# FLOSS: Free Lunch in Open-vocabulary Semantic Segmentation – Supplementary Material –

Yasser Benigmim[1] Mohammad Fahes[1] Tuan-Hung Vu[1,2] Andrei Bursuc[1,2] Raoul de Charette[1]
[1] Inria [2] Valeo.ai

## A. Additional experiments

In this section, we provide additional experimental results and analyses that complement our main findings. First, we showcase the consistent improvements brought by our method across different models and datasets (Section A.1). We also report the computational overhead analysis including inference time and GPU memory requirements (Section A.2). Then we extend our empirical analysis of class-expert templates to additional datasets and models (Section A.3). We follow with a detailed assessment of our expert identification method's accuracy (Section A.4) and an in-depth explanation of the different unsupervised metrics we investigated (Section A.5). We present a comprehensive study of the relationship between entropy and IoU across different settings (Section A.6). For reproducibility, we provide the complete list of ImageNet templates used across all experiments (Section A.7). Finally, we provide qualitative visual results that demonstrate the effectiveness of our class-expert fusion approach (Section A.8).

### A.1. Performance improvements across models

Fig. 7 provides a graphical view across datasets of the improvement in mIoU when plugging FLOSS on existing OVSS models. This demonstrates the general applicability of our approach, which can be seamlessly integrated with existing models without requiring any additional training.

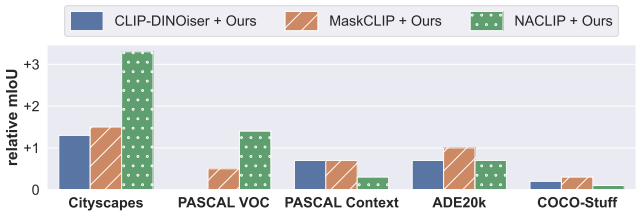



Figure 7. **Performance boost with FLOSS**. We report the mIoU difference when using FLOSS on top of an existing OVSS model. Our training-free method consistently improves all OVSS models across different datasets through class-expert template identification.

### A.2. Inference time and GPU memory

We provide the results of the inference time and GPU memory of FLOSS when it is plugged to existing open-vocabulary models across different datasets. As shown in Tab. 8, the computational overhead varies significantly based on the number of classes in each dataset. For datasets with a smaller number of classes such as CS [1] (19 classes) and VOC20 [2] (20 classes), FLOSS incurs minimal computational overhead, with modest increases in both inference time and GPU memory usage. However, for datasets with a larger number of classes like PC59 [4] (59 classes) and ADE [7] (150 classes), the overhead becomes more substantial, with notable increases in both inference time and peak GPU memory consumption, reflecting the scaling nature of cosine similarity computations across class-experts. It is worth noting that NACLIP [3] exhibits particularly high inference times on CS for both the baseline and FLOSS, which stems from the preprocessing pipeline in the original NACLIP implementation that resizes CS images to $1120\times560$ resolution and employs sliding window inference with $224\times224$ crops and a stride of 112, requiring 36 overlapping patch inferences to process each complete image.

| Method | CS | | VOC20 | | PC59 | | ADE | |
|---|---|---|---|---|---|---|---|---|
| | Inf. time | GPU mem. | Inf. time | GPU mem. | Inf. time | GPU mem. | Inf. time | GPU mem. |
| CLIP-DINOiser | 31 | 1203/1274 | 23 | 1184/1263 | 24 | 1204/1397 | 23 | 1252/1661 |
| + FLOSS | 50 | 1204/1940 | 35 | 1184/2034 | 83 | 1211/7225 | 339 | 1295/19249 |
| MaskCLIP | 30 | 1195/1259 | 23 | 1175/1251 | 23 | 1195/1436 | 22 | 1242/1652 |
| + FLOSS | 59 | 1196/1899 | 43 | 1175/1986 | 108 | 1202/7142 | 421 | 1286/19240 |
| NACLIP | 100 | 449/571 | 18 | 365/408 | 23 | 373/458 | 22 | 422/739 |
| + FLOSS | 249 | 449/571 | 41 | 366/1232 | 91 | 377/1092 | 274 | 444/4854 |

Table 8. **Inference time and GPU memory analysis.** We report inference time (ms) and GPU memory usage (current/peak in MB) for different OVSS models with and without FLOSS across four datasets.

### A.3. Empirical observations on other settings

The foundation of our work is the observation that there exist class-wise expert templates. We extend here our initial analysis to 4 datasets, for both CLIP-DINOiser [6] in Fig. 9 and NACLIP in Fig. 10, showing that our initial observations hold in these settings. Interestingly, we note that for

PASCAL CONTEXT 59, "template-averaging" performs already very well and most individual template underperform it. Arguably, this might be related to the proximity of PASCAL VOC 20 with ImageNet for which the templates of CLIP were engineered [5].

