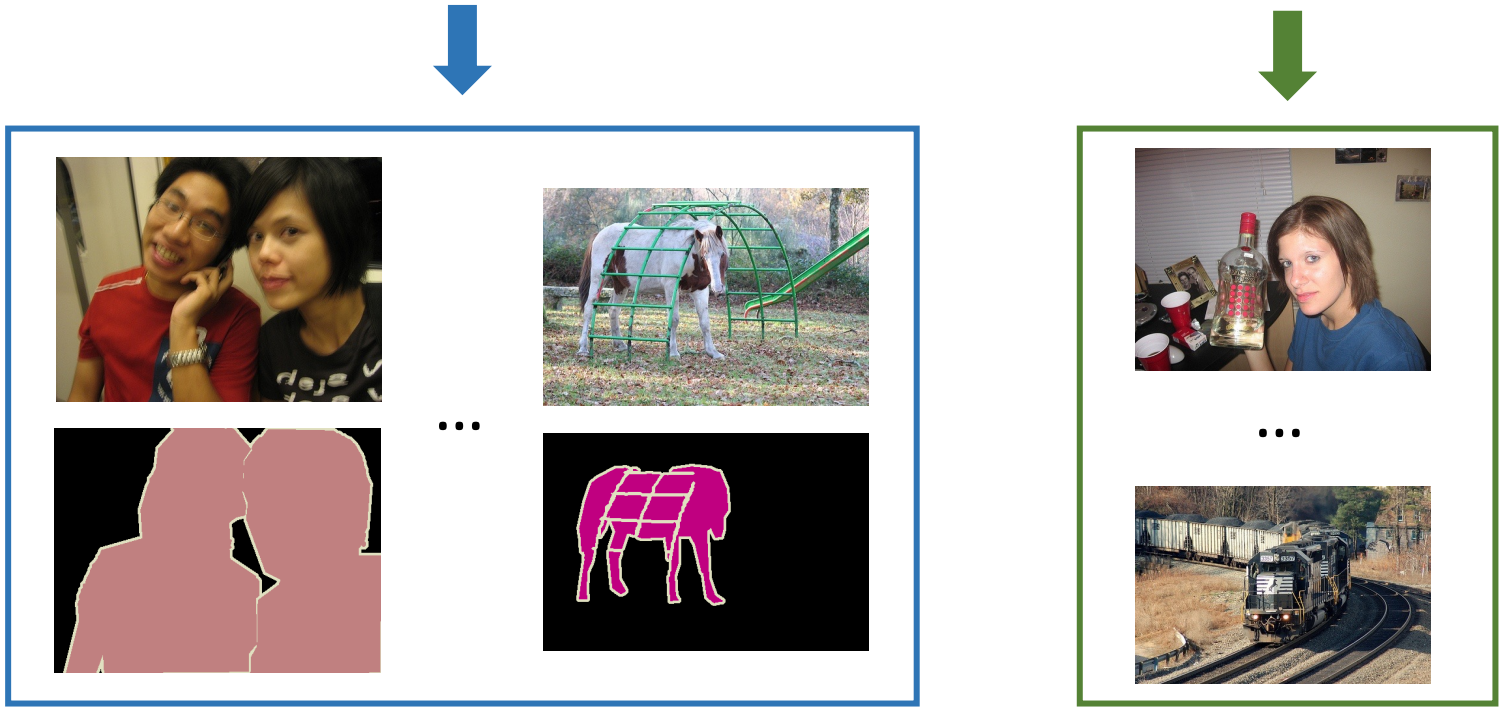
Hunting Sparsity: Density-Guided Contrastive Learning for Semi-Supervised Semantic Segmentation

