

TransFGU: A Top-down Approach to Fine-Grained Unsupervised Semantic Segmentation



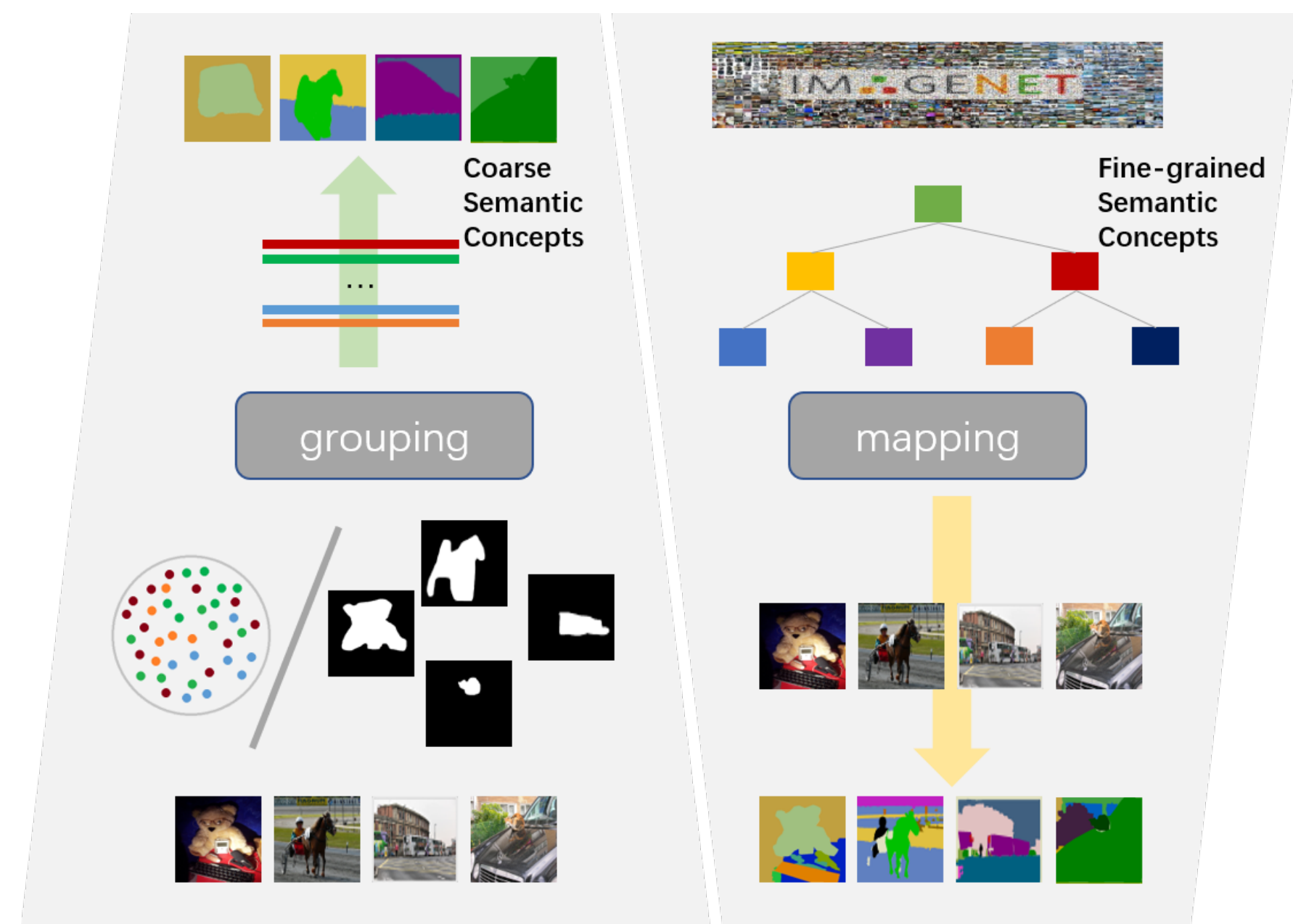
Zhaoyuan Yin¹, Pichao Wang², Fan Wang², Xianzhe Xu², Hanling Zhang³, Hao Li² Rong Jing²

¹College of Computer Science and Electronic Engineering, Hunan University, China

²Alibaba Group, China ³School of Design, Hunan University, China



Bottom-up vs. Top-down.