### A.4. Quality of experts

We report the quality of our estimated top-4 experts using entropy as unsupervised metric for both CLIP-DINOiser in Fig. 11 and NACLIP in Fig. 12. In details, for each class $k \in [1, K]$, the quality is computed as the normalized intersection between the set of estimated experts $\hat{\mathcal{E}}_k$ and the true set of experts $\mathcal{E}_k$, as detailed in Eq. (8). For each dataset, we also report the average accuracy over all classes, *i.e.*, $\frac{1}{K}\sum_k \hat{\rho}_k$, shown as a horizontal line.

While we observe a high variability of the quality across classes, our average top-4 quality is around 50%, meaning that half of our estimated expert templates are usually actual class-experts. We also note that the identification of experts on PASCAL VOC 20 is less accurate, which may stem from the rare class-experts on said dataset and the high performance of the "template-averaging" model.

### A.5. Unsupervised metrics for expert identification

In Sec. 5.2, we compared entropy to three other metrics for expert identification. We provide here details about these metrics. Given a single-template classifier $\boldsymbol{W}(\mathcal{T}_m)$ making predictions on unlabeled images, (i) *Avg. Probability* computes the average probability of all pixels predicted as class $k$ by $\boldsymbol{W}(\mathcal{T}_m)$. Higher values indicate better performance in this case. It writes:

$$\text{Avg.Prob}_k(\mathcal{T}_m) = \frac{1}{C_{m,k}} \sum_{i=1}^{C_{m,k}} \text{softmax}(\mathbf{q}_i)_k \qquad (1)$$

(ii) *MaNo* is based on low density separation (LDS) assumption. The LDS principle suggests a direct correlation between the magnitude of logit values and a model's generalization capabilities, so the higher the MaNo score, the better it is. For each class $k$, we compute the average $L_p$ norm of the probability vectors for pixels predicted as that class:

$$\text{MaNo}_k(\mathcal{T}_m) = \left( \frac{1}{C_{m,k}K} \sum_{i=1}^{C_{m,k}} \sum_{k=1}^{K} |\text{softrun}(\mathbf{q}_i)_k|^p \right)^{\frac{1}{p}}, \qquad (2)$$

where softrun is the normalization function used to avoid error accumulation, and we use $p = 4$ in practice, as done in the original paper.

(iii) *Inter-to-Intra (ITI)* is a class ratio that quantifies feature separability by examining the relationship between inter-class separation and intra-class compactness for each class $k$. The intra-class compactness measure $D_{\text{intra,k}}(\mathcal{T}_m)$ calculates the average squared Euclidean distance between the global class feature centroid $\bar{f}_k$ and the class-specific features $\bar{f}_k^i$ (denoting the average feature for class $k$ in image $i$) across all $N$ training images, capturing feature variability within the class.

Conversely, the inter-class separability measure $D_{\text{inter,k}}(\mathcal{T}_m)$ computes the average squared distance between the target class centroid $\bar{f}_k$ and all other class centroids $\bar{f}_i$, measuring how distinctly a class is represented in feature space. The ITI ratio $\text{ITI}_k(\mathcal{T}_m)$ then provides a normalized metric where higher values indicate well-separated and compact feature representations, with $\bar{f}_k$ representing the class $k$'s feature centroid across the entire dataset, $N$ representing the training dataset size, and $K$ signifying the total number of semantic classes.

$D_{\text{intra,k}}(\mathcal{T}_m)$, $D_{\text{inter,k}}(\mathcal{T}_m)$ and $\text{ITI}_k(\mathcal{T}_m)$ write as:

$$\begin{aligned} D_{\text{intra,k}}(\mathcal{T}_m) &= \frac{1}{N}\sum_{i=1}^{N} \|\bar{f}_k - \bar{f}_k^i\|_2^2 \\ D_{\text{inter,k}}(\mathcal{T}_m) &= \frac{1}{N-1}\sum_{i=1}^{N-1} \|\bar{f}_k - \bar{f}_i\|_2^2 \\ \text{ITI}_k(\mathcal{T}_m) &= \frac{D_{\text{inter,k}}(\mathcal{T}_m)}{D_{\text{intra,k}}(\mathcal{T}_m)} \end{aligned} \qquad (3)$$

### A.6. Studying the relation of entropy and IoU

We analyze the relationship between entropy and IoU across different models and datasets. Figs. 13 and 14 show this relationship for CLIP-DINOiser, while Figs. 15 and 16 present the same analysis for NACLIP. For each class, we plot the entropy and IoU of individual templates (colored dots) against the performance of the original CLIP "template-averaging" (dotted line). The number of valid templates (those predicting at least one pixel for the class) is indicated in parentheses next to each class name.

