Multi-group Vision Semantic Centroid for Semantic Segmentation

Jialei Chen[2][0009−0005−0654−4281], Daisuke Deguchi[2][0000−0003−0603−8790], Chenkai Zhang[2][0000−0002−7258−272X], Zhenzhen Quan[1,2][0000−0002−3981−7128], S eigo Ito[2], and Hiroshi Murase[2][0000−0002−8103−9294]

[1] Nagoya University, Nagoya 464-8601, Japan
[2] Shandong University, Qingdao 266237, Shandong, China
{chen.jialei.s6, zhang.chenkai.d4, quan.zhenzhen.g7}@s.mail.nagoya-u.ac.jp
{ddeguchi, murase}@nagoya-u.jp
iseigo@vislab.is.i.nagoya-u.ac.jp

Abstract. Multi-media processing has achieved great success based on semantic segmentation. Semantic segmentation can be viewed as pixel-clustering based on semantic prototypes. However, existing methods focus more on consistent semantics while ignoring the consistency in vision, making this task challenging. Motivated by the success of discrete visual representation learning, we propose Multi-group Visual Semantic Centroid (MVSC) to better cluster the pixels while maintaining consistent semantics of the dense features for any image encoder. Specifically, we randomly initialize multiple groups of prototypes as multi-groups in visual space. The visual features are also randomly split into the same groups and forced to be aligned with the corresponding prototypes. Then these visual prototypes are projected into the semantic space and supervised by the same classifier as the dense features. Compared with existing methods, MVSC further considers the visual space and thus facilitates the task. Experimental results on COCO-Stuff show great improvements compared with previous methods.

Keywords: Category centroid · Representation learning · Semantic segmentation.

1 Introduction

For multi-media processing, semantic segmentation improves the user experience by enhancing the content understanding in many aspects, *e.g.*, AR/VR scenarios. Different from the image classification task that recognizes the category of the whole image [19, 24], semantic segmentation aims to classify each pixel into its right category [5, 6, 22].Since FCN [22] treats semantic segmentation as a pixel-level classification, semantic segmentation has entered into a new era.

Before the wide use of transformer [26], researchers try to enlarge the receptive field, *e.g.*, DeepLab series [6, 7], deformable convolution [11], and fuse the

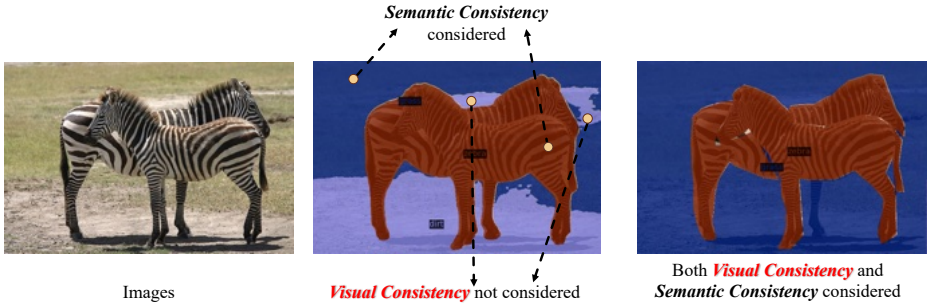

Fig. 1. Visual consistency and semantical consistency. Visual consistency indicates that the pixels belonging to the same object should belong to the same category. While the semantical consistency demonstrates that the pixels belonging to the same category should be segmented correctly. Existing methods pay much attention to the semantical consistency (the top caption in the middle image) while ignoring the visual consistency (the bottom caption in the middle image). Our methods consider both semantical and visual consistency.

multi-scale information, *e.g.*, FCN [22], PSPNet [32] to improve the segmentation performance. After the great success of ViT [13], designing transformer structures fitting for segmentation became popular, *e.g.*, Segformer [30], SETR [33].

When we revisit the semantic segmentation, we find that this task can be decoupled into two aspects: semantic consistency and visual consistency as shown in Fig. 1. Semantic consistency demonstrates that each pixel should be correctly classified, while visual consistency demonstrates that the pixels of one object should belong to the same category. Though remarkable, existing methods devoted all their efforts to maintaining semantical consistency, while ignoring visual consistency, leading to sub-optimal performance.