Bottom-up manners:

- deduce semantic concepts from pixel features.
- under the guide of pre-defined rules or visual cues.
- the result usually coarse.

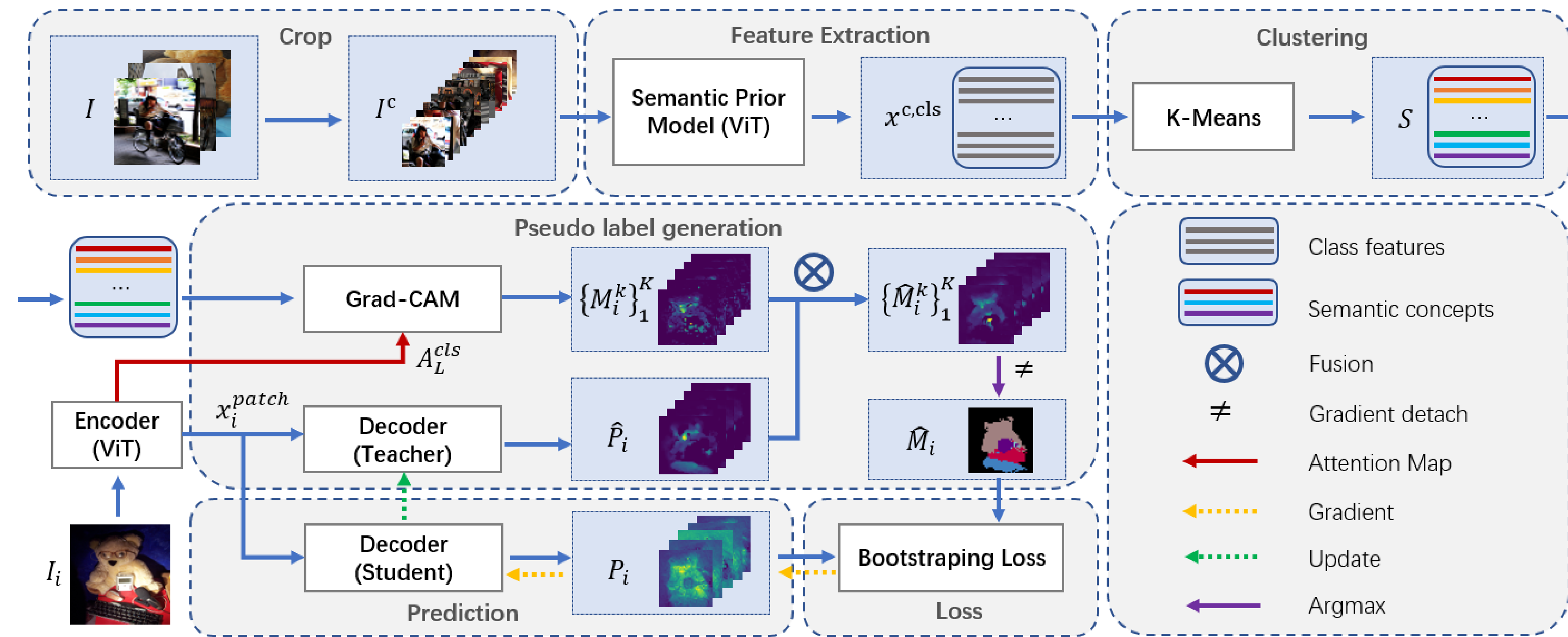
Top-down manner:

- induce semantic concepts from ImageNet (SSL).
- maps to pixel-level features.
- Robust to complicated scenarios and object appearance variations.