### A.7. ImageNet templates

For reproducibility, we provide the complete list of 80 ImageNet templates used consistently across all our experiments. These templates are applied in the exact order listed below, ensuring consistency across all reported results:

0. 'a bad photo of a {}.'
1. 'a photo of many {}.'
2. 'a sculpture of a {}.'
3. 'a photo of the hard to see {}.'
4. 'a low resolution photo of the {}.'
5. 'a rendering of a {}.'
6. 'graffiti of a {}.'
7. 'a bad photo of the {}.'
8. 'a cropped photo of the {}.'
9. 'a tattoo of a {}.'
10. 'the embroidered {}.'
11. 'a photo of a hard to see {}.'
12. 'a bright photo of a {}.'
13. 'a photo of a clean {}.'

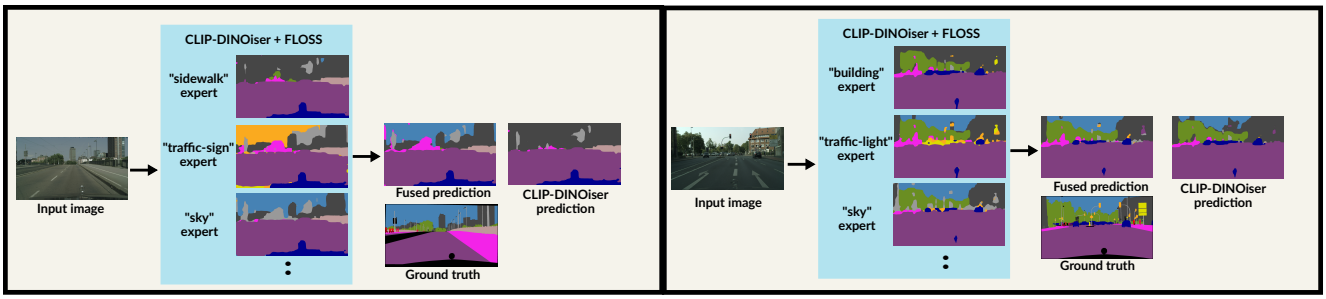
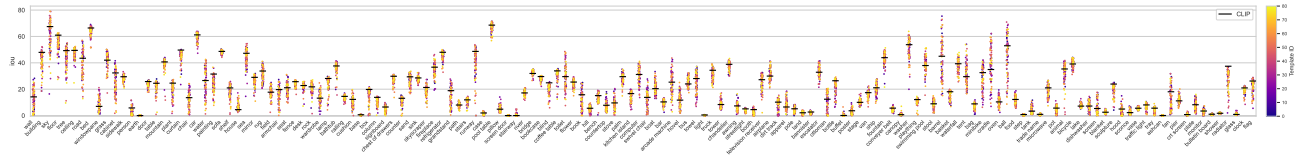



Figure 8. **Qualitative results on Cityscapes.** The figure shows qualitative results displaying predictions from individual class-experts when integrating FLOSS with CLIP-DINOiser on Cityscapes images. The visualization demonstrates how the fused prediction effectively combines the predictions made by class-experts within their respective domains of expertise, outperforming the baseline CLIP-DINOiser prediction.