Motivated by the success of discrete visual representation learning [14,25], we propose Multi-group Visual Semantic Centroid (MVSC) to consider both visual and semantical consistency for any image encoder. Specifically, we first randomly initialize a group of vectors that contains the same number as the categories in the dataset, acting as the prototypes for each category in the vision space. Then the visual features from the feature extractor are also randomly separated into G groups as different groups and aligned with the corresponding visual category prototypes. Finally, the visual category prototypes are mapped to the semantic space and supervised by pixel-level annotations with the same classifier as the dense representation. To further model the visual diversity, we expand the Visual Semantic Centroid to Multi-group Visual Semantic Centroid by expanding one group of centroids to multiple groups.

Different from the existing methods that devote all the efforts to semantical consistency, our methods further consider visual consistency. Experimental results on COCO-Stuff [2] dataset show great performance improvement compared with previous methods in mIoU.

In summary, our contributions can be listed as follows:
1) We propose MVSC to entangle the vision and semantics.
2) We propose a novel way to optimize the MVSC.
3) Experimental results show great improvements on COCO-Stuff.

2 Related work

2.1 CNN-based Semantic Segmentation

As one of the most significant tasks in computer vision, semantic segmentation aims to classify each pixel into its right category. Based on this idea, FCN [22], as the first fully convolutional network for semantic segmentation, inspires the following researchers. For example, DeepLab series [6,7] enlarges the receptive field to further improve the performance. Deformable convolution [11] and Non-local network [28] further expand the receptive fields by breaking the fixed geometric structures of CNN modules. There are also some methods that utilize the multi-scale features to enhance the representation ability, especially for the small and the boundary of different categories. For example, FCN [22] and SegNet [1] add the features of different scales from the encoder while recovering the size of the output features. Though remarkable, existing CNN-based segmentation methods fail to consider visual consistency leading to sub-optimal performance.

2.2 Transformer-based Semantic Segmentation

Since the great success of the self-attention mechanism in natural language processing [26], more and more research endeavors try to transfer the success from language to vision. ViT [13] as one of the representative works proves the potential of transformer in computer vision. Since then, transformers are applied to many downstream tasks [3,8,9,21]. For semantic segmentation, transformer-based methods can be grouped into two categories: backbone design and sparse prediction. Backbone design methods try to design the backbone that fits the semantic segmentation, *e.g.*, Segformer [30], SETR [33]. Sparse prediction methods, *e.g.*, maskformer [9] and mask2former [8], leverage the property of the transformer decoder where a group of trainable queries is first initialized. During training, these queries are matched with the ground truth through bipartite matching, and the classification and mask prediction are decoupled. However, the queries are data-agnostic which leads to sub-optimal performance.

3 Method

3.1 Preliminaries and Method Overview

Visual and semantical consistency in semantic segmentation. Given a dataset $\mathbb{D} = \{\mathbf{X}_i, \mathbf{Y}_i\}_{i=0}^{M}$ where $\mathbf{X}_i \in \mathbb{R}^{H \times W}$ is the images, $\mathbf{Y}_i \in$ is the corresponding pixel-level annotation, i is the index of the image, and M indicates

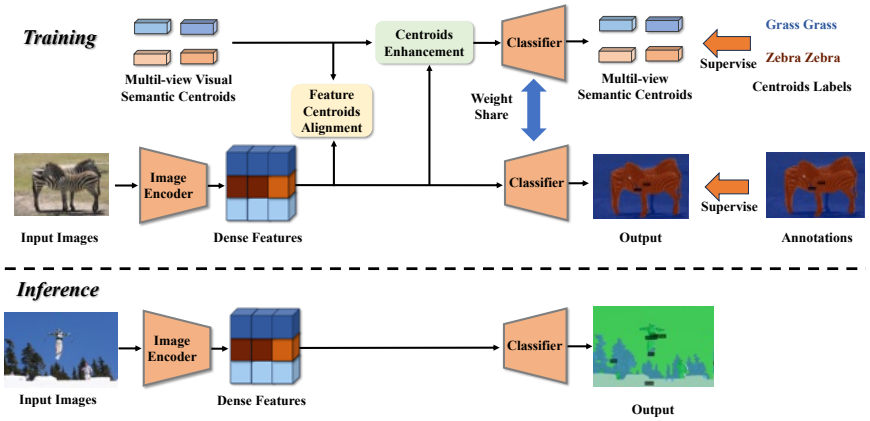

Fig. 2. The overview of the proposed methods. During training, We first initialize multi groups of vectors as the vision semantic centroids as multi-groups. Then, the image is fed into an image encoder to obtain the dense features. The dense features and the centroids are then aligned through the feature centroid alignment and centroid enhancement. Finally, the features and the centroids are put into the same classifier and supervised by the corresponding labels.

the size of the dataset. During training, $\mathbf{X}$ is first fed into an image encoder to obtain the pixel-level representation $\mathbf{R}^{D\times H\times W}$ where D indicates the channel numbers. Finally, $\mathbf{R}$ is fed into a trainable classifier $\mathbf{W}_c$ for prediction.

