*Table 1.* Performance Comparison Across CLIP Architectures and Post-processing Methods. Comparative evaluation of our approach against state-of-the-art methods (Trident and CLIPer) across different CLIP architectures (CLIP-ViT-B/16, CLIP-ViT-L/14, OpenCLIP-ViT-H/14) with and without post-processing. Best results are shown in bold, second-best results are underlined. Note: CorrCLIP is not included in the comparison as its implementation is not publicly available.

| Method | Post-processing | With background | | | Without background | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| | | VOC21 | Context60 | COCO-Obj | VOC20 | City | Context59 | ADE | Stuff | |
| **Without Post-processing** | | | | | | | | | | |
| ***CLIP-ViT-B/16*** | | | | | | | | | | |
| CLIPer(Sun et al., 2024a) | NO | 60.1 | 34.8 | 36.0 | 84.0 | - | 38.5 | 19.8 | 25.3 | - |
| Trident(Shi et al., 2024) | NO | 64.5 | 37.2 | **39.5** | 83.7 | 40.4 | 40.9 | 20.9 | **27.6** | 44.5 |
| **Ours** | NO | **66.1** | **37.7** | 39.4 | **85.4** | **42.8** | **41.2** | **21.0** | 27.3 | **44.8** |
| ***CLIP-ViT-L/14*** | | | | | | | | | | |
| CLIPer(Sun et al., 2024a) | NO | 61.2 | 34.3 | 39.6 | **88.2** | - | 39.8 | 21.8 | 25.8 | - |
| Trident(Shi et al., 2024) | NO | 61.4 | 36.4 | 40.2 | 84.8 | 40.4 | 39.8 | **23.2** | 26.4 | 42.5 |
| **Ours** | NO | **67.0** | **37.3** | **40.2** | 85.5 | **42.7** | **41.0** | 22.9 | **27.1** | **45.5** |
| ***OpenCLIP-ViT-H/14*** | | | | | | | | | | |
| CLIPer(Sun et al., 2024a) | NO | 58.0 | 34.1 | 39.2 | 85.8 | - | 36.9 | 22.1 | 25.2 | - |
| Trident(Shi et al., 2024) | NO | 68.6 | 38.2 | **40.8** | **87.7** | **43.6** | 42.6 | **25.4** | 28.0 | 46.6 |
| **Ours** | NO | **69.1** | **39.0** | 40.0 | 86.9 | 43.0 | **42.8** | 24.5 | **28.1** | **46.7** |
| **With Post-processing** | | | | | | | | | | |
| ***CLIP-ViT-B/16*** | | | | | | | | | | |
| CLIPer(Sun et al., 2024a) | YES | 65.9 | 37.6 | 39.0 | 85.2 | - | 41.7 | 21.2 | 27.5 | - |
| Trident(Shi et al., 2024) | YES | 67.1 | 38.6 | **41.1** | 84.5 | 42.9 | 42.2 | 21.9 | 28.3 | 45.8 |
| **Ours** | YES | **70.4** | **39.8** | 40.8 | **86.2** | **46.4** | **43.6** | **22.2** | **28.7** | **47.3** |
| ***CLIP-ViT-L/14*** | | | | | | | | | | |
| CLIPer(Sun et al., 2024a) | YES | 69.8 | 38.0 | **43.3** | **90.0** | - | **43.6** | **24.4** | **28.7** | - |
| Trident(Shi et al., 2024) | YES | 62.6 | 37.3 | 40.5 | 85.5 | 43.0 | 40.9 | 24.0 | 27.1 | 44.3 |
| **Ours** | YES | **71.3** | **39.6** | 40.7 | 86.5 | **46.4** | 43.3 | 23.8 | 28.5 | **47.5** |
| ***OpenCLIP-ViT-H/14*** | | | | | | | | | | |
| CLIPer(Sun et al., 2024a) | YES | **88.9** | 39.3 | **42.8** | **88.8** | - | 43.2 | 24.4 | 28.3 | - |
| Trident(Shi et al., 2024) | YES | 70.8 | 40.1 | 42.2 | 88.7 | **47.6** | 44.3 | **26.7** | 28.6 | **48.6** |
| **Ours** | YES | 71.7 | **40.5** | 41.4 | 87.6 | 47.0 | **44.8** | 25.6 | **28.8** | 48.4 |

*Table 2.* Ablation Study on Different VFM Feature Extractors. We compare four DINO variants: ViT-Base with patch sizes of 8 (B8) and 16 (B16), and ViT-Small with patch sizes of 8 (S8) and 16 (S16); two DINOV2 variants: ViT-B/14 and ViT-S/14; and the ViT-B/16 architecture from SAM. Results demonstrate that DINOV2 ViT-S/14 achieves the best performance on the VOC20 dataset, while DINO ViT-B/8 outperforms all other architectures across the remaining datasets.

| Model | Dataset | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | VOC21 | Context-60 | COCO | VOC20 | Cityscapes | Context-59 | ADE20K | COCO-Stuff |
| **DINO** | | | | | | | | |
| ViT-B/8 | **70.41** | **39.82** | **40.80** | 86.18 | **46.41** | **43.56** | **22.24** | **28.72** |
| ViT-B/16 | 69.26 | 39.67 | 40.46 | 85.99 | 45.56 | 42.47 | 22.11 | 27.88 |
| ViT-S/8 | 69.96 | 39.49 | 40.63 | 86.02 | 45.94 | 42.79 | 22.17 | 28.24 |
| ViT-S/16 | 68.91 | 39.12 | 40.27 | 85.88 | 45.23 | 42.01 | 21.96 | 27.42 |
| **DINOV2** | | | | | | | | |
| ViT-S/14 | 68.26 | 39.13 | 40.66 | **86.61** | 44.33 | 43.09 | 21.78 | 28.41 |
| ViT-B/14 | 68.17 | 39.03 | 40.57 | 86.50 | 44.59 | 43.07 | 21.87 | 28.48 |
| **SAM** | | | | | | | | |
| ViT-B/16 | 68.36 | 38.0 | 38.84 | 84.33 | 43.64 | 40.68 | 21.16 | 27.37 |