Contributions

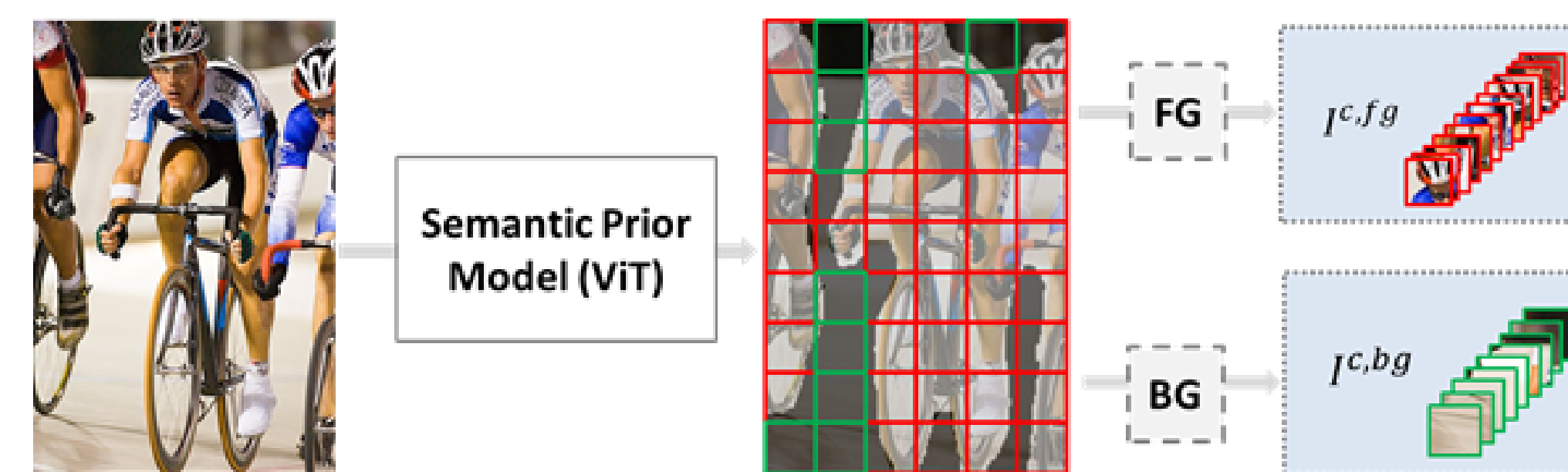
- We propose the first top-down framework for unsupervised semantic segmentation.
- We transfers the high-level semantic features obtained from SSL into low-level pixel-wise features to produce high-quality fine-grained segmentation results.
- This is the first work to apply unsupervised algorithms on real complex datasets (with a large number of classes).

Pipeline



Semantic concepts discovery.

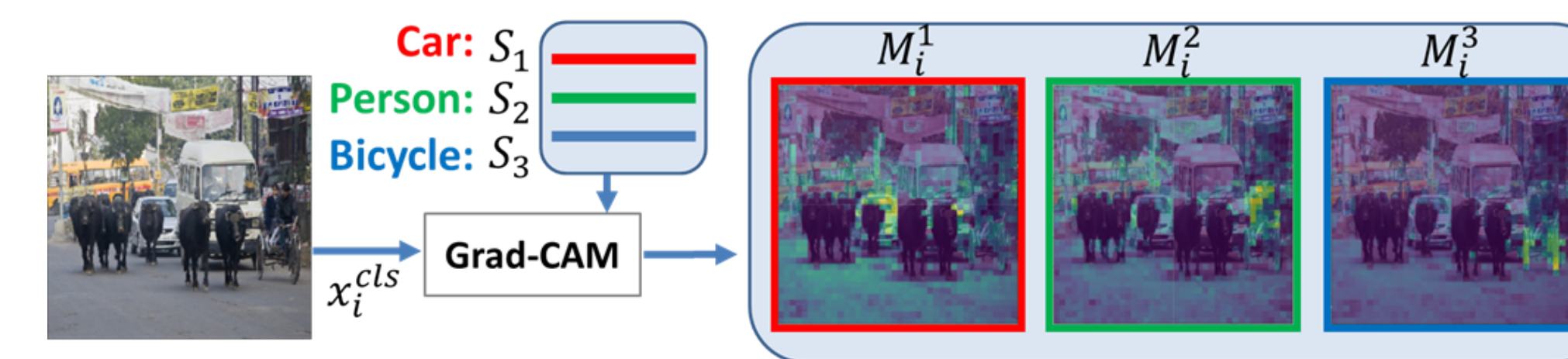
- crop each $I_i \in I$ with different sizes of sliding windows to form I^c .
- use attention map A_L^{cls} in class-token feature x_i^{cls} as foreground prior when foreground (things) and background (stuff) are required to segmented separately.



