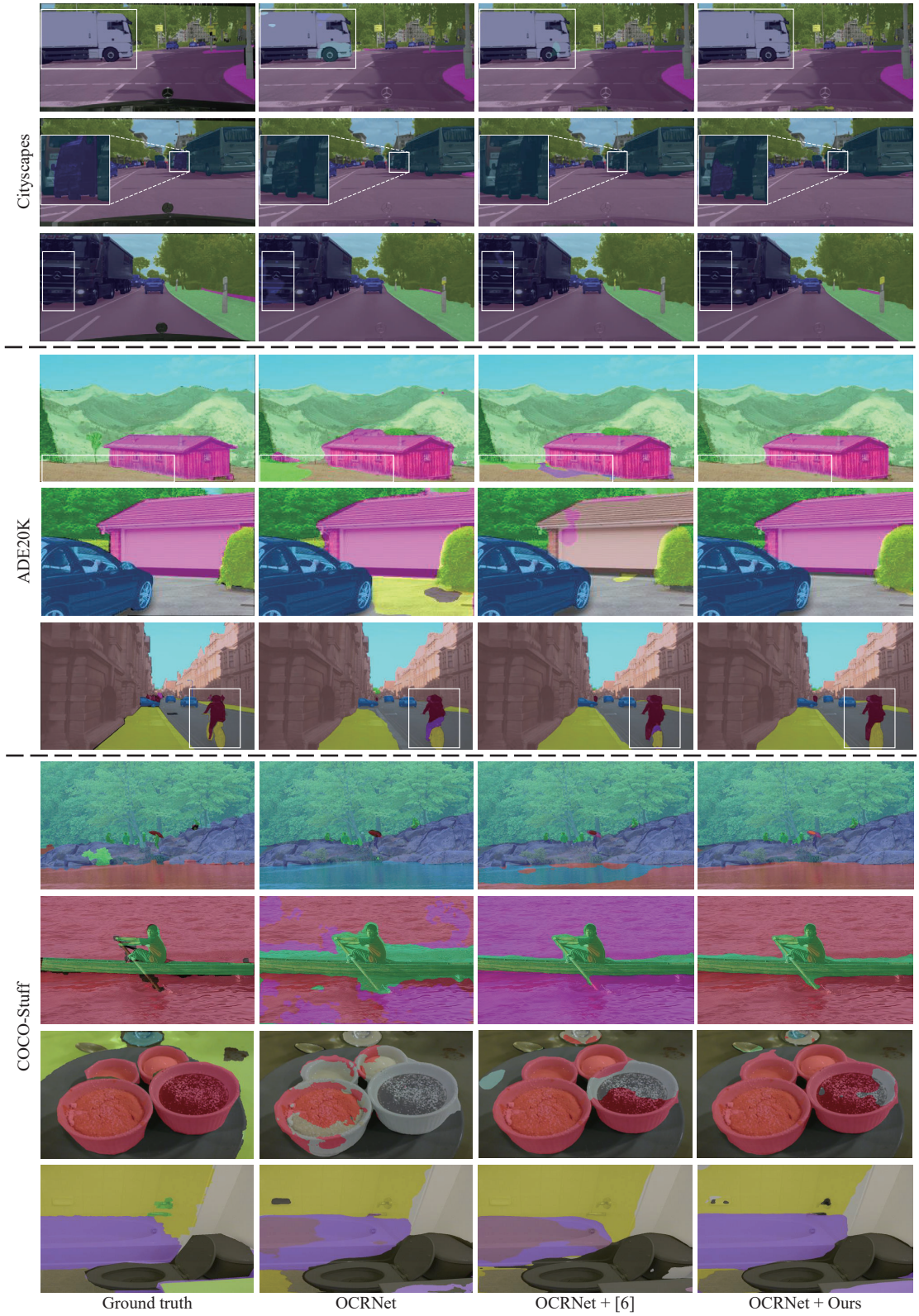


Figure 2. Qualitative results from OCRNet, OCRNet + [6], and OCRNet + Ours on Cityscapes, ADE20K, and COCO-Stuff datasets (best viewed on color).