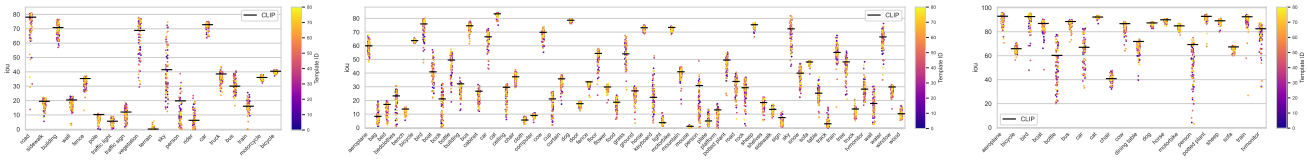
14. 'a photo of a dirty {}.'
15. 'a dark photo of the {}.'
16. 'a drawing of a {}.'
17. 'a photo of my {}.'
18. 'the plastic {}.'
19. 'a photo of the cool {}.'
20. 'a close-up photo of a {}.'
21. 'a black and white photo of the {}.'
22. 'a painting of the {}.'
23. 'a painting of a {}.'
24. 'a pixelated photo of the {}.'
25. 'a sculpture of the {}.'
26. 'a bright photo of the {}.'
27. 'a cropped photo of a {}.'
28. 'a plastic {}.'
29. 'a photo of the dirty {}.'
30. 'a jpeg corrupted photo of a {}.'
31. 'a blurry photo of the {}.'
32. 'a photo of the {}.'
33. 'a good photo of the {}.'
34. 'a rendering of the {}.'
35. 'a {} in a video game.'
36. 'a photo of one {}.'
37. 'a doodle of a {}.'
38. 'a close-up photo of the {}.'
39. 'a photo of a {}.'
40. 'the origami {}.'
41. 'the {} in a video game.'
42. 'a sketch of a {}.'
43. 'a doodle of the {}.'
44. 'a origami {}.'
45. 'a low resolution photo of a {}.'
46. 'the toy {}.'
47. 'a rendition of the {}.'
48. 'a photo of the clean {}.'
49. 'a photo of a large {}.'
50. 'a rendition of a {}.'
51. 'a photo of a nice {}.'
52. 'a photo of a weird {}.'
53. 'a blurry photo of a {}.'
54. 'a cartoon {}.'
55. 'art of a {}.'
56. 'a sketch of the {}.'
57. 'a embroidered {}.'
58. 'a pixelated photo of a {}.'
59. 'itap of the {}.'
60. 'a jpeg corrupted photo of the {}.'
61. 'a good photo of a {}.'
62. 'a plushie {}.'
63. 'a photo of the nice {}.'
64. 'a photo of the small {}.'
65. 'a photo of the weird {}.'
66. 'the cartoon {}.'
67. 'art of the {}.'
68. 'a drawing of the {}.'
69. 'a photo of the large {}.'
70. 'a black and white photo of a {}.'
71. 'the plushie {}.'
72. 'a dark photo of a {}.'
73. 'itap of a {}.'
74. 'graffiti of the {}.'
75. 'a toy {}.'
76. 'itap of my {}.'
77. 'a photo of a cool {}.'
78. 'a photo of a small {}.'
79. 'a tattoo of the {}.'

## A.8. Qualitative results

In this section, we provide qualitative visual results to complement our quantitative analysis. Fig. 8 presents example segmentation results demonstrating the effectiveness of our class-expert fusion approach when applied to CLIP-DINOiser on Cityscapes images. The visualizations clearly illustrate how FLOSS effectively combines predictions from different class-experts, each contributing their specialized knowledge for their respective classes of expertise. Notably, we observe the significant impact of class-specific experts, such as the sky-expert, on the final fused prediction, showcasing substantial improvements in segmentation quality compared to the baseline CLIP-DINOiser predictions.