*Table 3.* Performance comparison of our approach with other methods on eight semantic segmentation benchmarks. We report results on: **With background category**: VOC21, Context60, COCO-Obj; **Without background category**: VOC20, City, Context59, ADE, Stuff. The best and second-best results are marked in **bold** and <u>underline</u>, respectively. The shaded region ours(SD) represents results obtained by processing our final feature maps using the SD model following CLIPer's(Sun et al., 2024a) approach.

| Method | Post-processing | With background | | | Without background | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| | | VOC21 | Context60 | COCO-Obj | VOC20 | City | Context59 | ADE | Stuff | |
| GroupViT(Xu et al., 2022) | No | 50.4 | 18.7 | 27.5 | 79.7 | 11.1 | 23.4 | 9.2 | 15.3 | 27.7 |
| TCL(Cha et al., 2023) | No | 51.2 | 24.3 | 30.4 | 77.5 | 23.1 | 30.3 | 14.9 | 19.6 | 33.1 |
| C-Dsier(Wysoczańska et al., 2023) | No | 62.2 | 32.4 | 35.0 | 80.2 | 31.7 | 35.9 | 20.0 | 24.6 | 40.3 |
| CoDe(Wu et al., 2024) | No | 57.5 | 30.5 | 32.3 | - | 28.9 | - | 17.7 | 23.9 | - |
| CLIP(Radford et al., 2021) | No | 20.8 | 9.3 | 8.9 | 49.1 | 6.7 | 11.2 | 3.2 | 5.7 | 14.4 |
| ReCo(Shin et al., 2022) | No | 25.1 | 19.9 | 15.7 | 57.7 | 21.6 | 22.3 | 11.2 | 14.8 | 23.7 |
| MaskCLIP(Ding et al., 2023) | No | 38.8 | 23.6 | 20.6 | 74.9 | 12.6 | 26.4 | 9.8 | 16.4 | 27.9 |
| CLIPSurgery(Li et al., 2023) | No | 55.2 | 30.3 | 29.7 | 77.5 | 33.1 | 33.4 | 16.1 | 22.2 | 37.2 |
| GEM(Bousselham et al., 2024) | No | 46.2 | 32.6 | 33.9 | 79.9 | 21.2 | 35.9 | 15.7 | 23.7 | 36.1 |
| SCLIP(Wang et al., 2025) | No | 59.1 | 30.4 | 30.5 | 80.4 | 32.2 | 34.2 | 16.1 | 22.4 | 38.2 |
| ClearCLIP(Lan et al., 2025b) | No | 57.0 | 32.2 | 32.5 | 82.3 | 32.8 | 35.8 | 17.3 | 24.0 | 39.2 |
| CAR(Sun et al., 2024b) | No | 48.6 | 13.6 | 15.4 | 73.7 | - | 18.4 | 5.4 | - | - |
| OVDiff(Karazija et al., 2023) | No | 66.3 | 29.7 | 34.6 | 80.9 | 23.4 | 32.9 | 14.1 | 20.3 | 37.8 |
| LAVG(Kang & Cho, 2025) | No | 62.1 | 31.6 | 34.2 | 82.5 | 26.2 | 34.7 | 15.8 | 23.2 | 38.8 |
| ProxyCLIP(Lan et al., 2025a) | No | 61.3 | 35.3 | 37.5 | 80.3 | 38.1 | 39.1 | 20.2 | 26.5 | 42.3 |
| CLIPer(Sun et al., 2024a) | No | 60.1 | 34.8 | 36.0 | <u>84</u> | - | 38.5 | 19.8 | 25.3 | - |
| SCCLIP(Bai et al., 2024) | No | 64.6 | 36.8 | 37.7 | 84.3 | 41.0 | 40.1 | 20.1 | 26.6 | 43.9 |
| Trident(Shi et al., 2024) | No | <u>64.5</u> | <u>37.2</u> | **39.5** | 83.7 | <u>40.4</u> | <u>40.9</u> | <u>20.9</u> | <u>27.6</u> | <u>44.5</u> |
| **Ours** | No | **66.1** | **37.7** | <u>39.4</u> | **85.4** | **42.8** | **41.2** | **21.0** | **27.3** | **44.8** |
| **Post-processing** | | | | | | | | | | |
| NACLIP(Hajimiri et al., 2024) | Yes | 64.1 | 35.0 | 36.2 | 83.0 | 38.3 | 38.4 | 19.1 | 25.7 | 42.5 |
| CLIPer(Sun et al., 2024a) | Yes | 65.9 | 37.6 | 39.0 | <u>85.2</u> | - | 41.7 | 21.2 | 27.5 | - |
| Trident(Shi et al., 2024) | Yes | <u>67.1</u> | <u>38.6</u> | **41.1** | 84.5 | <u>42.9</u> | <u>42.2</u> | <u>21.9</u> | <u>28.3</u> | <u>45.8</u> |
| **Ours** | Yes | **70.4** | **39.8** | <u>40.8</u> | **86.2** | **46.4** | **43.6** | **22.2** | **28.7** | **47.3** |
| **Ours(SD)** | Yes | 71.0 | 40.2 | 41.4 | 86.4 | 46.9 | 43.9 | 22.3 | 29.1 | 47.7 |

*Table 4.* Efficiency Evaluation and Ablation Study of Post-processing Methods. Performance and computational efficiency across seven datasets without post-processing, conducted on a single RTX4090 24G GPU. The table compares various post-processing approaches: Trident SAM (from Trident(Shi et al., 2024)), Combination of Trident and CLIPer(Sun et al., 2024a) (Trident SAM+SD), and our speed-optimized version (Optimization SAM). The bottom rows present ablation experiments on Optimization SAM. Note: Coco-obj dataset experiments were omitted due to GPU memory constraints (24GB). Optimization SAM maintains the same core principles as Trident SAM.