A very natural observation is that a baby does not know any category information about the world, but the baby can tell which objects belong to the same category. We call this property visual consistency. Then, an adult can teach the child what category this object belongs to and be agnostic about the environment. We call this property semantical consistency. Though many works try to improve the segmentation performance in many ways, *e.g.*, designing powerful backbones [17,30,33], enlarging the receptive fields [6,11] or utilizing multi-scale features [1, 22, 29], they still achieve sub-optimal performance due to the mere focus on semantic consistency and the ignorance on visual consistency. Similar works are maskformer and mask2former, however, the queries are data-agnostic, and our method is not.

Method overview. The overall structure is shown in Fig. 2. To consider the visual consistency, we propose Multi-group Visual Semantic Centroids (MVSC). First, we randomly initialize G groups of vectors as the category centroid as multi-groups in visual space, and each group contains the same number of categories in a dataset. Then, to ensure visual consistency, the visual features are aligned with the centroids through Feature Centroid Alignment. Next, inspired by the great success of discrete representation in image generation [14, 25], we enhance the MVSC by the centroid enhancement. Finally, the MVSC and the visual features are projected to the semantic space by the same classifier and

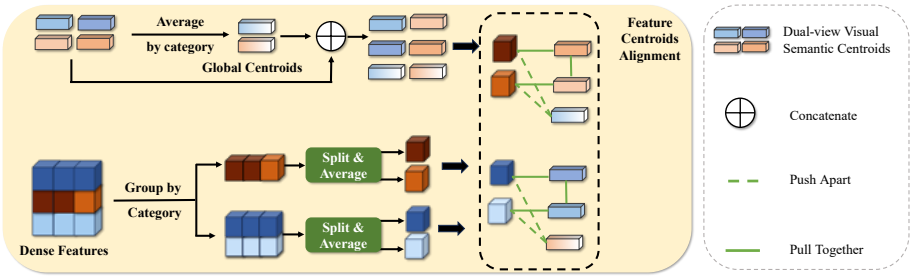

Fig. 3. The overview of Multi-group Visual Semantic Centroids (MVSC). Multiple groups of category centroids are first averaged within the same category across different groups to obtain global centroids. Next, the features are divided into the same groups as the category centroids and averaged based on their ground-truth labels. Finally, contrastive learning is applied to pull centroids from the same group closer together while pushing centroids from different groups further apart.

supervised by the corresponding category and pixel-level annotations. We will introduce the details in the following sections.

3.2 Feature Centroid Alignment

Before the training procedure, we randomly initialize G groups of vectors $\mathbf{V}_1, \mathbf{V}_2 ... \mathbf{V}_g$ as the centroids of each category in vision space where $\mathbf{V}_1, \mathbf{V}_2, ... \mathbf{V}_g \in \mathbb{R}^{C \times D}$ and C indicates the number of categories. Given one mini-batch of input $\mathbf{X}_b^{B \times H \times W}$ where B is the size of the mini-batch and the corresponding annotations $\mathbf{Y}_b$, suppose there are N unique categories in $\mathbf{X}_b$. First, based on $\mathbf{Y}_b$, we group the visual features $\mathbf{R}_b$ by the categories to obtain a set $\mathbb{C}$. Each element in $\mathbb{C}$ is:

$$\mathbb{C} = \{\mathbf{R}_n\} \text{ where } \mathbf{R}_n = \mathbf{R}_b \cdot \mathbb{1}(\mathbf{Y}_b = n) \tag{1}$$

where $n \in N$ indicates the nth category in $\mathbf{Y}_b$. Then the $\mathbf{R}_n$ is randomly split into the same groups and the elements in each group are averaged to obtain the centers of each group $\mathbf{R}_{n1}, \mathbf{R}_{n2} ... \mathbf{R}_{ng}$.