Xiaoyang Wang, Bingfeng Zhang, Limin Yu, Jimin Xiao

Semi-supervised Semantic Segmentation

- Limited annotated data + Massive unlabeled data

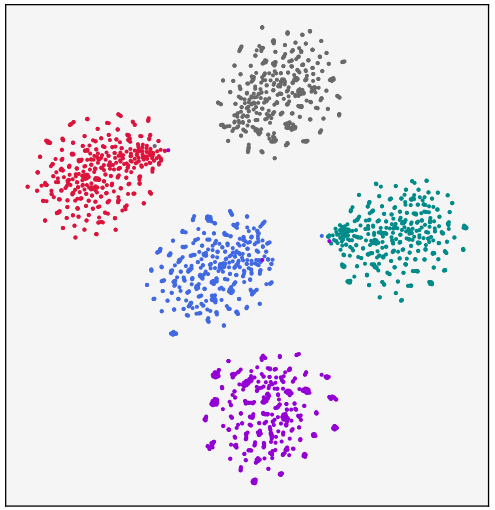


- Training for better generalization performance

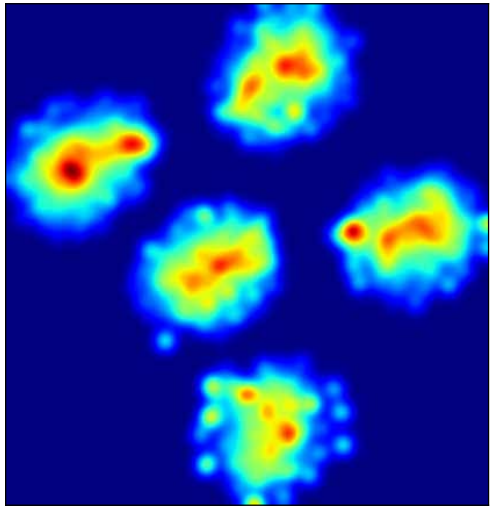
Where supervision comes from ?

On unlabeled data:

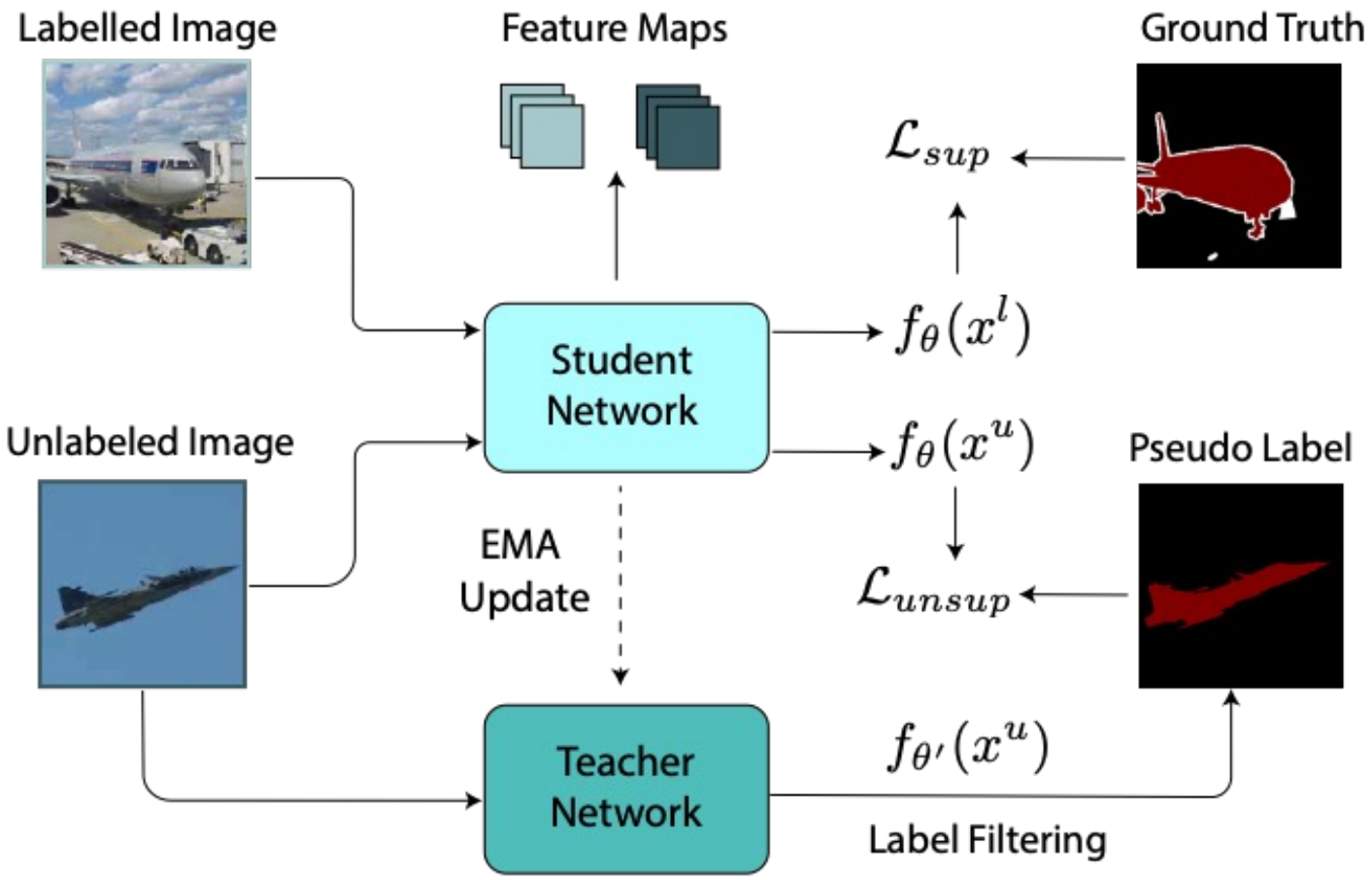
- Model-generated pseudo labels
- Feature local density