| Method | With background | | Without background | | | | | Avg. | Total Time |
|---|---|---|---|---|---|---|---|---|---|
| | VOC21 | Context60 | VOC20 | City | Context59 | ADE | Stuff | | |
| **Main Methods** | | | | | | | | | |
| **Trident SAM+SD** | **71.0** | **40.2** | **86.4** | **46.9** | **43.9** | **22.3** | **29.1** | **48.5** | – |
| Time | 22min 2s | 87min 44s | 22min | 41min 16s | 88min 27s | 29min 31s | 93min 51s | – | 384min 51s |
| **Trident SAM** | 70.4 | 39.8 | 86.2 | 46.4 | 43.6 | 22.2 | 28.7 | 48.2 | – |
| Time | 2min 50s | 20min 03s | 2min 39s | 6min 33s | 20min 25s | 5min 26s | 24min 11s | – | 82min 7s |
| **Optimization SAM** | 70.4 | 39.8 | 86.2 | 46.4 | 43.6 | 22.2 | 28.7 | 48.2 | – |
| Time | **2min 30s** | **19min 28s** | **2min 26s** | **5min 52s** | **19min 46s** | **4min 41s** | **22min 03s** | – | **76min 46s** |
| NO SAM | 66.1 | 37.7 | 85.4 | 42.8 | 41.2 | 21.0 | 27.3 | 45.9 | - |
| Time | 1min 36s | 15min 33s | 1min 33s | 4min 8s | 15min 23s | 2min 23s | 16min 16s | - | 56min 52s |
| **Ablation Studies (based on Optimization SAM)** | | | | | | | | | |
| w/o CCA | 69.1 | 39.1 | 86.0 | 42.4 | <u>42.9</u> | <u>21.5</u> | <u>28.16</u> | <u>47.0</u> | – |
| w/o box | 65.7 | 38.0 | 84.7 | 42.9 | 41.4 | 21.3 | 27.2 | 46.0 | – |
| w/o mask | **71.0** | <u>39.7</u> | <u>86.1</u> | <u>46.2</u> | **43.6** | **22.2** | **28.7** | **48.2** | – |
| **full model** | <u>70.4</u> | **39.8** | **86.2** | **46.4** | **43.6** | **22.2** | **28.7** | **48.2** | – |

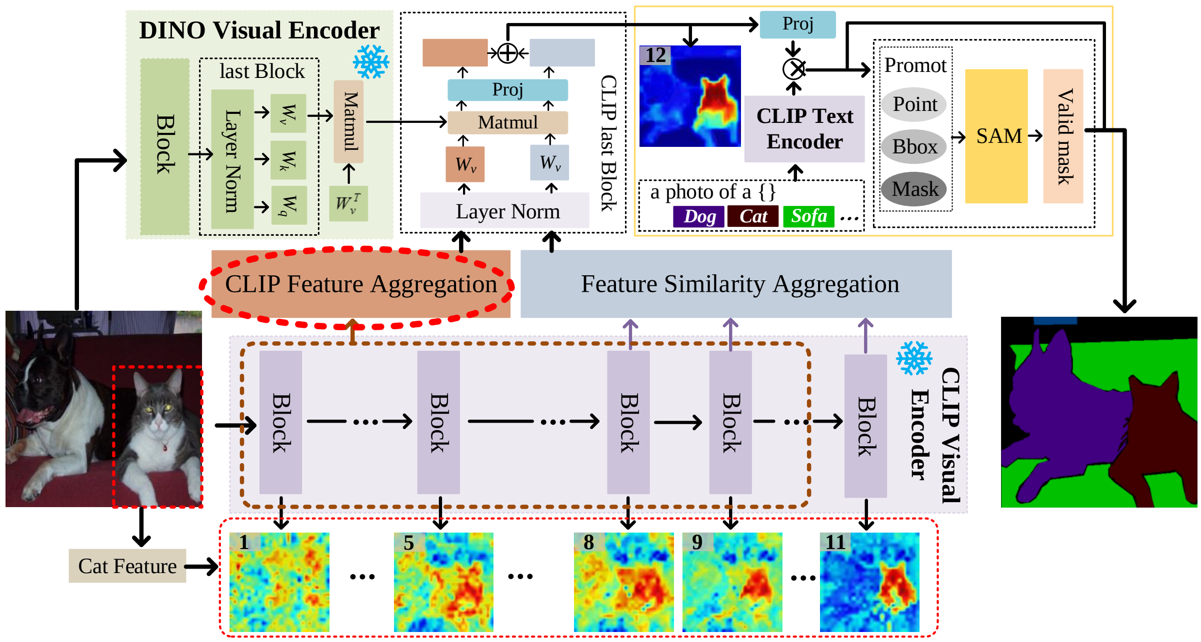


*Figure 1.* Overall architecture of the proposed S2CLIP. The framework consists of hierarchical progressive feature extraction from CLIP, feature fusion between CLIP and DINO, and SAM post-processing stage.