For each group of centroids, we also average them to obtain the global centroid $\mathbf{V}_G \in \mathbb{R}^{C \times D}$. The Feature Centroid Alignment can be represented as:

$$\mathcal{L}_{fca}(R, V) = \sum_{i}^{N} \frac{r_i^T \cdot v_i}{r_i^T \cdot v_i + \sum_{j \neq i}^{N} r_i^T \cdot v_j'}, \tag{2}$$

where $v' \in \mathbf{V}_G$ indicates the global centroid of a specific category. Besides, to further enhance the visual consistency, we further pull close the centroids belonging to the same category. Specifically, we randomly separate the centroids except the global centroid into two groups. Then we reduce the cosine similarity of the two groups:

$$\mathcal{L}_{cos}(V_1, V_2) = 1 - cos(V_1, V_2), \tag{3}$$

3.3 Centroid Enhancement

Though Feature Centroids Alignement improves the visual consistency, the Visual Semantic Centroids are still data-agnostic, leading to sub-optimal visual consistency. In this section, we propose a simple but effective Centroid Enhancement to solve this problem inspired by the great success of discrete representation in image generation [14, 25].

Specifically, after obtaining $\mathbf{R}_n$, we pick the centroids $\mathbf{R}'$ whose labels appear in $\mathbf{Y}_b$. Then we update the $\mathbf{V}'$ as:

$$\mathbf{V}' = \mathbf{V} + \mathbf{R} - SG(\mathbf{R}) \tag{4}$$

where SG indicates the stop gradient operation. Then, the augmented centroids are fed into the classifier and supervised by the corresponding category labels.

Discussions on Centroid Enhancement. Though simple, we want to discuss the reasons why this operation works. The first reason is that this operation inexplicitly extracts the region-level information. Even if the value of the centroid does not change, the gradient of the visual features that are assigned to the centroid is copied. The copied gradient helps the image encoder learn the region-level information. The second reason is that the copied gradient helps the upgrade of the vision semantic centroids. To make the analysis easier, we ignore the gradient of $\mathcal{L}_{fca}$. Without the copied gradient, the gradient for the visual semantic centroids is $\partial\mathcal{L}/\partial W_c$ where $\mathcal{L}$ indicates the total loss. With the copied gradient the new gradient for the visual semantic centroid is $\partial\mathcal{L}/\partial W_c \cdot (1 + \partial W_c/\partial E)$. Note that $\partial W_c/\partial E$ indicates the scale of how the encoder increases. Therefore, the visual semantic centroids can update at the same scale as the encoder and avoid the encoder's collapse to meaningless centroids.

3.4 Training Objectives and Inference

During training, the total loss functions are listed as follows:

$$\mathcal{L} = \lambda_f \cdot \mathcal{L}_{fca} + \lambda_c \cdot \mathcal{L}_{cos} + \mathcal{L}_{ce} + \lambda_v \cdot \mathcal{L}_v \tag{5}$$

where λ_c, λ_f and λ_v are hyperparamters to control the scale of the corresponding loss. During Inference, as shown in Fig. 2 (bottom), only the image encoder and the classifier is needed for the final output.

4 Experiments

4.1 Experiment Setup

Dataset and Implementation Details. We conducted experiments on the COCO-Stuff dataset [2] where 9K images are used for training and 1K images are used for evaluating the final performance. For the categories, this dataset contains 80 object categories, *e.g.*, person, car, and 91 stuff categories, *e.g.*, grass, wall. For semantic segmentation, all the 171 categories are used.

Table 1. Comparison with State-of-the-Art Methods on COCO-Stuff Dataset where the highest performance is highlighted in **bold**.

Publication	Method	Backbone	mIoU	Gain
CVPR15	FCN [22]	RN101 [19]	33.0	+10.6
CVPR19	SVCNet [12]	RN101 [19]	39.6	+4.0
ECCV18	DANet [15]	RN101 [19]	39.7	+3.9
ICCV19	SpyGR [20]	RN101 [19]	39.9	+3.7
ICCV19	ACNet [16]	RN101 [19]	40.1	+3.5
ECCV20	OCR [31]	HRNetV2 [27]	40.5	+3.1
NeurIPS21	MaskFormer [9]	RN101 [19]	39.8	+3.8
TPAMI21	HRNet [27]	HRNetV2 [27]	38.7	+4.9
CVPR22	ProtoSeg [35]	Swin-B [21]	42.4	+1.2
CVPR22	ProtoSeg [35]	MiT-B4 [30]	43.3	+0.3
CACML24	CM [4]	MiT-B4 [30]	43.2	+0.4
NeurIPS21	SegFormer [30]	MiT-B4 [30]	42.9	+0.7
	SegFormer + Ours		**43.6**	

We use MMsegmentation [10] as the base to complete our algorithm. For each experiment in one dataset, we follow the default settings. Specifically, for COCO-Stuff, all the backbones are first pre-trained on ImageNet1K [23] and the decode head initialized by He initialization [18]. We rescale the short scale of the image to train crop size while keeping the aspect ratio unchanged. Random scale jittering with a factor in [0.5, 2], random horizontal flipping, random cropping, and random color jittering are applied as the data augmentation. The optimizer is AdamW. The learning rate is scheduled following the polynomial annealing policy. In addition, the batch size is 16 and the crop size is 512 $\times$ 512 pixels. 80K iterations are needed to obtain the best performance. τ is set to 0.07. λ_f is set as 0.1 and λ_c is set as 1, and λ_v is set as 0.05 in training.