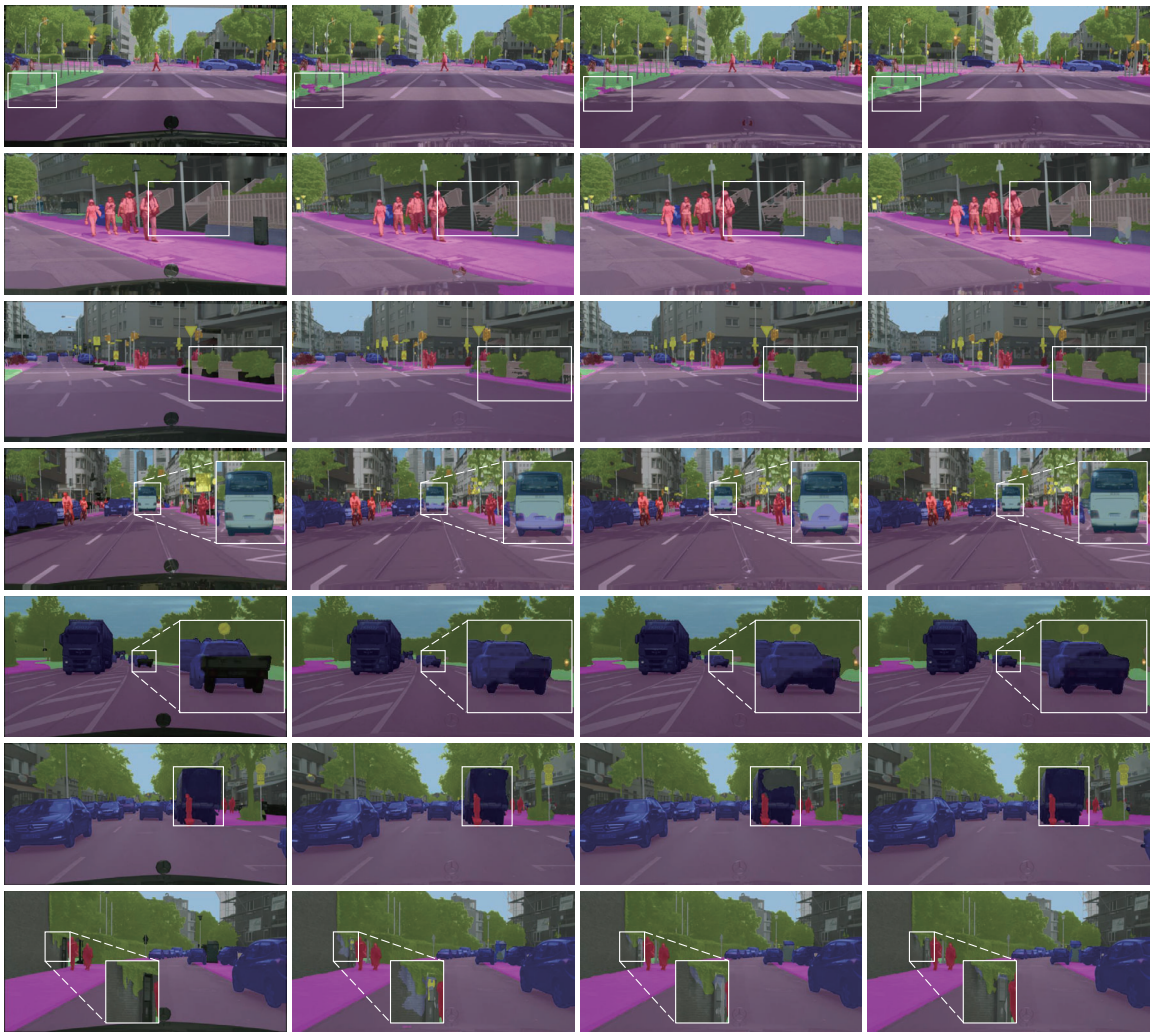


Figure 3. Qualitative results from UPerNet, UPerNet + [6], and UPerNet + Ours on Cityscapes (best viewed on color).

6. Qualitative comparisons for feature-level analyses

This section demonstrates additional feature-level analyses with qualitative results. We visualized the gradient-weighted class activation mapping (Grad-CAM), feature maps of the last layer, and t-distributed stochastic neighbor embedding (t-SNE). Grad-CAM highlights important regions in the image for prediction, as shown in Fig. 4, which demonstrates Grad-CAM for bicycle, bus, car, motorcycle, person, pole, and rider classes on Cityscapes. As illustrated in Fig. 4, Contextrast focuses more on the correct regions and does not focus on the unlabeled regions, i.e. poles.

