Hunting Sparsity: Density-Guided Contrastive Learning for Semi-Supervised Semantic Segmentation

Xiaoyang Wang[1,2,3] Bingfeng Zhang[4] Limin Yu[1] Jimin Xiao[1*]
[1]XJTLU [2]University of Liverpool [3]Metavisioncn [4]China University of Petroleum (East China)
wangxy@liverpool.ac.uk, bingfeng.zhang@upc.edu.cn, {limin.yu, jimin.xiao}@xjtlu.edu.cn

1. Overview

This supplementary material first presents a summary of the workflow of DGCL. Then it provides more details for the proposed memory bank to show its settings and the updating rules, which is not mentioned in the original paper. It also shows visual results on Cityscapes [1] dataset. More ablation studies on hyper-parameters are reported. Finally, we discuss the limitation of our DGCL and directions for future exploration.

2. Workflow of Feature Contrast

This section presents the workflow of performing density-guided feature contrast across classes in each mini-batch. The whole process is summarized in Algorithm 1. Note that the equation indices in the algorithm refer to those in the original paper.

Algorithm 1: Feature contrast in each iteration

Input:
(x^l, y^l): labeled images and ground truth
$(x^u, \hat{y}^u)$: Unlabeled images with filtered pseudo labels
$\mathcal{P}$: Categorical memory banks
Output: Updated student model

Extract features: $\mathcal{V} \leftarrow h([x^l, x^u])$;
Initialize contrastive loss: $\mathcal{L}_{contra} \leftarrow 0$;
for $c \in \mathcal{C}$ **do**
  Get in-batch class features: $\mathcal{V}^c \leftarrow \{v_i | y_i = c\}$;
  Build nearest neighbor graphs for each v^c with $\mathcal{P}^c$;
  Calculate density $\{d(v^c)\}$ using Eq. (7) and Eq. (8) ;
  Sample anchors $\mathcal{Q}^c$ from $\mathcal{V}^c$ using Eq. (9);
  Sample positive keys $\mathcal{R}^{c,+}$ from $\mathcal{V}^c$ and $\mathcal{P}^c$ using Eq. (10) and Eq. (11) ;
  Randomly sample out-of-class negative keys $\mathcal{R}^{c,-}$;
  Update feature memory $\mathcal{P}^c$ with $\{(v^c, d(v^c))\}$;
  Contrastive loss on projected features: $\mathcal{L}_{contra} \leftarrow \mathcal{L}_{contra} + \ell(g(\mathcal{Q}^c), g(\mathcal{R}^{c,+}), g(\mathcal{R}^{c,-}))$;
end

*Corresponding author.

3. More Details of Memory Bank

The memory bank consists of a collection of feature representations and their corresponding density values. The feature collection is in the form of a $C \times N \times D$ matrix where C refers to the number of classes. N denotes the number of features per class, and we set its value as 10000. D is the dimension of each feature vector which is 256 in this work. The density value collection is in the form of a $C \times N$ matrix, which matches the first two dimensions with feature memory.

