A. Appendix

In this document, we include supplementary materials for PPL. Firstly, we provide methodological details on PPL (Sec. B) and Pseudo-code of PPL (Sec. C). Furthermore, we provide the additional qualitative results for dense prediction tasks (Sec. D).

B. Uncertainty on PPL

In this section, we provide how to measure uncertainty of text representations with visual context. In addition, we report the cause of uncertainty.

B.1. Uncertainty estimation

In PPL, the whole distribution of each class of given input image is estimated as a Mixture of Gaussian (MoG) with K-attribute prompts. To compute uncertainty of images, we describe the computation of mean and variance of each class of image. The PDF of a MoG of each class is represented by the average PDF of its attribute distributions of image given.

$$f_{\boldsymbol{w}_c}(z) = \frac{1}{K}\sum f_{w_c^k}(z). \tag{15}$$

Then, the mean of the MoG is formulated follow as:

$$\begin{aligned}\mu(\boldsymbol{w}_c) &= \int z f_{\boldsymbol{w}_c}(z)dz \\ &= \frac{1}{K}\sum \int z f_{w_c^k}(z)dz \\ &= \frac{1}{K}\sum \mu_c^k.\end{aligned} \tag{16}$$

The standard deviation $\sigma(\boldsymbol{w}_c)^2$ is derived as follow:

$$\begin{aligned}\sigma(\boldsymbol{w}_c)^2 &= \int z^2 f_{\boldsymbol{w}_c}(z)dz - \mu(\boldsymbol{w}_c)^2 \\ &= \frac{1}{K}\sum \int z^2 f_{w_c^k}(z)dz - \mu(\boldsymbol{w}_c)^2 \\ &= \frac{1}{K}\sum((\mu_c^k)^2 + (\sigma_c^k)^2) - (\frac{1}{K}\sum \mu_c^k)^2.\end{aligned} \tag{17}$$

We define the geometric mean of the variance $\sigma(\boldsymbol{w}_c)$ of each class c is used as uncertainty of its class. Finally, we formulate the total uncertainty of image is derived as follow:

$$\bar{\sigma}(\boldsymbol{w}_{1:C})_{\mathrm{G}} = \prod(\sigma(\boldsymbol{w}_c))^{1/c}. \tag{18}$$

B.2. Uncertainty Analysis

To better understand uncertainty, we provide a brief analysis of the causes of uncertainty. Although it is impossible to estimate all causes, we studied the correlation between the number of classes in an image and uncertainty. As shown in Fig. 1, the number of classes in image have positive relationship with uncertainty. Based on this results, the predicted uncertainty can be used to remove ambiguous visual-context and leverage the useful context in image.

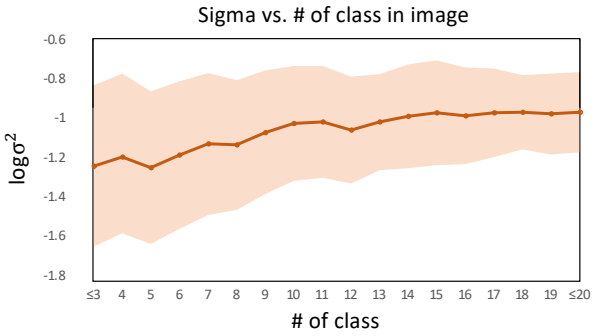


Figure 1. We measure the uncertainty of images on ADE20k [2] dataset according to the number of classes included in the image.

C. Algorithm

Algorithm 1: Pseudo-code of PPL Training.