- Among confidently predicted features:
  - High density features are more likely to be cluster centers
  - Low density features are less representative
  - High-density features attract low-density features to form compact clusters



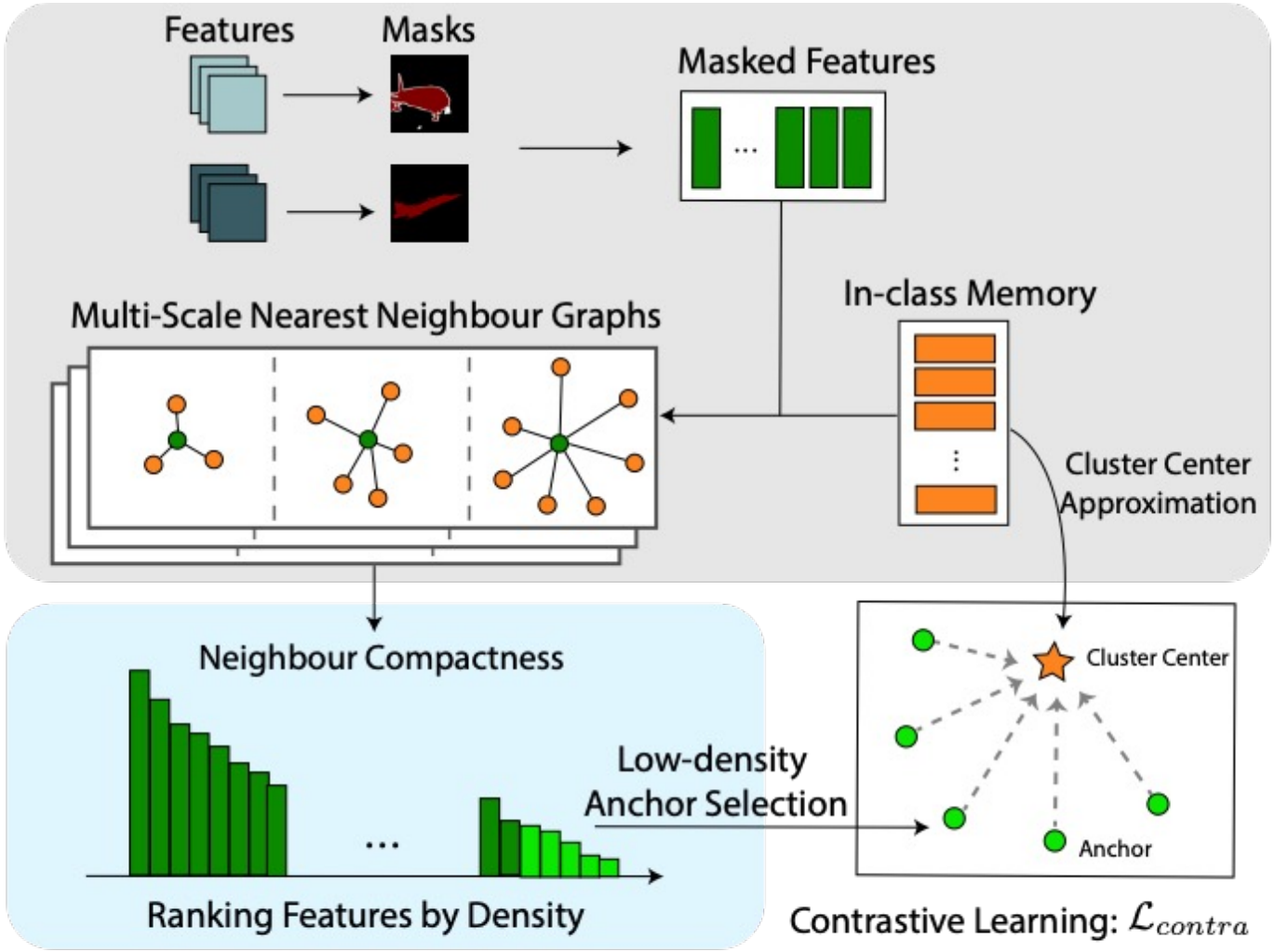
Label-Guided Pixel Classification



Corss-entropy loss on:

- Labeled Images
- Unlabeled data with filtered pseudo labels (Low entropy predictions)

Density-Guided Feature Contrast

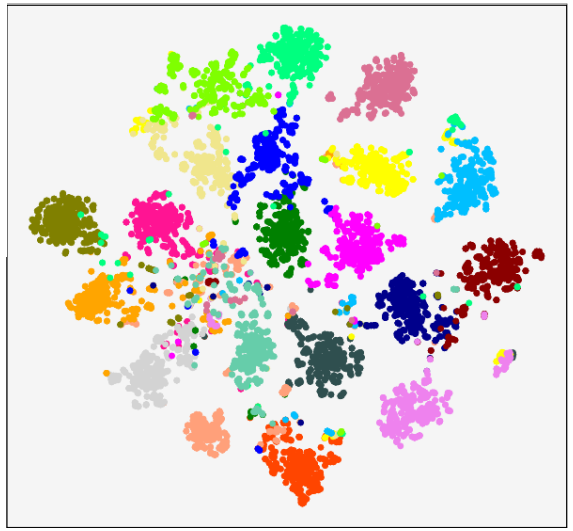


- Memory bank of categorical features to model class distribution
- Multi-scale KNN to capture robust local feature density
- Contrastive learning to form more compact categorical clusters

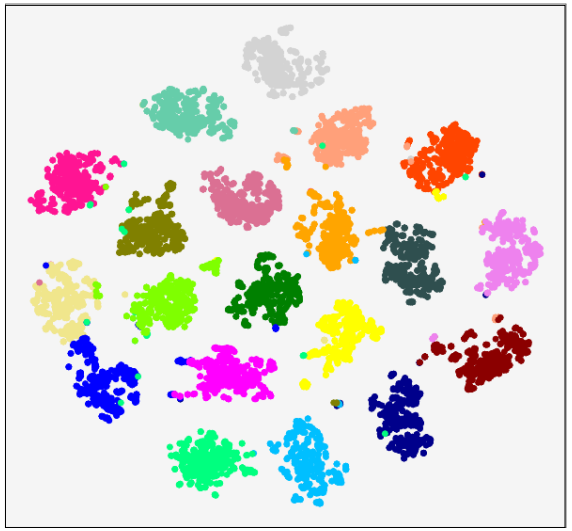
Impact on feature clustering

	Silhouette ↑	Calinski-Harbasz ↑	Davies-Boulding ↓
Baseline	0.46	3421.94	1.53
DGCL	**0.70**	**7937.87**	**1.13**

- Baseline: Plain self-training
- DGCL: Self-training with density-guided contrastive learning



(a) Baseline method



(b) DGCL

Ablations on sampling strategy

Anchors	Positive Keys	183	1323
Random	Average	71.82	76.73
Low Conf	High Conf	71.42	77.11
Low Denisty	High Density	**77.14**	**78.37**

Pulling low-density samples towards high-density samples achieves the most effective training.

Comparison with state-of-the-art

DGCL achieves overall best performance compared with previous SOTA

Table 2. Comparison with state-of-the-art methods on PASCAL VOC 2012 *val* set with mIoU results (%) ↑. Methods are trained under *blended* setting. Labeled images are randomly sampled from the extended training set, which consists of 10582 samples.