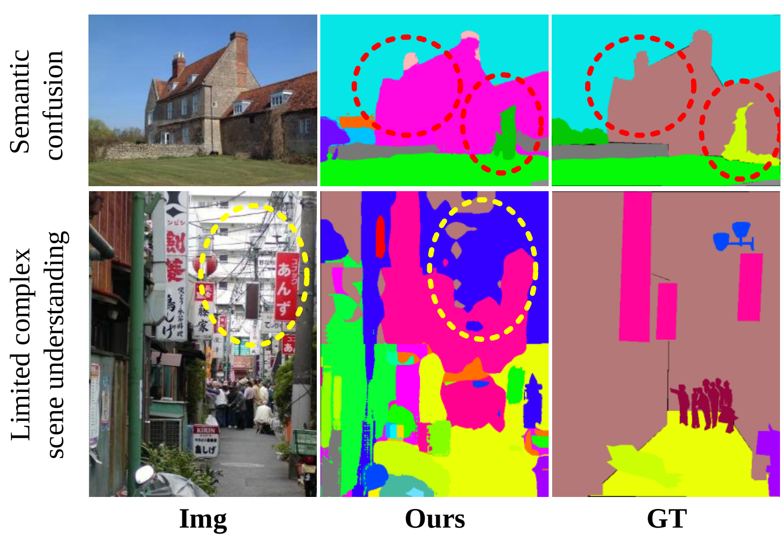


*Figure 2.* Failure Cases of S2CLIP. The upper section demonstrates semantic category confusion in S2CLIP, while the lower section illustrates that S2CLIP's understanding of extremely complex scenes requires further improvement. These limitations stem from two primary factors: (1) S2CLIP's use of simple CLIP text templates fails to adequately capture subtle semantic variations and state descriptions of objects in complex scenes, and (2) S2CLIP lacks the capability to explicitly model inter-object relationships.

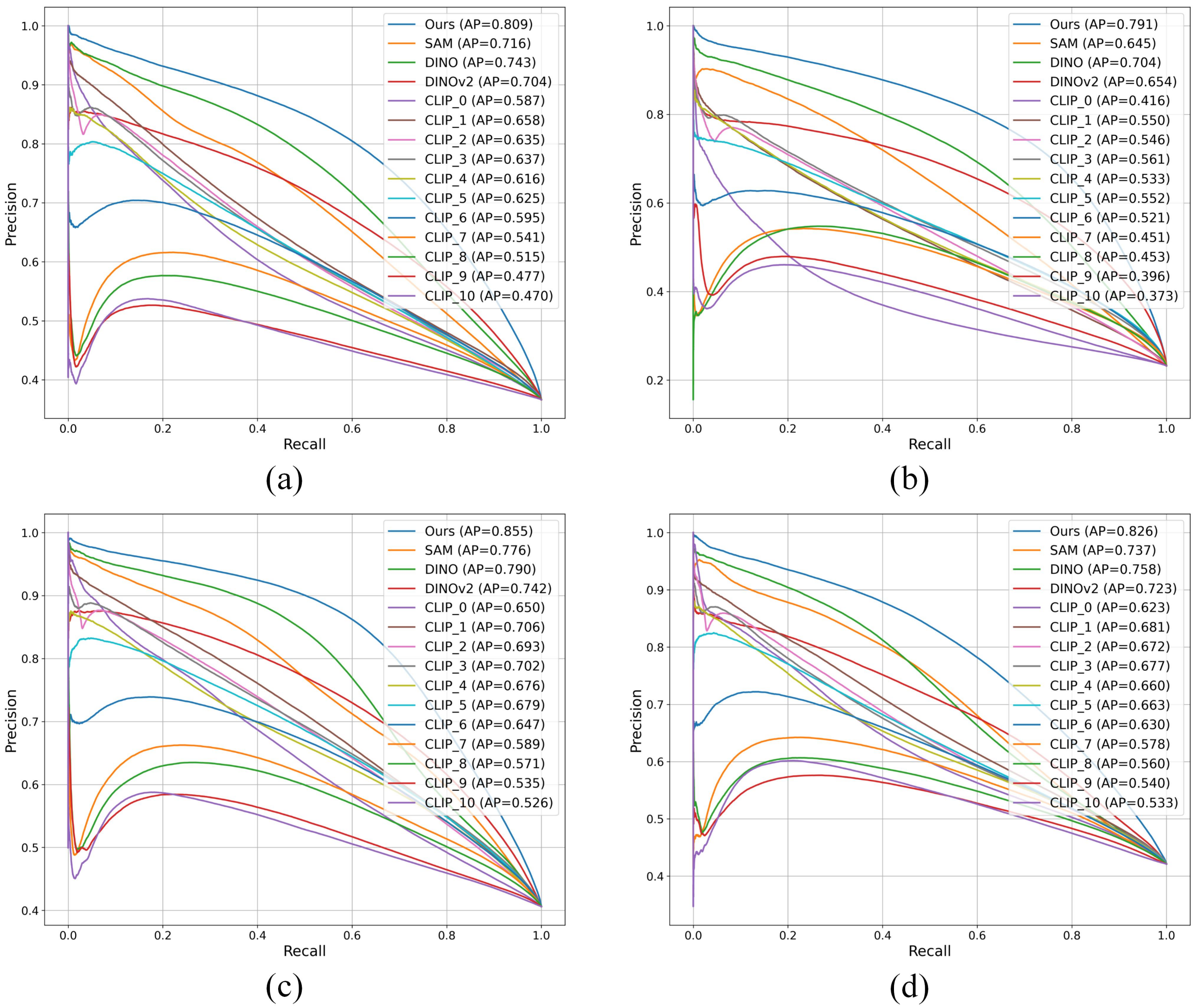


*Figure 3.* Precision-Recall Performance of Vision Foundation Models for Semantic Consistency Classification. Comparison of Average Precision (AP) between CLIP (layers 0-10) and other Vision Foundation Models (DINO, DINOV2, and SAM) across (a) ADE20K, (b) Cityscapes, (c) COCO-Stuff, and (d) Pascal Context datasets. Higher AP indicates better performance in distinguishing whether image patches belong to the same semantic category.