For simplicity, we do not apply any test-time data augmentation. Our model is implemented in PyTorch and trained on 8 Tesla V100 GPUs with 32GB memory per card. Testing is conducted on the same machine. We report each model's mean intersection over the union (mIoU) score.

4.2 Comparasion with State-of-the-Art

We conduct experiments on the COCO-Stuff dataset to compare the performance with other methods as shown in Table 1. From the table, we observe that traditional methods like FCN and SVCNet, which utilize the ResNet-101 backbone, achieve mIoUs of 33.0% and 39.6%, respectively. More recent methods, such as MaskFormer and ProtoSeg, demonstrate improvements, with mIoUs of 39.8 and 43.3%. Notably, SegFormer, which uses the MiT-B4 backbone, achieves a strong performance with a mIoU of 42.9%. First, we compare the performance with the baseline method Segformer [30]. For Segformer-B4, they achieve 42.9%

Table 2. Ablation on the CE where 'CE' indicates the centroid enhancement.

$\mathcal{L}_{cfa}$	$\mathcal{L}_{cos}$	$\mathcal{L}_{v}$	**CE**	**mIoU**	**mAcc**
✓	✓	✓	✓	**32.6**	**44.0**
-	✓	✓	✓	31.4	41.8
✓	-	✓	✓	32.2	43.0
✓	✓	-	✓	32.3	43.3
✓	✓	✓	-	32.1	42.8
-	-	-	-	30.9	41.1

Table 3. Ablation on the group of prototypes.

Group Num	Backbone	mIoU	mAcc
1	Segformer-B0	32.2	43.1
2		32.0	43.4
4		**32.6**	**44.0**
8		32.3	43.8

mIoU. After adding our proposed methods, the performance boosts up to 43.6%, which is a 0.7% improvement, indicating our effectiveness.

Compared with other SOTA methods, *e.g.*, ProtoSeg [35], under the same backbone, *i.e.*, Segformer-B4, our method can still be 0.3% higher than them. When we compare one of the latest methods, *i.e.*, CM [4], we can achieve 0.4% higher mIoU than them. In summary, Our approach achieves the highest mIoU score of 43.6%. This demonstrates the superiority of our method compared to both classical and modern segmentation techniques.

Visualization of Predictions. We further visualize the prediction of our model as shown in Fig. 6. We can find that our method has very similar results with the ground truth label indicating the effectiveness of our method.

4.3 Ablation Study

To evaluate whether each proposed method can work as expected, we conduct ablation experiments on all the parts of the proposed methods. Note that for the ablation studies, to further indicate the generalization of our method, the dataset we use is ADE20K dataset [34] which is comprised of 150 categories to be segmented. For this dataset, 20K images are used for training and 2K images for evaluation. The model we use is Segformer-B0 and trains 20K iterations. λ is set as 0.05 and γ is set as 0.1 during training.

First, we conducted a comprehensive evaluation of the effectiveness of each component in our model, with the results detailed in Table 4.3. The baseline model, which serves as our point of comparison, does not incorporate any of the proposed enhancements and achieves a baseline performance of 30.9% mIoU and 41.1% mAcc. This serves as a foundational reference for understanding the impact of each subsequent modification.