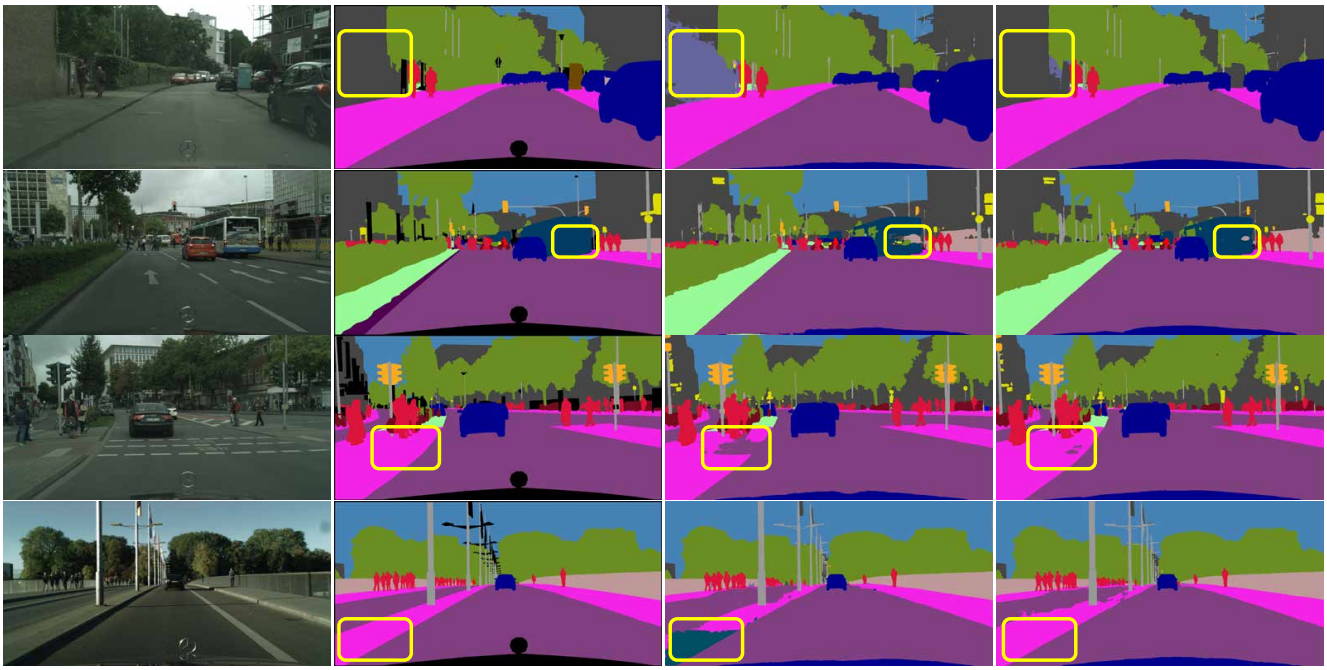
During training, we set a threshold N_{memo} to control the pace of memory updating across mini-batches. It means, for each class in one mini-batch, at most N_{memo} samples can be extracted to update its corresponding memory. We perform random sampling on features if their quantity is above the threshold. We set N_{memo} as 1000 in this work, which means at least 10 mini-batches per class are absorbed in the memory to guarantee the diversity in memory. The features $\{v\}$ are updated into the memory along with their density values $\{d(v)\}$. Note that we estimate $\{d(v)\}$ in feature memory before getting $\{v\}$ absorbed. In such a setting, feature density is always estimated without in-batch features to guarantee robust estimation under feature-to-memory style.

4. Qualitative Results on Cityscapes

We present qualitative results on Cityscapes [1] *val* set in Fig. 1. It can be observed that the baseline model predicts poorly on specific categories such as *building* and *sidewalk*. Compared with the baseline, the model equipped with DGCL can make more accurate predictions and generate cleaner masks with less noise.

5. Additional Ablation Studies

This section reports additional ablation studies on hyper-parameters. The following indices for equations and sections refer to those in the original paper. Tab. 1 shows the influence of different weights of contrastive loss denoted as λ_{contra} in Eq. (2). Results for different temperature coeffi-



(a) Images (b) Ground Truth (c) Baseline (d) Ours

Figure 1. Qualitative results on Cityscapes [1] *val* set based on model trained on 1/8 (372) data set. The baseline model is trained solely with self-training strategy. Ours with additional DGCL strategy shows overall better qualitative results.

Table 1. Ablation study on weight of contrastive loss λ_{contra} under PASCAL VOC 2012 *classic* 1/4 (366) and 1/2 (732) data splits.

λ_{contra}	0.1	0.5	1	1.5	2
1/4 (366)	78.53	78.56	**78.73**	78.47	78.21
1/2 (732)	79.12	78.88	**79.23**	78.69	78.53

Table 2. Ablation study on temperature coefficient τ in contrastive loss under PASCAL VOC 2012 *classic* 1/4 (366) and 1/2 (732) data splits.

τ	0.01	0.1	0.5	1
1/4 (366)	74.61	78.54	**78.73**	78.54
1/2 (732)	75.50	78.73	**79.23**	79.08

cients τ in Eq. (13) are shown in Tab. 2. Tab. 3 ablates the initial entropy percentile β_0 in Section 3.3. Tab. 4 studies the number of anchors N_q in Section 3.4.2. The experiments are conducted on PASCAL VOC 2012 [2] dataset.

6. Limitations and Future Work

The proposed density-guided contrastive learning strategy has shown its superiority, but its connection to consistency regularization is not fully explored. Further exploration should focus on embedding DGCL deeply into the consistency-regularization-based framework, for example, to apply DGCL on the perturbed samples to explore possible improvement.

Table 3. Ablation study on initial entropy percentile β_0 which is used to filter out noisy predictions. Results are under *classic* 1/4 (366) set in PASCAL VOC 2012 dataset.

β_0	0%	20%	40%	60%	80%
mIoU	78.45	**78.73**	77.67	75.93	74.63

Table 4. Ablation study on N_q which is the number of anchors per class each mini-batch. Results are under *classic* 1/4 (366) set in PASCAL VOC 2012 dataset.

N_q	32	64	128	256	512
mIoU	77.74	77.73	77.22	**78.73**	77.88

References

[1] 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 *CVPR*, 2016. 1, 2

[2] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2009. 2