- obtain K target semantic concepts S based on class-token feature of $x^{c,cls}$ by K-Means, K set as the number of classes w.r.t desired granularity levels.

Training.

- generate gradient on A_L^{cls} w.r.t S_k by maximizing the cosine similarity: $\min(1 - \frac{x_i^{cls} S_k^T}{\sqrt{d}})$.



- aggregate the output of Grad-CAM $\{M_i^k\}_1^K$ and teacher network \hat{P}_i to refine pseudo label $\{\hat{M}_i^k\}_1^K$.
- $M_i^{bg} = \text{RELU}(T^{bg} - \max_{k \in [0, K]} M_i^k)$ is background probability if only foreground need to segmented:
- bootstrapping loss $\mathcal{L} = \mathcal{L}_{peer} + \omega_1 \cdot \mathcal{L}_{div} + \omega_2 \cdot \mathcal{L}_{unc}$.
 - peer loss: $\mathcal{L}_{peer} = \text{CE}(P, \hat{M}) - \alpha \cdot \text{CE}(P, \hat{M}')$
 - diversity loss: $\mathcal{L}_{div} = 1 + \frac{1}{K^2} \sum \frac{C \cdot C^T}{\sqrt{d}}$.
 - uncertainty loss: $\mathcal{L}_{unc} = 1 - \frac{1}{hw} \sum (h', w') (p' - p'')$.

Quantitative Results

Table 1: Results on four benchmarks. * indicates the results are evaluated on the “curated” samples. † denotes PiCIE trained without auxiliary clustering.

| Dataset | Method | mIoU | Acc. |
|----------------|--------------|--------------|--------------|
| COCO-Stuff-27* | IIC | 6.71 | 21.79 |
| | PiCIE† | 13.84 | 48.09 |
| | PiCIE | 14.36 | 49.99 |
| | TransFGU | 17.47 | 52.66 |
| COCO-Stuff-27 | IIC | 2.36 | 21.02 |
| | PiCIE | 11.88 | 37.20 |
| | TransFGU | 16.19 | 44.52 |
| COCO-Stuff-171 | IIC | 0.64 | 8.67 |
| | PiCIE | 4.56 | 24.66 |
| | TransFGU | 11.93 | 34.32 |
| COCO-80 | MaskContrast | 3.73 | 8.81 |
| | TransFGU | 12.69 | 64.31 |
| Cityscapes | IIC | 6.35 | 47.88 |
| | PiCIE | 12.31 | 65.50 |
| | TransFGU | 16.83 | 77.92 |
| Pascal-VOC | MaskContrast | 35.00 | 79.84 |
| | TransFGU | 37.15 | 83.59 |
| LIP-5 | TransFGU | 25.16 | 65.76 |
| LIP-16 | TransFGU | 15.49 | 60.08 |
| LIP-19 | TransFGU | 12.24 | 42.52 |