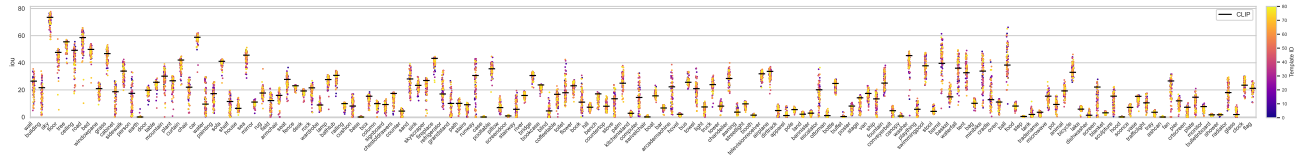


(a) ADE20K

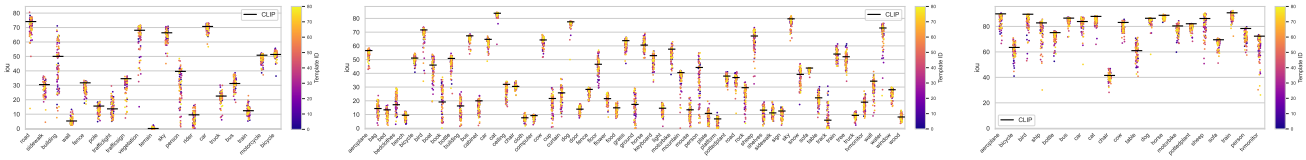


(b) Cityscapes (c) PASCAL CONTEXT 59 (d) PASCAL VOC 20

Figure 9. **Individual template vs average templates (CLIP-DINOiser).**

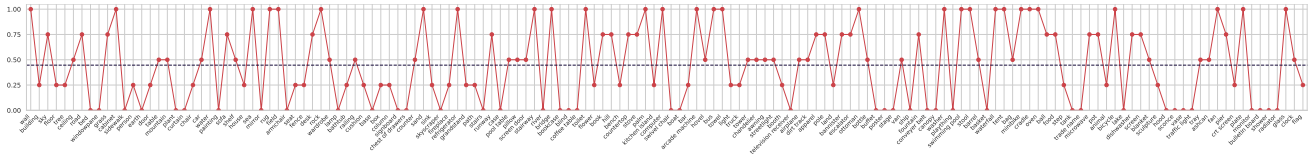


(a) ADE20K

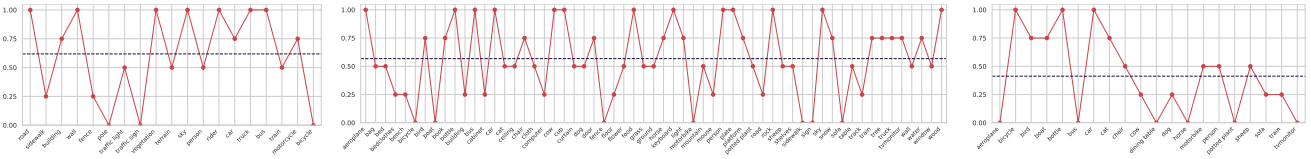


(b) Cityscapes (c) PASCAL CONTEXT 59 (d) PASCAL VOC 20

Figure 10. **Individual template vs average templates (NACLIP).**

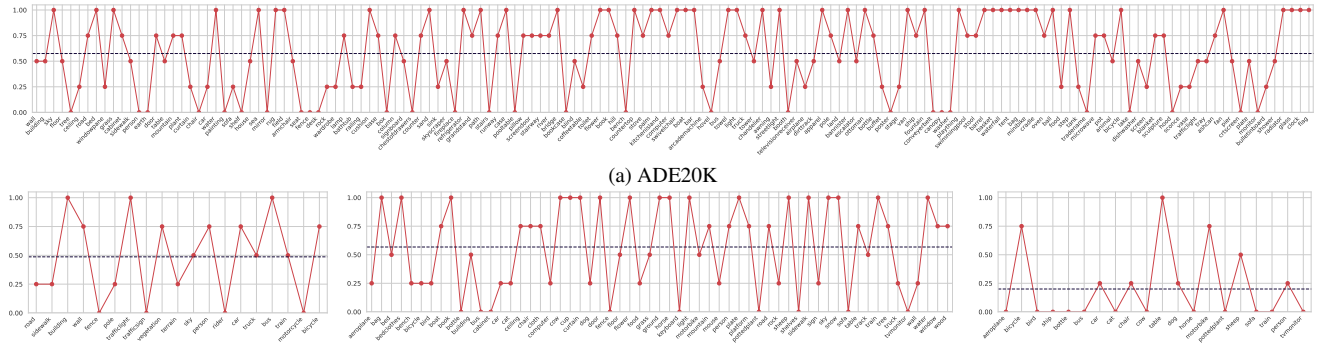


(a) ADE20K



(b) Cityscapes (c) PASCAL CONTEXT 59 (d) PASCAL VOC 20

Figure 11. **Quality of the estimated templates (CLIP-DINOiser).** For each class $k$, we report the accuracy of the estimated top-4 experts $\hat{\mathcal{E}}_k$ as the normalized intersection with the set of true experts from that class, i.e., $\mathcal{E}_k$. The dash line indicates the average across classes.



(a) ADE20K

(b) Cityscapes (c) PASCAL CONTEXT 59 (d) PASCAL VOC 20

Figure 12. **Quality of the estimated templates (NACLIP).** For each class $k$, we report the accuracy of the estimated top-4 experts $\hat{\mathcal{E}}_k$ as the normalized intersection with the set of true experts from that class, i.e., $\mathcal{E}_k$. The dash line indicates the average across classes.