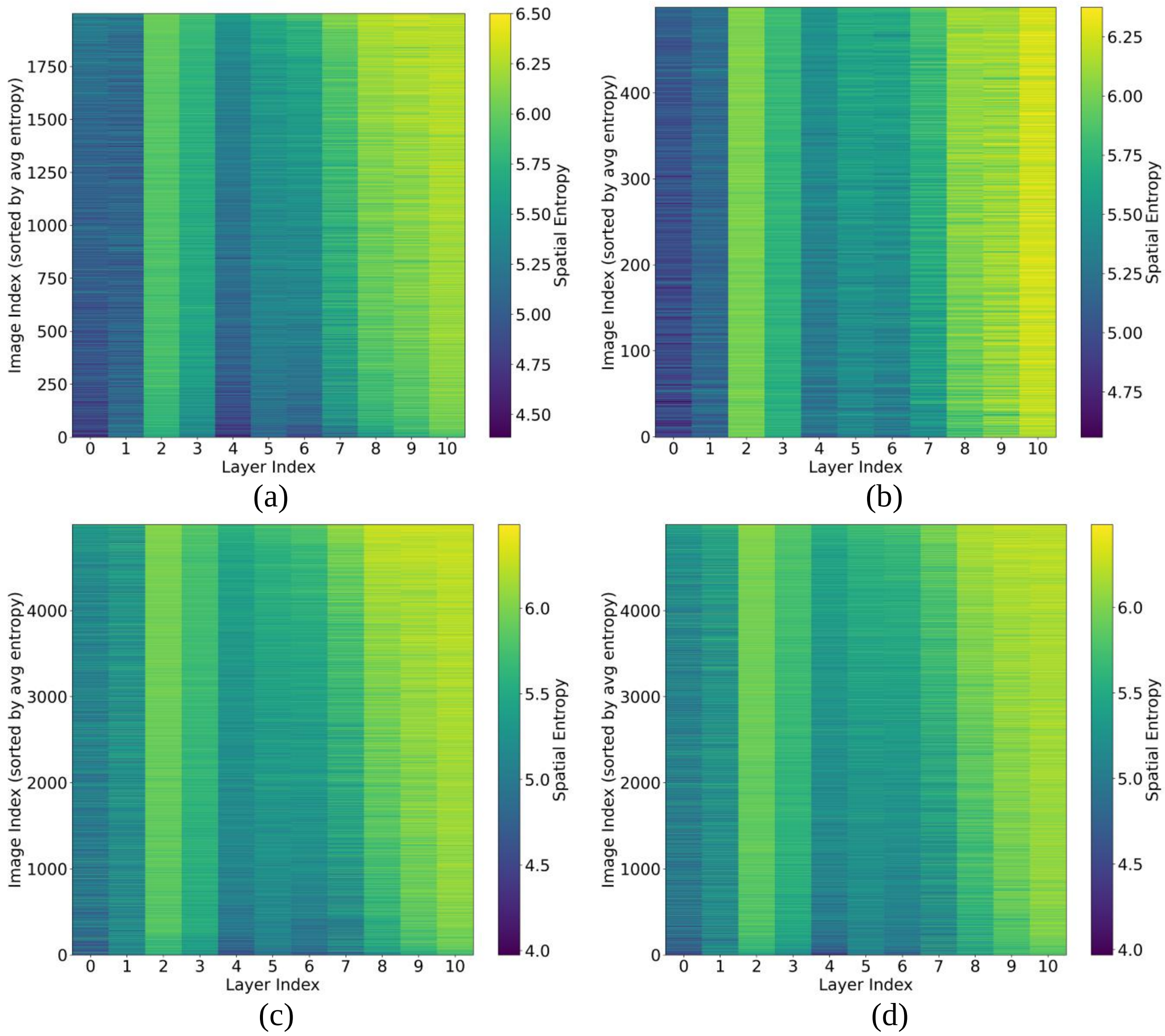


*Figure 4.* Spatial Entropy Distribution Across CLIP Layers on Multiple Datasets. Visualization showing spatial entropy values across CLIP layers for (a) ADE20K, (b) Cityscapes, (c) COCO-Stuff, and (d) Pascal Context datasets. Darker regions indicate lower spatial entropy. The spatial entropy computation methodology is described in Algorithm 1.

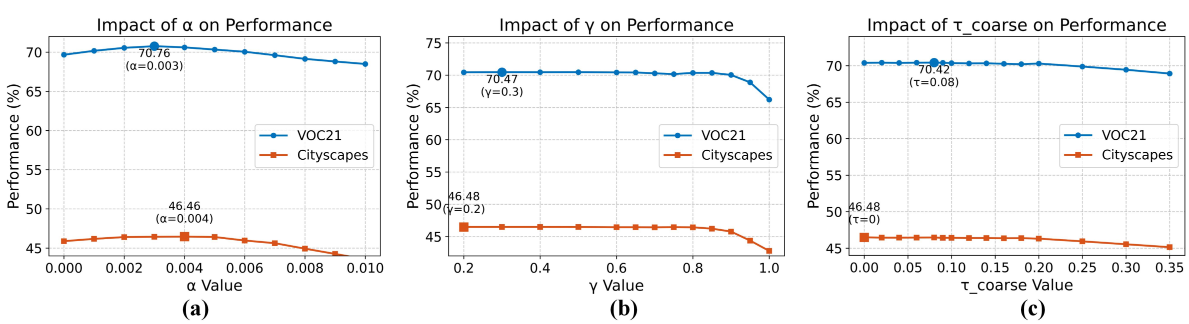


*Figure 5.* Overall architecture of the proposed S2CLIP. The framework consists of hierarchical progressive feature extraction from CLIP, feature fusion between CLIP and DINO, and SAM post-processing stage.

