VR-CLIP: Visual Refinement of CLIP for Zero-Shot Semantic Segmentation

Supplementary Material

6. Implementation Details

For fair comparison, we follow the same experimental settings as prior works [44, 50], including dataset splits, text templates.

Split of unseen classes. The specific unseen classes used for evaluation on each dataset are listed in Tab. 7. These splits follow the widely adopted ZS3 protocol and ensure consistent comparison across different methods.

Table 7. Unseen class names for each dataset.

Dataset	Unseen Classes
VOC	*pottedplant, sheep, sofa, train, tvmonitor*
COCO	*cow, giraffe, suitcase, frisbee, skateboard, carrot, scissors, cardboard, clouds, grass, playingfield, river, road, tree, wall concrete*
Context	*cow, motorbike, sofa, cat, boat, fence, bird, tvmonitor, keyboard, aeroplane*

Text templates. We follow the standard CLIP prompting strategy to construct text embeddings. For PASCAL VOC 2012, we use a single prompt template due to its small and generic label space:

```
A photo of a {}.
```

For large-scale datasets with more diverse category names such as COCO-Stuff 164K and PASCAL-Context, we adopt 15 augmented templates to obtain more robust class embeddings. The templates are listed as follows:

```
A photo of a {}.
A photo of a small {}.
A photo of a medium {}.
A photo of a large {}.
This is a photo of a {}.
This is a photo of a small {}.
This is a photo of a medium {}.
This is a photo of a large {}.
A {} in the scene.
A photo of a {} in the scene.
There is a {} in the scene.
There is the {} in the scene.
This is a {} in the scene.
This is the {} in the scene.
This is one {} in the scene.
```

For each class, the textual embeddings produced by all templates are averaged to form the final class representation used during training and inference.

7. Comparisons with other methods

As discussed in Sec.1, several prior works [6, 22, 27, 35] have explored adapting CLIP's visual features for dense prediction tasks. To further support our analysis, we conducted preliminary experiments to examine whether these adaptation strategies, although effective in fully supervised dense prediction, can transfer to the ZS3 setting.

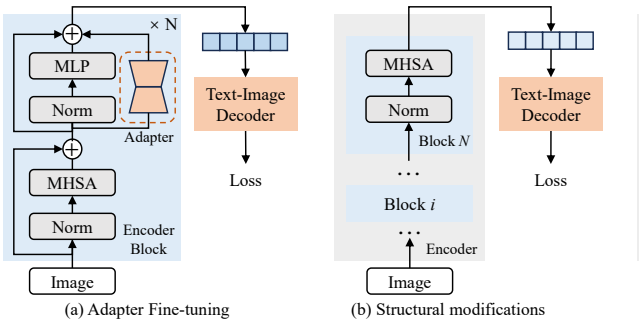


Figure 6. Methods for optimizing CLIP's visual features for dense prediction tasks. (a) Add parallel Adapter to the MLP layer of the VIT Block to adapt features for downstream tasks. (b) We employ the previously method [22] to eliminate residual connections and remove the multi-layer MLP from the final layer to reduce noise.