The feature map of the last layer, which is just before the segmentation head, is illustrated as Fig. 5. The feature of the baseline model has less context information and too many fine details that are likely to be noisy, which causes over-segmentation problems. The feature of the multi/cross-scale contrastive learning method has too few fine details, which causes under-segmentation problems. In contrast, our proposed method balances both fine details and global context in feature maps, so Contextrast achieved better semantic segmentation performances.

Figs. 6 and 7 demonstrate features learned with baseline model and Contextrast by t-SNE. Each class label is colored differently. Contextrast better aligns intra-class features and separates inter-class features in each layer compared with the baseline semantic segmentation model in the feature space, as shown in Figs. 6 and 7.

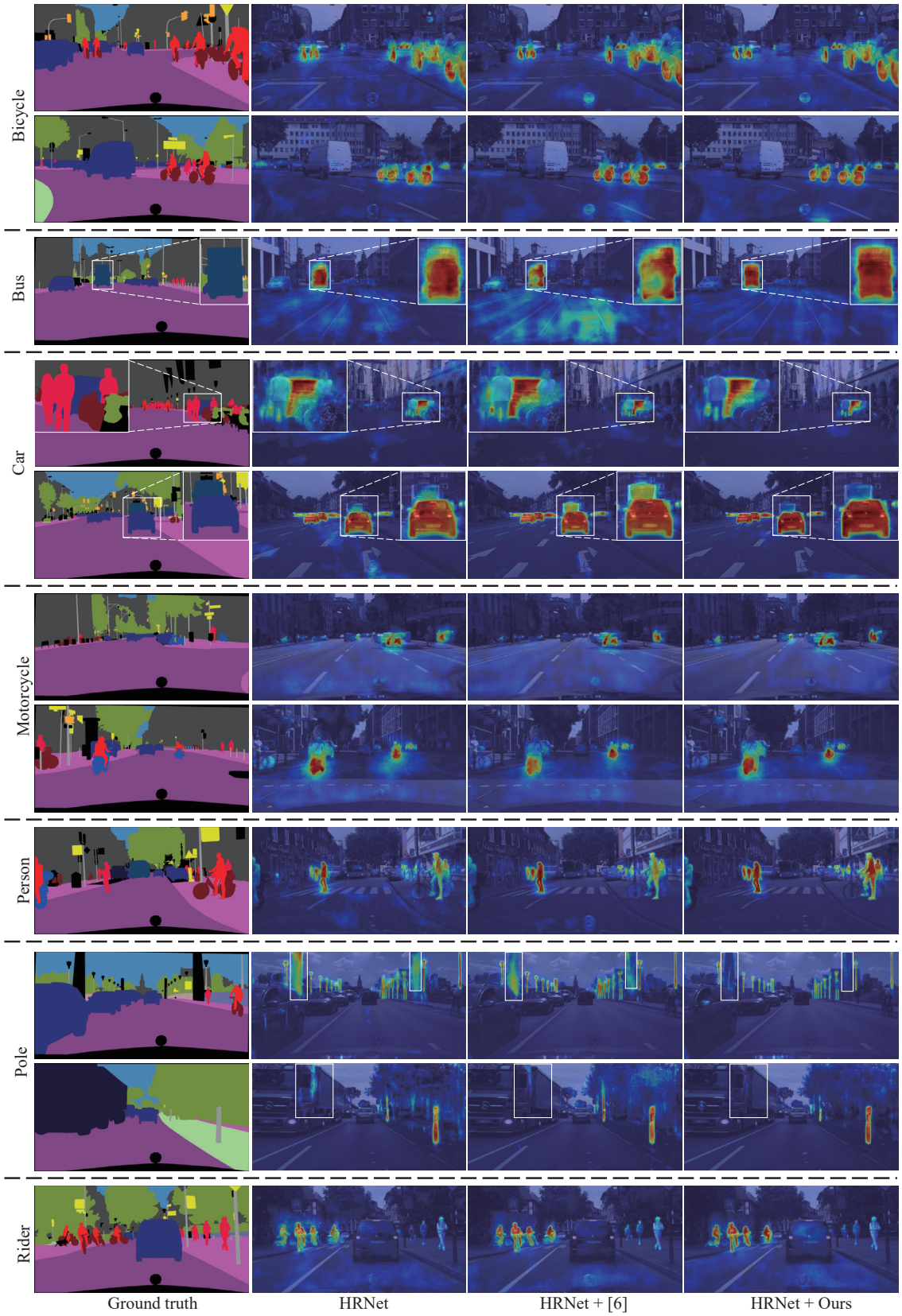


Figure 4. Grad-CAM results from HRNet, HRNet + [6], and HRNet + Ours on the Cityscapes dataset (best viewed on color).