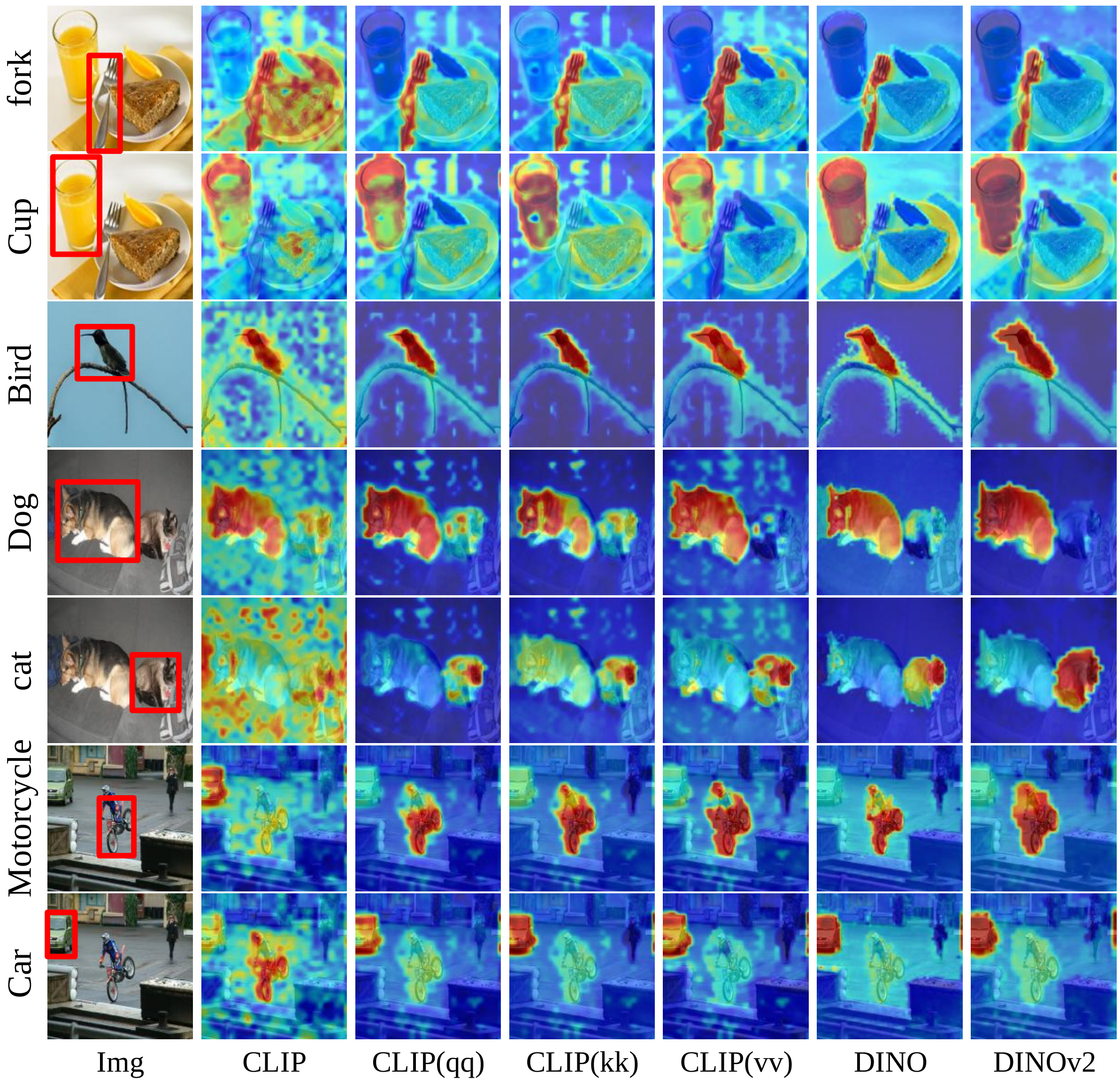


*Figure 6.* Feature Visualization Comparison of Vision Encoder Architectures. Qualitative comparison of features extracted by CLIP, DINO, and DINOV2 visual encoders. CLIP(qq), CLIP(kk), and CLIP(vv) represent different attention weight recombination strategies in CLIP's final layer. Standard CLIP outputs (column 2) show limited localization capability, while self-self attention modifications (columns 3-5) improve localization despite noise artifacts. DINO and DINOV2 (columns 6-7) demonstrate superior object-level feature localization.

**Algorithm 1** Spatial Entropy Calculation for Feature Maps

**Input:** $\mathbf{F} \in \mathbb{R}^{B \times C \times H \times W}$ {Feature map}

Preprocess and reshape $\mathbf{F}$ to $\mathbf{F} \in \mathbb{R}^{B \times C \times N}$ where $N = H \times W$ {$N$: spatial dimensions}

$\boldsymbol{\mu} \leftarrow \mathbb{E}[\mathbf{F}]_N$

$\boldsymbol{\sigma} \leftarrow \sqrt{\mathbb{V}[\mathbf{F}]_N}$

$\varepsilon \leftarrow \max(\min(\boldsymbol{\sigma}) \cdot 10^{-2}, 10^{-6})$

$\hat{\mathbf{F}} \leftarrow \frac{\mathbf{F} - \boldsymbol{\mu}}{\boldsymbol{\sigma} + \varepsilon}$

$\mathbf{F}^* \leftarrow \hat{\mathbf{F}} - \max(\hat{\mathbf{F}})_N$

$\mathbf{P} \leftarrow \frac{\exp(\mathbf{F}^*)}{\sum_{n=1}^{N} \exp(\mathbf{F}^*_n) + \varepsilon}$ {Normalized probability distribution}

$\mathbf{P} \leftarrow \frac{\max(\mathbf{P}, \varepsilon)}{\sum_{n=1}^{N} \max(\mathbf{P}_n, \varepsilon)}$

$\mathbf{H} \leftarrow -\sum_{n=1}^{N} \mathbf{P}_n \log(\mathbf{P}_n)$ {Entropy per channel}

$\mathbf{P}_{sort} \leftarrow \text{sort}(\mathbf{P})_N$

$\mathbf{L} \leftarrow \text{cumsum}(\mathbf{P}_{sort})_N$

$\mathbf{G} \leftarrow 1 - \frac{2}{N} \sum_{i=1}^{N} \frac{i \cdot \mathbf{L}_i}{\sum_{j=1}^{N} \mathbf{P}_{sort,j}}$ {Gini coefficient}