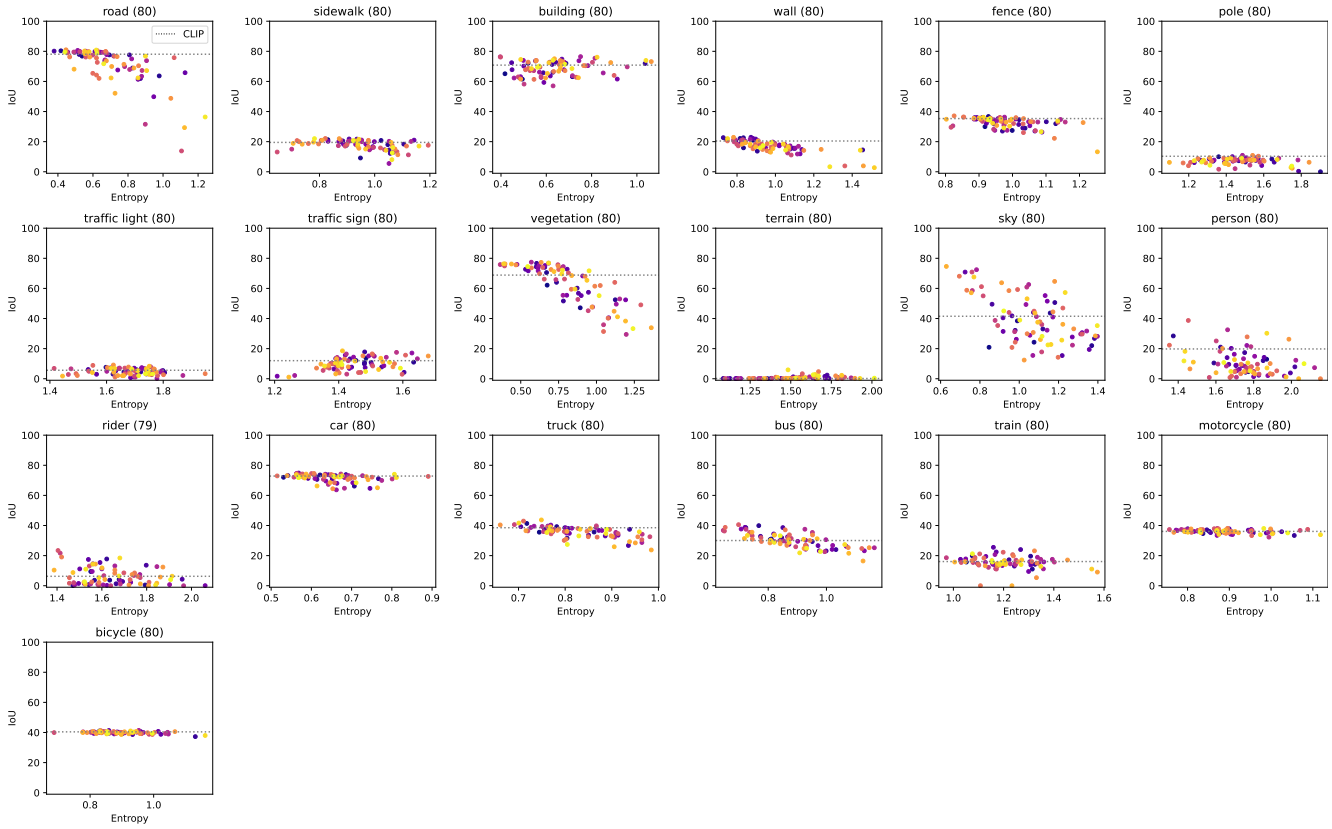



Figure 13. **Entropy and IoU per template (CLIP-DINOiser, Cityscapes).** Each plot reports, for a given class, the entropy and IoU of all individual template (colored dots) as well as the original CLIP averaged-templates performance (dotted line). Note that we consider templates valid only if they predict more than 0 pixels for that class, and therefore report the number of valid templates in parenthesis next to the class name.