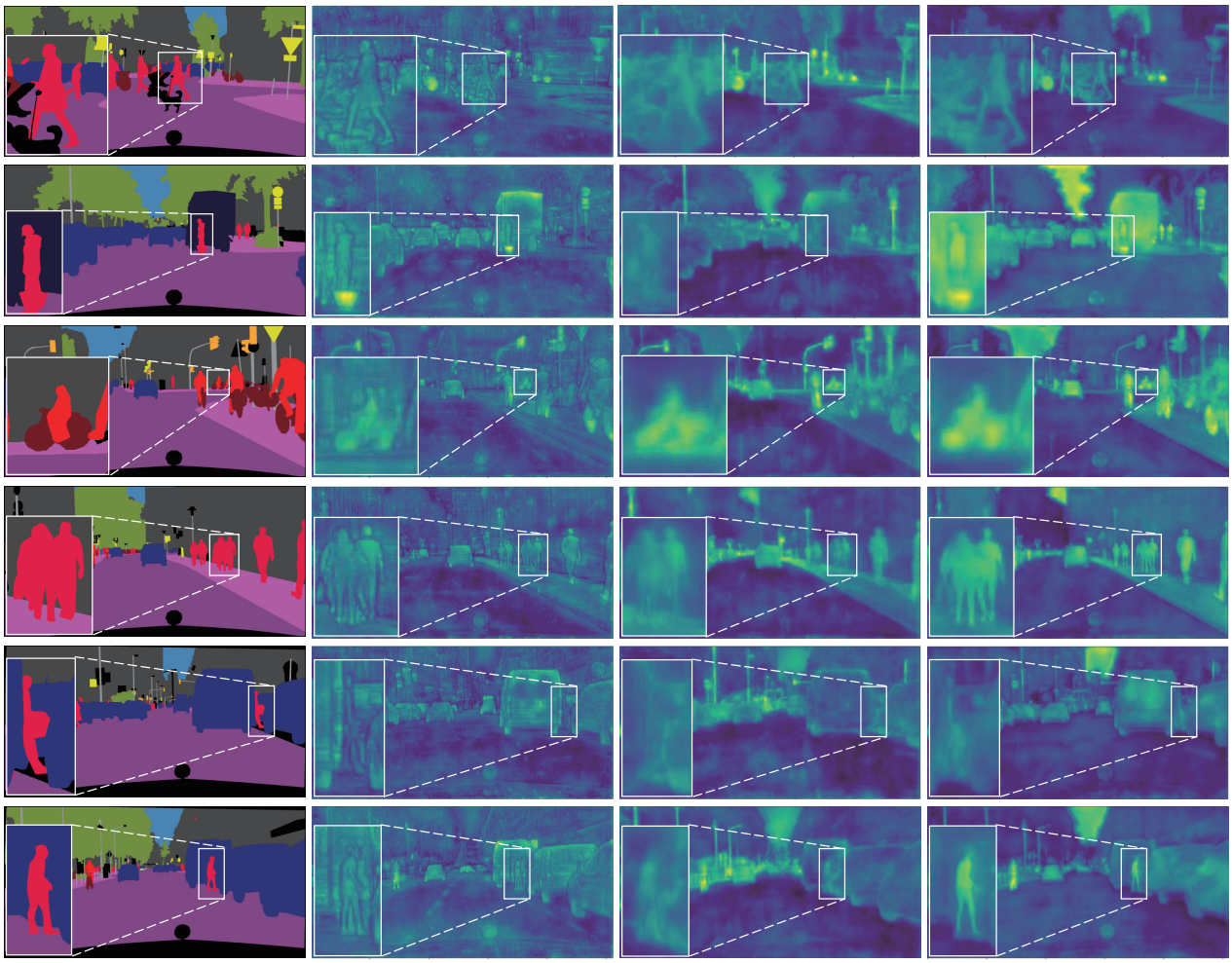
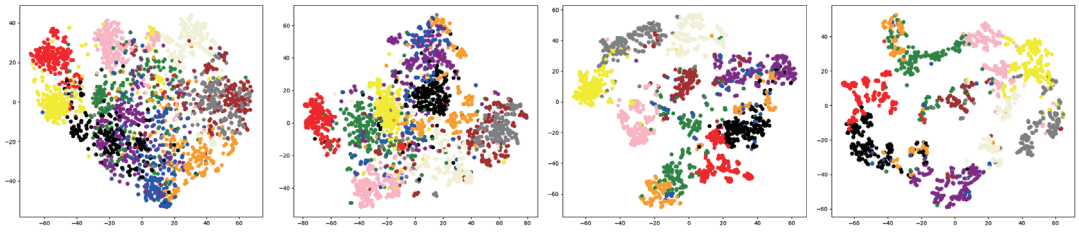
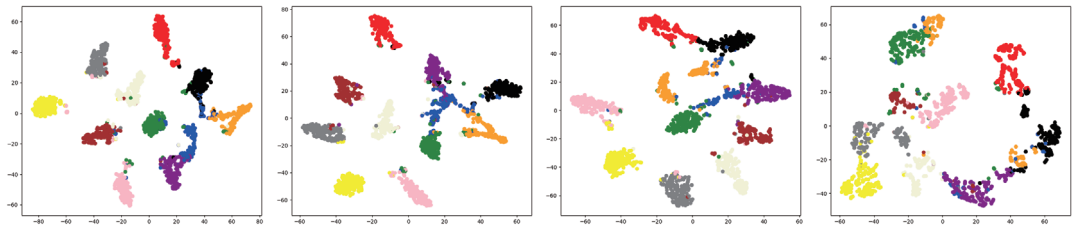