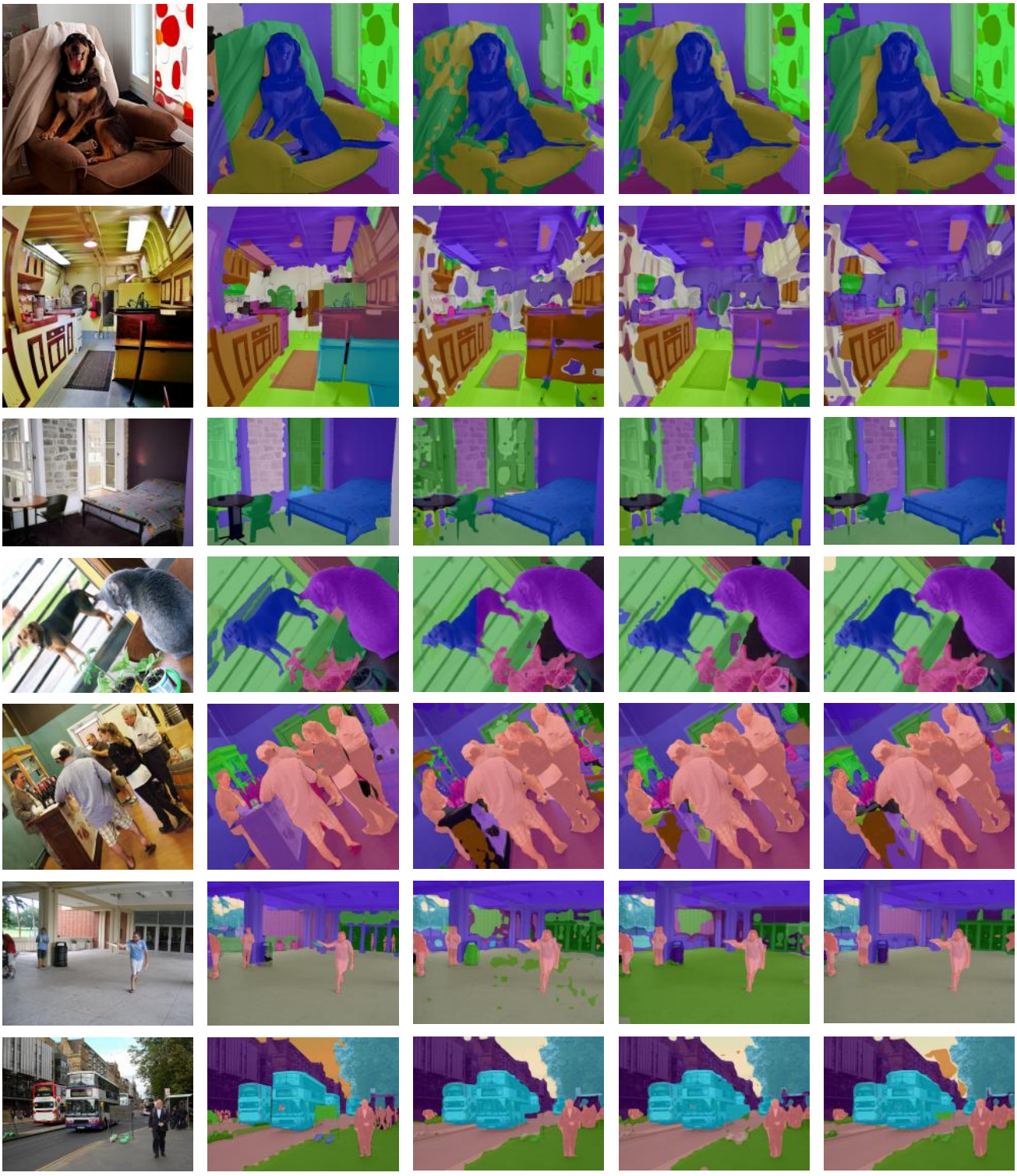
Table 8. Comparisons with different methods (Fig. 6).

Methods	mIoU^S↑	mIoU^U↑	hIoU↑
Baseline	93.1	85.4	88.7
+(a) in Shallow layers	93.5	78.9	85.6
+(a) in all layers	93.9	64.1	76.2
+(b)	84.6	72.5	78.1
+(Ours)	**94.2**	**87.8**	**90.9**

Our results in Tab. 8 show that methods commonly used for enhancing CLIP in dense prediction (e.g., adapter-based tuning or structural modifications to the encoder (Fig. 6) do not yield meaningful improvements under ZS3. Instead, they tend to either overfit to seen classes or disrupt the delicate visual–language alignment of CLIP, both of which are critical for zero-shot inference. These observations reinforce the motivation of our work: effective ZS3 requires a refinement strategy that preserves CLIP's global alignment while enhancing pixel-level discriminability.

8. More visualizations

More visualization results under the transductive setting on COCO-Stuff 164K dataset are provided in Fig. 7.



(a) Image (b) Ground truth (c) ZegCLIP (d) SPT-SEG (e) Ours

Figure 7. Qualitative results on COCO-Stuff164K dataset.