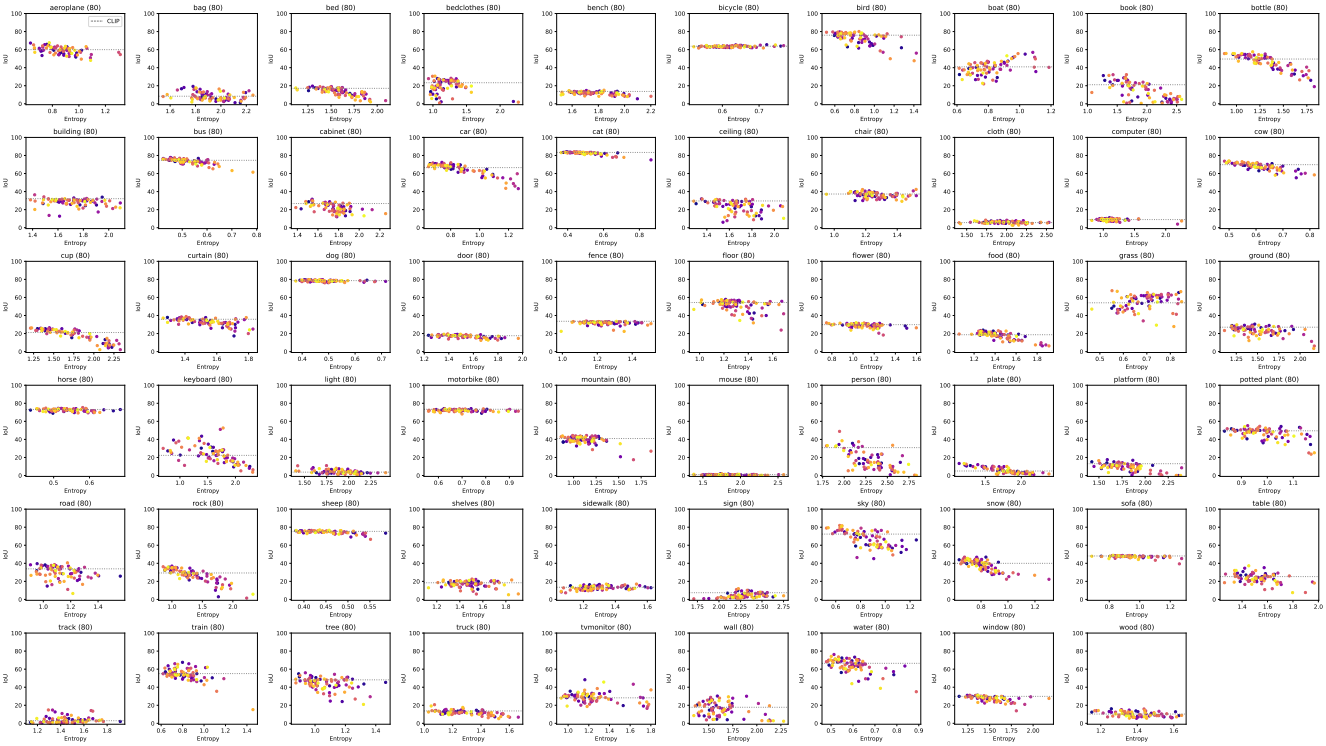



Figure 14. **Entropy and IoU per template (CLIP-DINOiser, PASCAL CONTEXT 59).** Each plot reports, for a given class, the entropy and IoU of all individual template (colored dots) as well as the original CLIP averaged-templates performance (dotted line). Note that we consider templates valid only if they predict more than 0 pixels for that class, and therefore report the number of valid templates in parenthesis next to the class name.