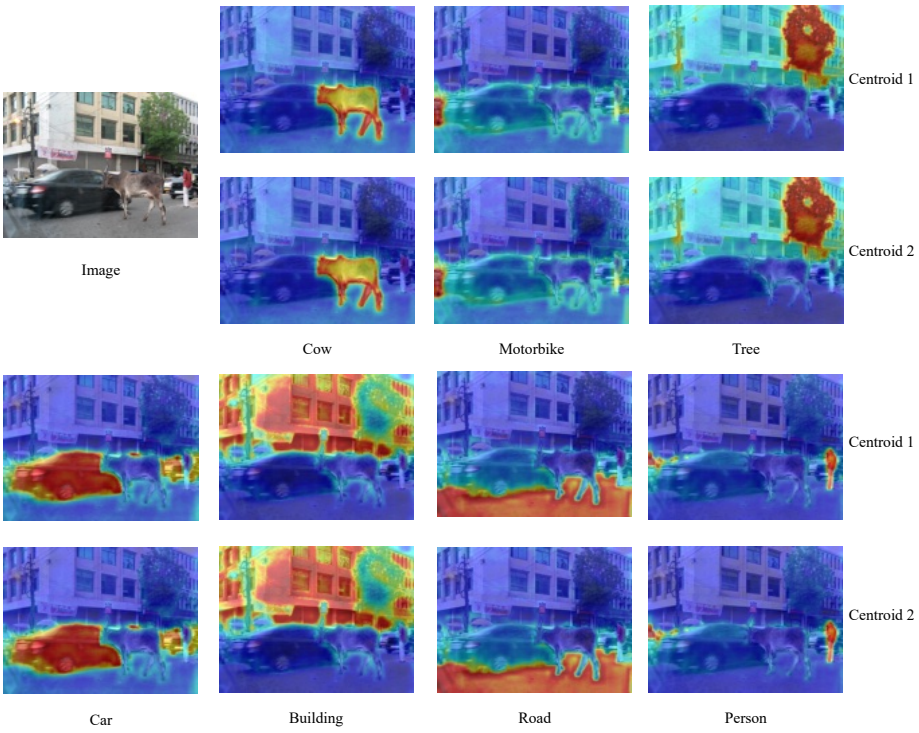

Fig. 4. Visualization of centroid heatmap where the red color indicates that the pixel is similar to the centroid and the blue color indicates the opposite.

As illustrated in the table, when all the proposed methods are integrated, our model achieves its peak performance, with mIoU rising to 32.6% and mAcc to 44.0% (first row). This demonstrates the synergistic effect of combining our techniques, yielding substantial improvements in both segmentation accuracy and mean accuracy. To better understand the contribution of each individual component, we systematically ablated them one by one. First, we removed the centroid-feature alignment mechanism, which led to a noticeable decrease in performance, with mIoU dropping to 31.4% and mAcc to 42.4% (second row). This indicates the critical role that clustering plays in organizing and refining the visual space, ensuring that similar features are grouped effectively, which directly enhances segmentation performance. Next, we evaluated the impact of the centroid-pulling mechanism, specifically the loss function $\mathcal{L}_{cos}$, which is designed to ensure the compactness of centroids in the feature space. When this component was ablated, the model's performance suffered a significant reduction, with mIoU decreasing by 0.4% and mAcc by 1.0%. This substantial drop underscores the importance of maintaining tight, well-defined clusters in the feature space, which is crucial for accurate and consistent segmentation. Following this, we examined the effect of removing the projection back to the semantic

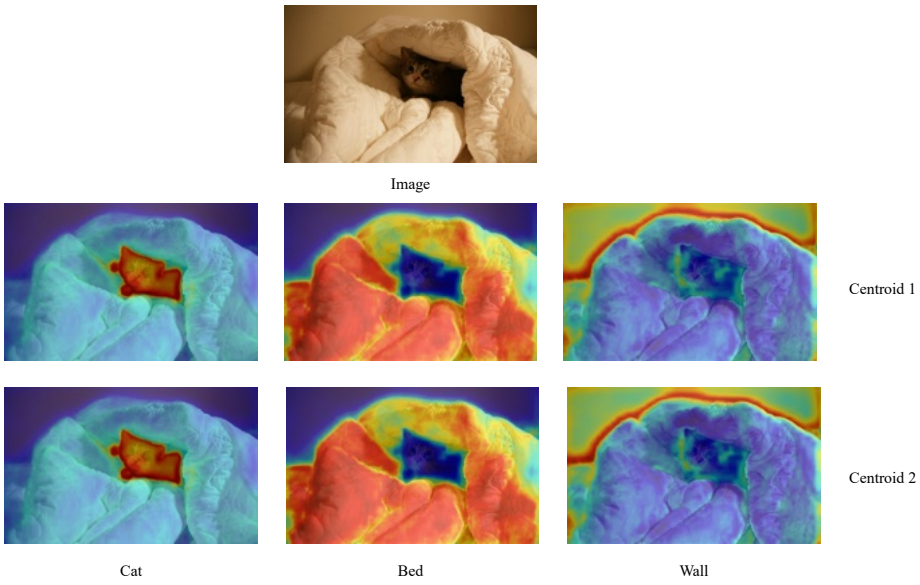

Fig. 5. Visualization of centroid heatmap where the red color indicates that the pixel is similar to the centroid and the blue color indicates the opposite.