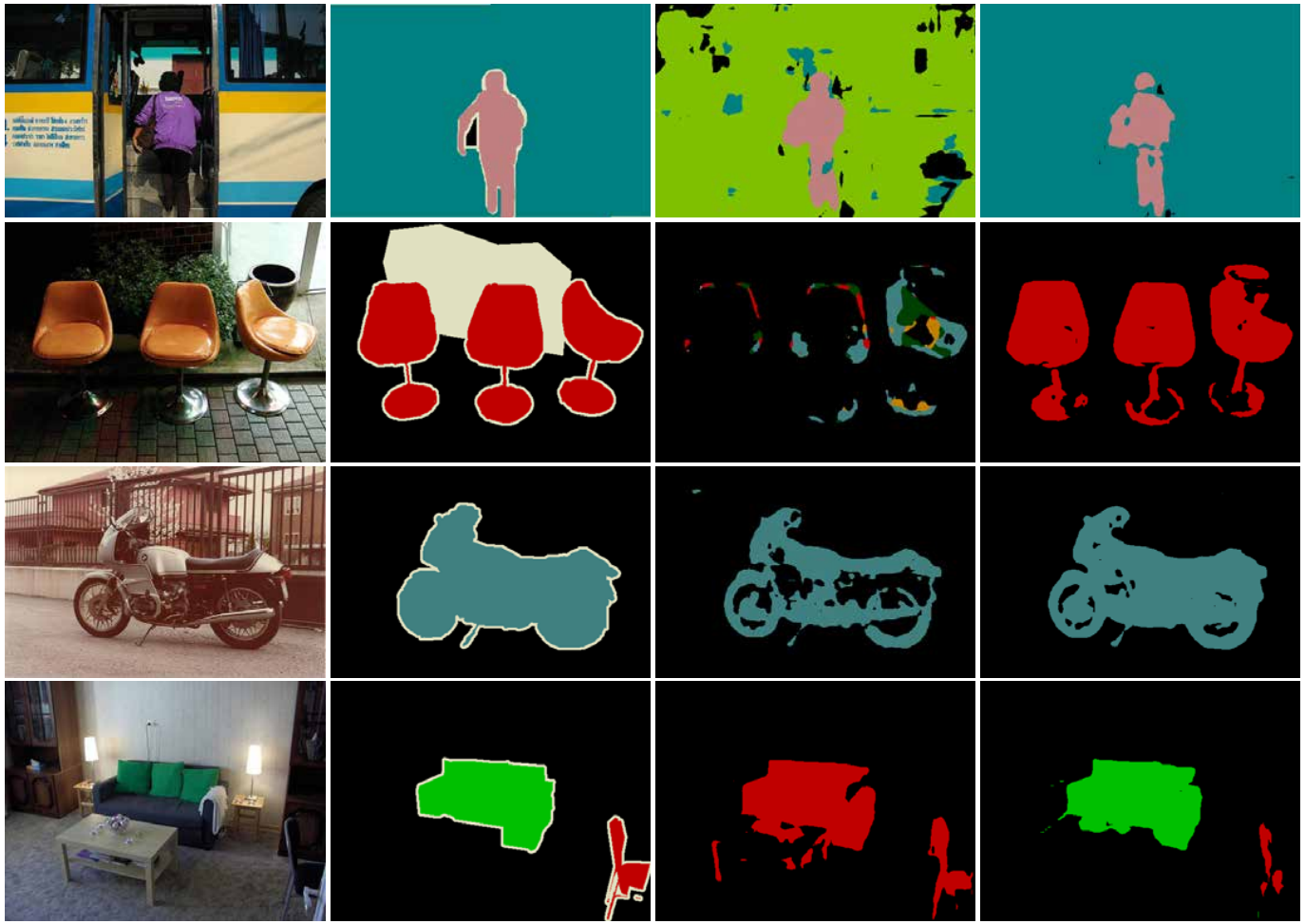
Method	1/16 (662)	1/8 (1323)	1/4 (2646)	1/2 (5291)
Supervised	67.87	71.55	75.80	77.13
MT [38]	70.51	71.53	73.02	76.58
CutMix [13]	71.66	75.51	77.33	78.21
CCT [32]	71.86	73.68	76.51	77.40
GCT [23]	70.90	73.29	76.66	77.98
CPS [7]	74.48	76.44	77.68	78.64
U^2PL^* [40]	74.43	77.60	78.70	79.94
PS-MT [29]	75.50	78.20	78.72	79.76
Ours	**76.61**	**78.37**	**79.31**	**80.96**

$*$ denotes that the results are reproduced with CPS [7] splits.

Table 3. Comparison with state-of-the-art methods on Cityscapes *val* set with mIoU results (%) ↑. Labeled images are selected from Cityscapes *train* set, which contains 2975 samples.

Method	1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)
Supervised	65.74	72.53	74.43	77.83
MT [38]	69.03	72.06	74.20	78.15
CutMix [13]	67.06	71.83	76.36	78.25
CCT [32]	69.32	74.12	75.99	78.10
GCT [23]	66.75	72.66	76.11	78.34
CPS [7]	69.78	74.31	74.58	76.81
U^2PL [40]	70.30	74.37	76.47	79.05
PS-MT [29]	-	76.89	77.60	79.09
PCR [42]	**73.41**	76.31	78.40	79.11
GTA-Seg [22]	69.38	72.02	76.08	-
Ours	73.18	**77.29**	**78.48**	**80.71**

Qualitative Results



Images Ground Truth Self-training Self-training + DGCL