Figure 5. Feature map of the last layer from HRNet, HRNet + [6], and HRNet + Ours on the Cityscapes dataset (best viewed on color).

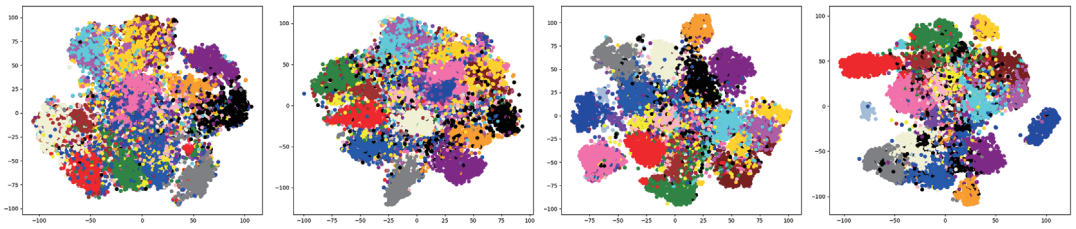


(a)

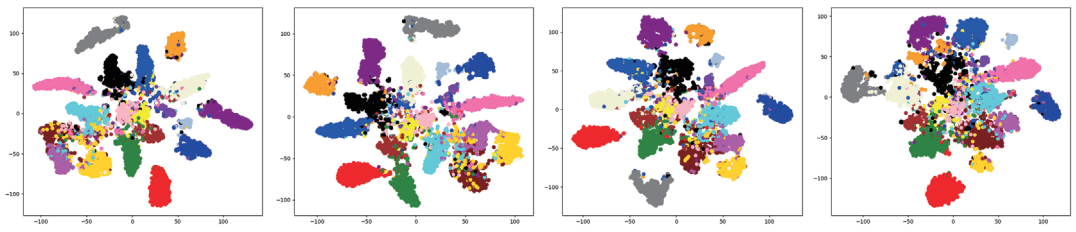


(b)

Figure 6. Visualization of features learned with HRNet [7] and Contextrast on CamVid. Each class label is colored differently. (a) t-SNE results of the baseline model. (b) t-SNE results of Contextrast. Note that the distributions of features corresponding to each class become more distinguishable (best viewed in color).



(a)



(b)

Figure 7. Visualization of features learned with HRNet [7] and Contextrast on Cityscapes. Each class label is colored differently. (a) t-SNE results of the baseline model. (b) t-SNE results of Contextrast. Note that the distributions of features corresponding to each class become more distinguishable (best viewed in color).