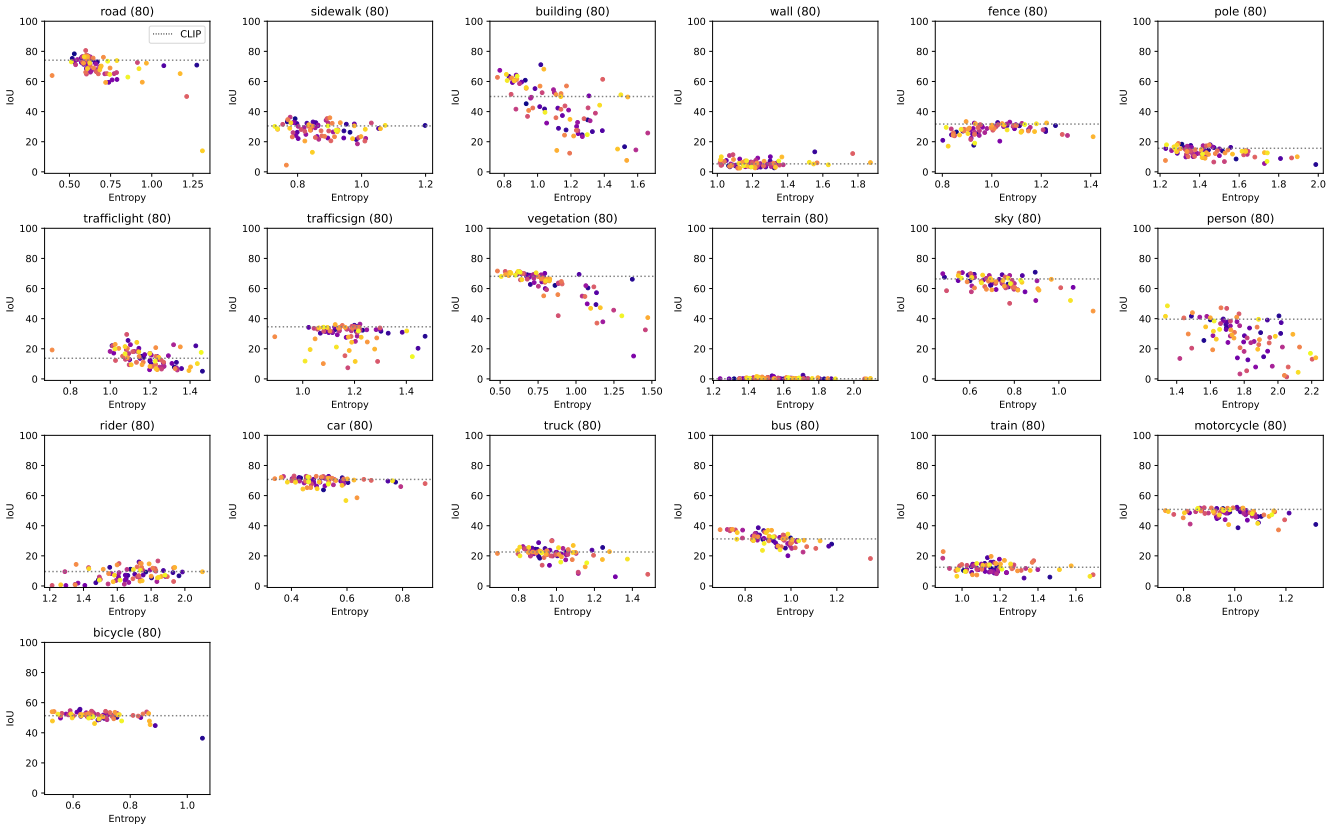



Figure 15. **Entropy and IoU per template (NACLIP, Cityscapes).** Each plot reports, for a given class, the entropy and IoU of all individual template (colored dots) as well as the original CLIP averaged-templates performance (dotted line). Note that we consider templates valid only if they predict more than 0 pixels for that class, and therefore report the number of valid templates in parenthesis next to the class name.

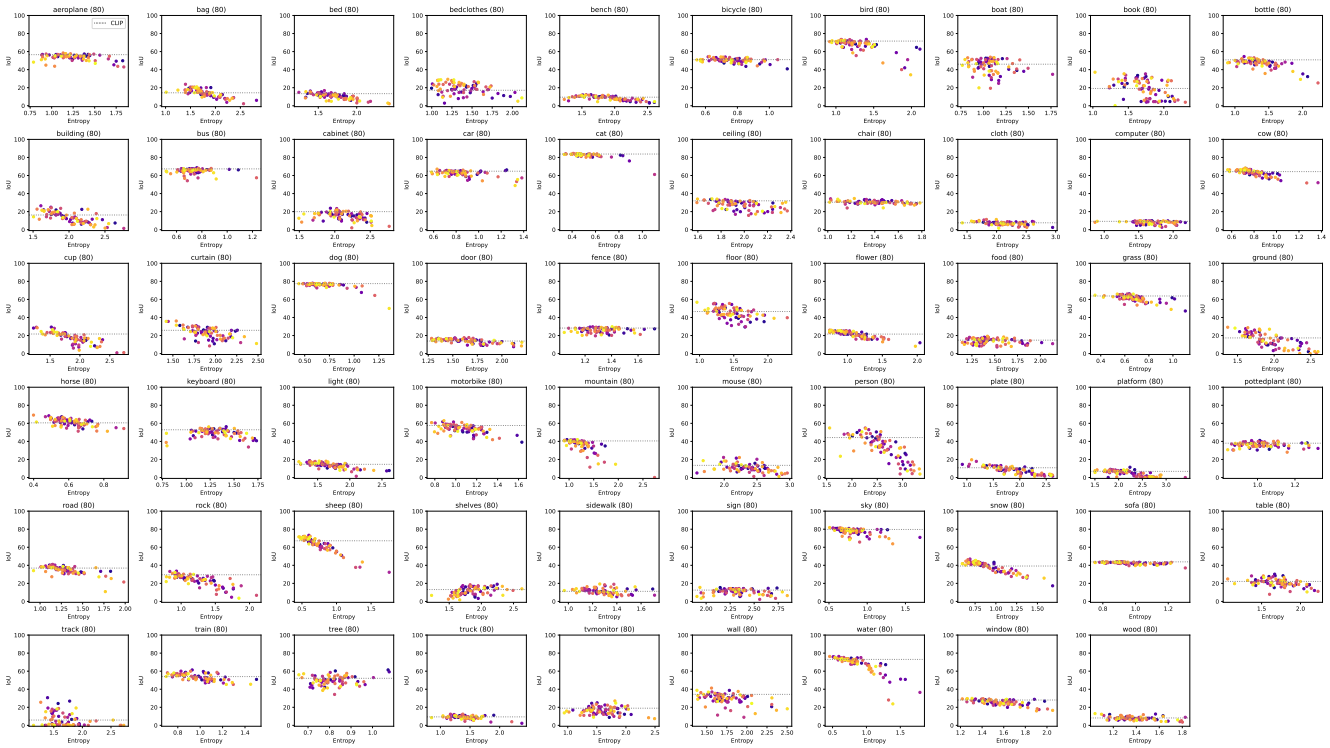



Figure 16. **Entropy and IoU per template (NACLIP, PASCAL CONTEXT 59).** Each plot reports, for a given class, the entropy and IoU of all individual template (colored dots) as well as the original CLIP averaged-templates performance (dotted line). Note that we consider templates valid only if they predict more than 0 pixels for that class, and therefore report the number of valid templates in parenthesis next to the class name.

## 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] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, 2012. 1

[3] Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *WACV*, 2025. 1

[4] 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 *CVPR*, 2014. 1

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[6] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In *ECCV*, 2024. 1

[7] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019. 1