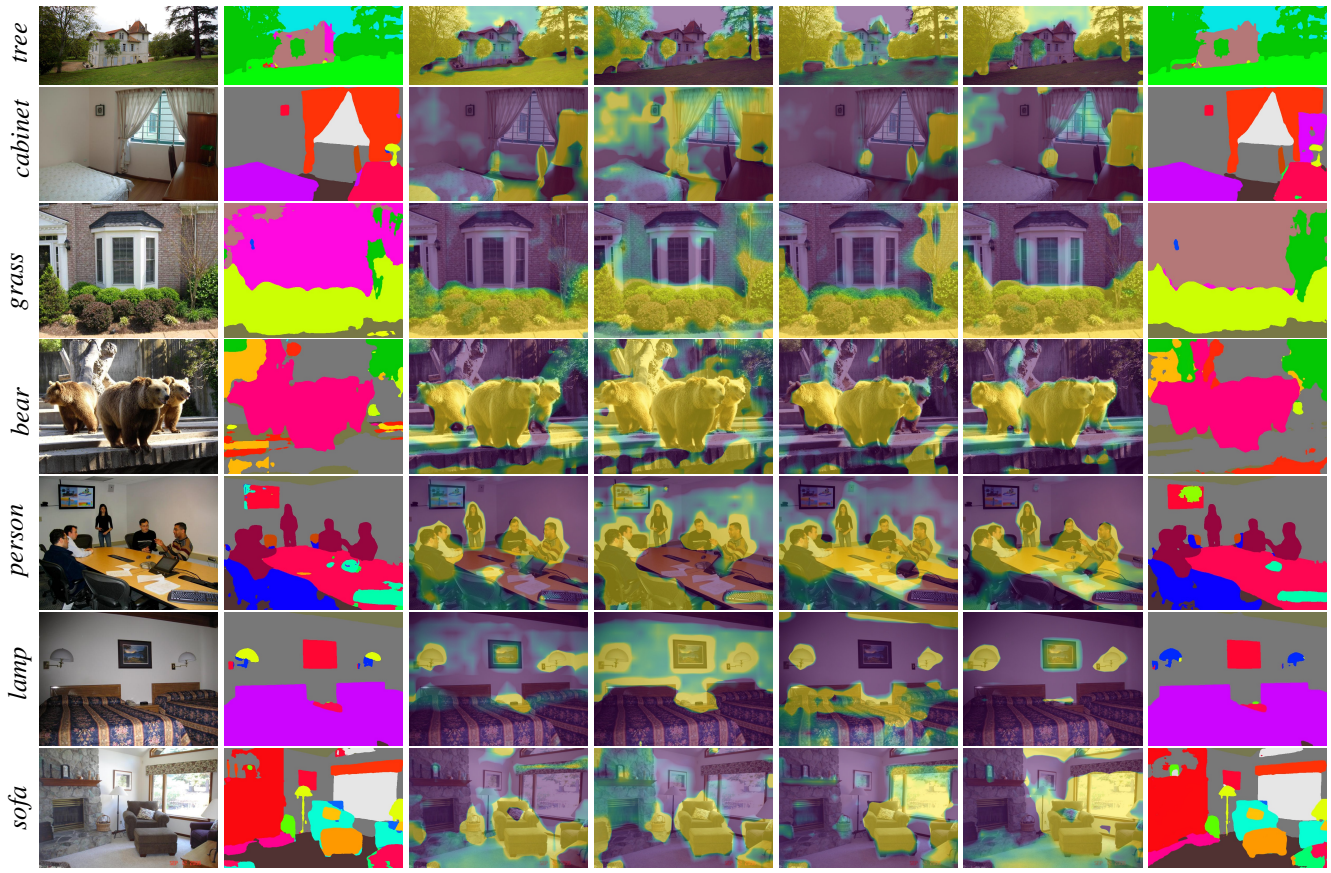
Require: The pre-trained CLIP text encoder $\mathcal{G}$,image encoder $\mathcal{F}$, and visual-context probabilistic decoder $\mathcal{M}$
Require: Class descriptions $\mathbf{t}_{1:C}(\cdot)$ and randomly initialized prompts set $\mathbf{P} = [\mathbf{p}^1, ...\mathbf{p}^N]$

for *t to T do* **do**
- Draw a mini-batch ($\mathbf{x}$, y).
- Compute $v = \mathcal{F}(\mathbf{x})$ and $\boldsymbol{w}_{1:C} = \mathcal{G}(\mathbf{t}_{1:C}(\mathbf{P}))$
- Let $\boldsymbol{w}_c = [w_c^1, ..., w_c^K]$
- Compute $\mathcal{L}_{div}$ according to Eq. (5)
- Compute $\boldsymbol{\sigma}_c^k = \mathcal{M}(w_c^k, v)$
- Compute $p(z|\boldsymbol{w}_c)$ according to Eq. (8)
- Compute $\mu(\boldsymbol{w}_c)$ and $\sigma(\boldsymbol{w}_c)$ according to Eq. (16), (17)
- Sample text embedding $\boldsymbol{z}_c$ from $p(z|\boldsymbol{w}_c)$
- Compute uncertainty $\log\sigma^2$ according to Eq. (18)
- Compute $\mathcal{L}_{pixel}$ according to Eq. (10)
- Compute $\mathcal{L}_{prob}$ according to Eq. (11)
- Compute $\mathcal{L}_{KL}$ according to Eq. (12)
- Compute total loss $\mathcal{L}$ according to Eq. (13)
- Update $\mathbf{P}$ and $\mathcal{M}$ by gradient descent

D. Additional Visualization

In this section, we provide more visualization results of our method and comparison our method with DenseCLIP [1]. As shown if Fig. 2, we showed that each similarity map with different visual context represent

the target class object as different ways. Specifically, combining different similarity maps remove undesirable prediction and improves performance. We report the qualitative results with given score maps compared to DenseCLIP [1].



(a) Image (b) DenseCLIP [1] (c) `Samples 1` (d) `Samples 2` (e) `Samples 3` (f) `Samples 4` (g) PPL(ours)

Figure 2. **Visualization of activation maps and segmentation results**. We visualize the activation maps (c), (d) (e), and (f) of sampled representation of different classes indicated on the left side, with $K = 3$ on the ADE20k dataset [2]. We report qualitative results of segmentation of both (b) DenseCLIP [1], (g) PPL(ours).

References

[1] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, 2022. 1, 2, 3

[2] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 127(3):302–321, 2019. 1, 3