Representative anchor $\hat{\mathbf{A}}$	Dataset [mIOU (%)]	
	Cityscapes	CamVid
Lowest layer	81.29	83.33
Highest layer (Ours)	**82.20** (+0.91)	**84.33** (+1.00)

Table 5. Ablation study: performance variation according to the selection of representative anchor $\hat{\mathbf{A}}$ which is shared in each layer. Contextrast shares the highest representative anchor. To test the performance variation depending on the selection of the representative anchor $\hat{\mathbf{A}}$, we have experimented with the case sharing the lowest representative anchor in each layer.

7. Additional ablation study

In this section, we demonstrate two ablation studies. First, Table 5 presents the rationale why Contextrast shares the representative anchor of the highest-level features instead of the representative anchor of the lowest-level features. When the representative anchor is set as the lowest layer, higher-level features are aligned based on the characteristics of fine details and features in that layer. Thus, it loses global context information in higher-level features. On the other hand, the proposed method shares the global context information, so lower-level features are aligned based on the characteristics of the global context and features in that layer. Thus, it maintains global context information in all layers. Table 5 demonstrates that our proposed method comprehends global contexts in all layers, thus achieving better semantic segmentation performances.

Second, as shown in Table 6, we explain why Contextrast uses the representative anchor information in all layers. When the representative anchor information was used in partial layers, it showed worse semantic segmentation performance than the proposed method that utilized the representative anchor information in all layers. Therefore, the representative anchor of the highest layer should be shared in all layers to align features consistently with global context information.

Third, Table 7 demonstrates that the proposed contrastive learning method slightly increases complexity and memory cost during the training phase. However, our approach does not impose additional burdens during inference, aligning with our objective for efficiency.

Lastly, we identified the optimal λ_i with various combinations of hyperparameters as detailed in Table 8a. Upon optimizing λ_i, we adjusted α to balance the scale between cross-entropy loss and PA loss as shown in Table 8b. Despite non-optimized hyperparameters, our method's performance surpassed that of state-of-the-art methods, achieving an 83.14 mIoU, as shown in Table 1 in our manuscript, except for one experiment that mostly utilizes low-level features as demonstrated in the fifth-row of Table 8a.

	Layer 4	Layer 3	Layer 2	Layer 1	mIOU (%)
Cityscapes	✓				81.15 (-1.05)
	✓	✓			81.14 (-1.06)
	✓	✓	✓		81.52 (-0.68)
	✓	✓	✓	✓	**82.20 (Ours)**
CamVid	✓				83.42 (-0.91)
	✓	✓			83.13 (-1.20)
	✓	✓	✓		83.17 (-1.16)
	✓	✓	✓	✓	**84.33 (Ours)**

Table 6. Ablation study: performance according to the layers that utilize the representative anchor of the last layer. Our proposed method demonstrates the best semantic segmentation performance with HRNet [7].

	Baseline	Contextrast	Increase rate (%)
Params (M)	70.01	70.39	+0.54
FLOPs (G)	1295.86	1300.13	+0.33
Training time (sec/epoch)	352.53	415.63	+17.90

Table 7. Computational complexity and memory usage in the training phase.

$\lambda_{4\to1}$				mIoU
1.0	1.0	1.0	1.0	83.63
1.0	0.8	0.6	0.4	83.53
1.0	0.75	0.5	0.25	83.35
1.0	0.7	0.4	0.1	84.33
0.1	0.4	0.7	1.0	82.92
0.25	0.5	0.75	1.0	83.42
0.4	0.6	0.8	1.0	83.53

(a)

α	mIoU
0.1	84.33
0.2	83.75
0.3	83.59
0.4	83.31
0.5	83.25

(b)

Table 8. Comparison with different hyperparameter settings with CamVid dataset.

8. Limitation Analysis

This paper proposed Contextrast, which utilizes representative anchors in a hierarchical structure. Thus, it enables sharing the global context of high-level features in each layer. It mostly achieved state-of-the-art performance on public datasets, but the improvements are not as large in COCO-Stuff [2] and PASCAL-C [5] as in CamVid [1], Cityscapes [3], and ADE20K [10]. We believe that there is a limitation in having generalized representative anchors in the last layer because some datasets have so many different classes in the scene; Contextrast only has limited features for each class with a limited training batch size. In the future, we plan to research further on how to generalize the representative anchor in many datasets without increasing training batch size.

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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.