space, represented by $\mathcal{L}_v$. The absence of this component led to a decline in performance, with mIoU falling to 32.2% and mAcc to 43.3%. This result highlights the necessity of semantic consistency, as projecting features back to the semantic space ensures that the learned features remain aligned with the semantic labels, thereby preserving the interpretability and accuracy of the segmentation output. Lastly, we considered the effect of removing the centroid enhancement technique. Without this enhancement, the model's performance further declined, with mIoU reduced to 32.1% and mAcc to 42.8%. This demonstrates the value added by the centroid enhancement process, which strengthens the representation of feature centroids and thus contributes to the overall robustness and precision of the segmentation model. Collectively, these ablation studies provide strong evidence for the effectiveness of each proposed method. The consistent improvements across various metrics, when these methods are applied, affirm their importance in achieving superior segmentation performance. Each component plays a vital role in refining the model's ability to accurately segment and classify objects within complex visual scenes, leading to SOTA results.

Next, we conduct experiments on the group number of prototypes, and the results are shown in Table 3. As can be seen from this table, when the group number is set as 4, the model can reach its best mIoU and mAcc performance to 32.6% and 44.0%. Note that when multi-group strategy is not used, *i.e.*, the group is set as 1, the performance drops to 32.2% for mIoU and 43.1% to mAcc indicating the effectiveness of multi-group strategies.

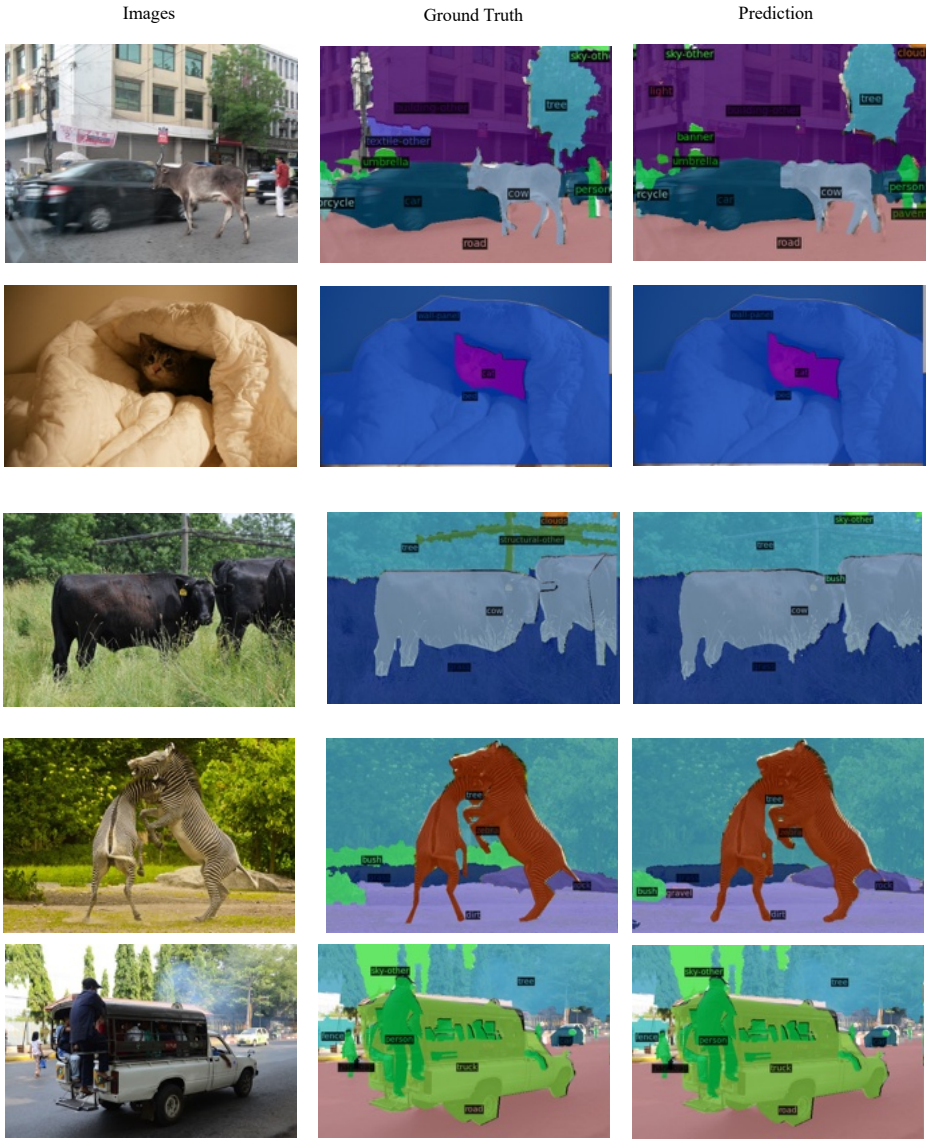

Fig. 6. Visualization of the prediction.