Quantitative Results

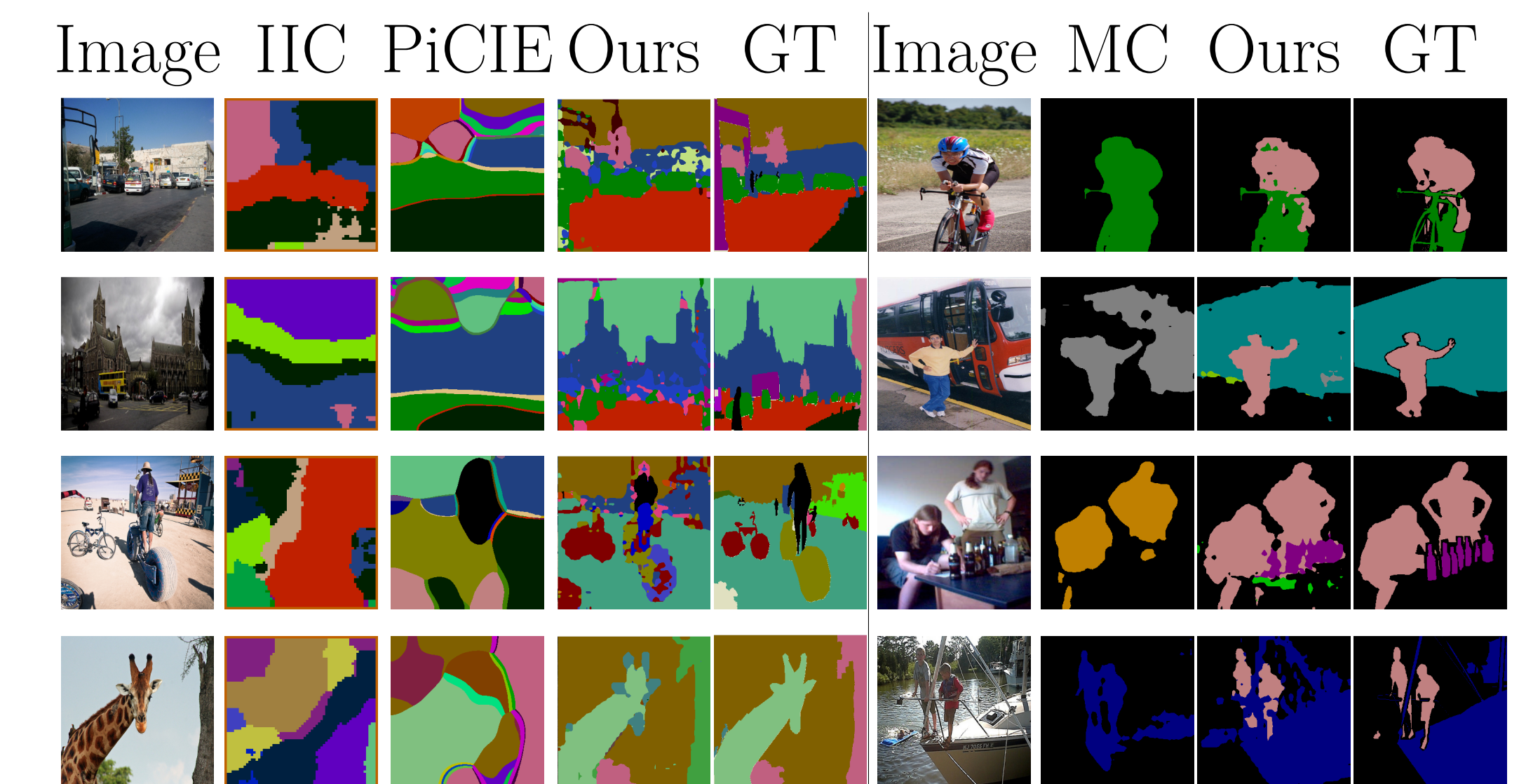


Figure 1: Qualitative comparison on COCO-STuff-171 (left) and Pascal-VOC (right).