$\boldsymbol{\rho} \leftarrow \mathbf{0}_B$ {Channel correlation}

**if** $C > 1$ **then**

  **for** $b = 1$ **to** $B$ **do**

    $\mathbf{R}_b \leftarrow \text{corr}(\mathbf{P}_b)$

    $\boldsymbol{\rho}_b \leftarrow \frac{1}{C(C-1)} \sum_{i \neq j} \mathbf{R}_{b,i,j}$

  **end for**

**end if**

$\bar{H} \leftarrow \frac{1}{BC} \sum_{b=1}^{B} \sum_{c=1}^{C} \mathbf{H}_{b,c}$ {Mean spatial entropy}

**if** $\bar{H} \notin \mathbb{R}$ **then**

  $\bar{H} \leftarrow 0$

**end if**

**Output:** $\bar{H}$

**Algorithm 2** Original Connected Component Analysis

**Input:** $\mathbf{S} \in \mathbb{R}^{C\times H\times W}$ {Segmentation mask}
**Input:** $\mathbf{L} \in \mathbb{R}^{C\times H\times W}$ {Segmentation logits}
**Input:** $\tau$ {Coarse threshold}
**Input:** $A_{\min}$ {Minimal area threshold}
**Input:** split_last $\in \{\text{True}, \text{False}\}$

$\mathcal{R} \leftarrow \emptyset, \mathcal{B} \leftarrow \emptyset, \mathcal{S} \leftarrow \emptyset, \mathcal{P} \leftarrow \emptyset$
**for** $c \leftarrow 0$ **to** $C-1$ **do**
  **if** $\neg$split_last $\wedge\, c = C-1$ **then**
    **continue**
  **end if**
  $\mathcal{R}[c] \leftarrow \emptyset, \mathcal{P}[c] \leftarrow \emptyset, \mathcal{S}[c] \leftarrow \emptyset, \mathcal{B}[c] \leftarrow \emptyset$
  $\mathbf{M}_c \leftarrow \mathbf{S}[c]$ {Class mask}
  $\mathbf{L}_c \leftarrow \mathbf{L}[c]$ {Class logit}
  **if** $\max(\mathbf{L}_c) < \tau$ **then**
    **continue**
  **end if**
  $\mathbf{M}_{\text{labeled}} \leftarrow \mathcal{F}_{\text{label}}(\mathbf{M}_c, 2)$ {2-connected labeling}
  $N_{\text{regions}} \leftarrow \max(\mathbf{M}_{\text{labeled}})$
  **for** $r \leftarrow 1$ **to** $N_{\text{regions}}$ **do**
    $\mathbf{M}_r \leftarrow \mathbb{I}[\mathbf{M}_{\text{labeled}} = r]$ {Region indicator mask}
    **if** $\sum_{i,j} \mathbf{M}_r[i,j] < A_{\min}$ **then**
      **continue**
    **end if**
    $s_r \leftarrow \frac{1}{|\Omega_r|}\sum_{(i,j)\in\Omega_r} \mathbf{L}_c[i,j]$ {$\Omega_r$ is region $r$}
    $\mathbf{L}_r \leftarrow \mathbf{L}_c \odot \mathbf{M}_r$
    $(i^*, j^*) \leftarrow \arg\max_{i,j} \mathbf{L}_r[i,j]$
    $\mathcal{P}[c] \leftarrow \mathcal{P}[c] \cup \{(j^*, i^*)\}$
    $\mathcal{R}[c] \leftarrow \mathcal{R}[c] \cup \{\mathbf{M}_r\}$
    $\mathcal{S}[c] \leftarrow \mathcal{S}[c] \cup \{s_r\}$
  **end for**
  **if** $|\mathcal{R}[c]| > 0$ **then**
    $\mathcal{R}[c] \leftarrow \text{stack}(\mathcal{R}[c])$ {Stack region masks}
    **if** $\mathcal{R}[c] = \emptyset$ **then**
      $\mathcal{B}[c] \leftarrow \mathcal{R}[c]$
    **else**
      $\mathcal{B}[c] \leftarrow \mathcal{F}_{\text{masks\_to\_boxes}}(\mathcal{R}[c])$
    **end if**
  **end if**
**end for**
**Output:** $\mathcal{R}, \mathcal{B}, \mathcal{S}, \mathcal{P}$ {Regions, boxes, scores, points}

**Algorithm 3** Optimized Connected Component Analysis

**Input:** $\mathbf{S} \in \mathbb{R}^{C\times H\times W}$ {Segmentation mask}
**Input:** $\mathbf{L} \in \mathbb{R}^{C\times H\times W}$ {Segmentation logits}
**Input:** $\tau$ {Coarse threshold}
**Input:** $A_{\min}$ {Minimal area threshold}
**Input:** split_last $\in \{\text{True}, \text{False}\}$