4.4 Qualification Results

Visualization of centroid heatmap where the red color indicates that the pixel is similar to the centroid and the blue color indicates the opposite. First, we visualize the centroid to validate if the features are clusters in the visual space. We randomly choose two images as the visualization sam-

ple. Next, we compute the cosine similarity between the dense features and the centroid by group and visualize it. The results are shown in Fig. 4 and Fig. 5. We randomly choose two of the four prototypes to visualize. From Fig. 4, we can find that no matter the large objects, *e.g.*, buildings and roads, or small objects, *e.g.*, person and motorbike, the visual features are all very compact with the centroids. In Fig. 5, even the categories with similar appearance, *i.e.*, wall and bed, the visual features can be grouped into the right category, indicating the effectiveness of our methods. **Visualization of Predictions.** Finally, we visualize the predictions of our model as shown in Fig. 6. First, we give the prediction of the images in Fig. 4 and 5. As can be seen from the figure, the corresponding objects can be segmented correctly. Besides, we give another 4 images and all of them can achieve satisfying results.

5 Conclusion

In this paper, we presented a Multi-group Visual Semantic Centroid (MVSC) method, aimed at improving semantic segmentation performance for any image encoder. MVSC is motivated by the consideration that semantic segmentation should consider both visual consistency and semantic consistency. To achieve these two consistencies, we propose Feature Centroid Alignment and Centroids Enhancement to optimize the DSVC. Feature Centroid Alignement clusters the dense features in the visual space and Centroids Enhancement helps the centroids be updated at the same scale as the image encoder. Through extensive experiments and comparative analysis with state-of-the-art methods, our approach demonstrated superior performance, achieving a new benchmark mIoU of 43.6. This result not only surpasses the original SegFormer model by a significant margin of +0.7 mIoU but also outperforms other leading segmentation models. Future work could explore the application to additional segmentation tasks, further solidifying its role in next-generation computer vision systems.

6 Acknowledgment

This work was supported by the annual project funding of the Smart State Governance Laboratory, Shandong University, and JSPS KAKENHI Grant Number 23K28164 and 24H00733, and JST CREST Grant Number JPMJCR22D1. All the computations of this paper are carried out on the supercomputer “Flow” at the Information Technology Center, Nagoya University.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)

2. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
4. Chen, J., Deguchi, D., Zhang, C., Murase, H.: Centroid module for shaping feature space in semantic segmentation. In: Proceedings of the 2024 3rd Asia Conference on Algorithms, Computing and Machine Learning. p. 71–75. CACML '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3654823.3654837, https://doi.org/10.1145/3654823.3654837
5. Chen, J., Deguchi, D., Zhang, C., Zheng, X., Murase, H.: Frozen is better than learning: A new design of prototype-based classifier for semantic segmentation. Pattern Recognition **152**, 110431 (2024)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: 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)
7. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
8. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
9. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems **34**, 17864–17875 (2021)
10. Contributors, M.: Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark (2020)
11. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
12. Ding, H., Jiang, X., Shuai, B., Liu, A.Q., Wang, G.: Semantic correlation promoted shape-variant context for segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8885–8894 (2019)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
15. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3146–3154 (2019)
16. Fu, J., Liu, J., Wang, Y., Li, Y., Bao, Y., Tang, J., Lu, H.: Adaptive context network for scene parsing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6748–6757 (2019)

17. Guo, M.H., Lu, C.Z., Hou, Q., Liu, Z., Cheng, M.M., Hu, S.M.: Segnext: Rethinking convolutional attention design for semantic segmentation. Advances in Neural Information Processing Systems **35**, 1140–1156 (2022)
18. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
20. Li, X., Yang, Y., Zhao, Q., Shen, T., Lin, Z., Liu, H.: Spatial pyramid based graph reasoning for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8950–8959 (2020)
21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
22. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
23. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**, 211–252 (2015)
24. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
25. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
27. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3349–3364 (2020)
28. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
29. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018)
30. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34** (2021)
31. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. pp. 173–190. Springer (2020)
32. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)

33. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6881–6890 (2021)
34. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)
35. Zhou, T., Wang, W., Konukoglu, E., Van Gool, L.: Rethinking semantic segmentation: A prototype view. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2582–2593 (2022)