$\mathcal{R} \leftarrow \emptyset, \mathcal{B} \leftarrow \emptyset, \mathcal{S} \leftarrow \emptyset, \mathcal{P} \leftarrow \emptyset$
$\mathcal{D} \leftarrow \text{device}(\mathbf{L})$ {Device information}
$\mathbf{L}_{\max} \leftarrow \max_{h,w} \mathbf{L}_{c,h,w} \quad \forall c \in \{0, 1, \ldots, C-1\}$ {Class-wise max values}
$\mathcal{C}_{\text{valid}} \leftarrow \{c \mid \mathbf{L}_{\max}[c] \geq \tau\}$ {Pre-filter valid classes}
**for** $c \leftarrow 0$ **to** $C-1$ **do**
  **if** $\neg$split_last $\wedge\, c = C-1$ **then**
    **continue**
  **end if**
  $\mathcal{R}[c] \leftarrow \emptyset, \mathcal{P}[c] \leftarrow \emptyset, \mathcal{S}[c] \leftarrow \emptyset, \mathcal{B}[c] \leftarrow \emptyset$
  **if** $c \notin \mathcal{C}_{\text{valid}}$ **then**
    **continue**
  **end if**
  $\mathbf{M}_c \leftarrow \mathbf{S}[c]$ {Class mask (CPU memory)}
  $\mathbf{L}_c \leftarrow \mathbf{L}[c]$ {Class logit}
  $(\mathbf{M}_{\text{labeled}}, N_{\text{labels}}, \mathbf{\Gamma}) \leftarrow \mathcal{F}_{\text{CC}}(\mathbf{M}_c, 8)$ {8-connected components with stats}
  **if** $N_{\text{labels}} \leq 1$ **then**
    **continue**
  **end if**
  $\mathcal{V} \leftarrow \emptyset$ {Valid masks}
  **for** $r \leftarrow 1$ **to** $N_{\text{labels}} - 1$ **do**
    $\mathbf{M}_r \leftarrow \mathbb{I}[\mathbf{M}_{\text{labeled}} = r]$ {Region indicator mask}
    $A_r \leftarrow \mathbf{\Gamma}[r, \text{AREA}]$ {Region area from component stats}
    **if** $A_r < A_{\min}$ **then**
      **continue**
    **end if**
    $\mathbf{M}_r^{\mathcal{D}} \leftarrow \mathcal{F}_{\text{to\_device}}(\mathbf{M}_r, \mathcal{D})$ {Transfer to device}
    $s_r \leftarrow \frac{1}{|\Omega_r|}\sum_{(i,j)\in\Omega_r} \mathbf{L}_c[i,j]$ {$\Omega_r$ is region $r$}
    $\mathbf{L}_r \leftarrow \mathbf{L}_c \odot \mathbf{M}_r^{\mathcal{D}}$
    $(i^*, j^*) \leftarrow \arg\max_{i,j} \mathbf{L}_r[i,j]$
    $\mathcal{P}[c] \leftarrow \mathcal{P}[c] \cup \{(j^*, i^*)\}$
    $\mathcal{R}[c] \leftarrow \mathcal{R}[c] \cup \{\mathbf{M}_r\}$
    $\mathcal{S}[c] \leftarrow \mathcal{S}[c] \cup \{s_r\}$
    $\mathcal{V} \leftarrow \mathcal{V} \cup \{\mathbf{M}_r\}$ {Track valid masks}
  **end for**
  **if** $|\mathcal{R}[c]| > 0$ **then**
    $\mathcal{R}[c] \leftarrow \mathcal{F}_{\text{to\_device}}(\text{stack}(\mathcal{V}), \mathcal{D})$ {Stack valid masks}
    $\mathcal{B}[c] \leftarrow \mathcal{F}_{\text{masks\_to\_boxes}}(\mathcal{R}[c])$ {Convert masks to boxes}
  **else**
    $\mathcal{B}[c] \leftarrow \mathcal{R}[c]$ {Empty collection for invalid classes}
  **end if**
**end for**
**Output:** $\mathcal{R}, \mathcal{B}, \mathcal{S}, \mathcal{P}$ {Regions, boxes, scores, points}

**Algorithm 4** Optimized SAM Refinement

**Input:** $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$, $\mathcal{S} \in \mathbb{Z}^{H\times W}$, $\mathcal{L} \in \mathbb{R}^{C\times H\times W}$, $C \in \mathbb{Z}^{+}$, $\mathcal{P}$
**Output:** $\mathcal{S}' \in \mathbb{Z}^{H\times W}$, $\mathcal{Q} \in \mathbb{R}^{N}$, $\mathcal{L}' \in \mathbb{R}^{C\times H\times W}$, $\mathcal{B}' \in \mathbb{R}^{N\times 4}$

$d \leftarrow \mathcal{S}$.device, $r \leftarrow 2$, $(h_d, w_d) \leftarrow (\lfloor H/r \rfloor, \lfloor W/r \rfloor)$
**with** no_grad() and mixed_precision($d$):
$\mathcal{S}_d \leftarrow \mathcal{F}_{\text{down}}(\mathcal{S}, h_d, w_d)$, $\mathcal{L}_d \leftarrow \mathcal{F}_{\text{down}}(\mathcal{L}, h_d, w_d)$
**if** $\tau_c > 0$ **then**
  $\mathcal{S}_d[\max_c(\mathcal{L}_d) < \tau_c] \leftarrow C$, $\mathcal{M} \leftarrow \Phi(\mathcal{S}_d, C+1)$ {One-hot encoding}
**else**
  $\mathcal{M} \leftarrow \Phi(\mathcal{S}_d, C)$
**end if**
$\mathcal{M}_{\text{cpu}} \leftarrow \Pi_{\text{mem}}(\mathcal{M}, d)$, $(\mathcal{R}, \mathcal{B}, \mathcal{Q}_b, \mathcal{P}) \leftarrow$ SplitRegions$(\mathcal{M}_{\text{cpu}}, \mathcal{L}_d, \tau_c)$

**if** $\forall i \in [0, C) : |\mathcal{B}_i| = 0$ **then return** $(\mathcal{S}, \emptyset, \mathcal{L}, \emptyset)$
**end if**
  $b_{\text{size}} \leftarrow \min(32, \max(1, \lfloor M_{\text{free}}/(3\cdot 10^9)\rfloor))$, $\mathcal{P}$.PrecomputeTransforms$(h_d, w_d, H, W)$
  $\mathbb{I} \leftarrow \emptyset$, $\mathcal{B}_c, \mathcal{M}_v, \mathcal{Q}_v, \mathcal{L}_v, \mathcal{R}_f \leftarrow \emptyset, \emptyset, \emptyset, \emptyset, \emptyset$ {Collections init}
  **for** $i \in [0, C)$ **do**
    **if** $|\mathcal{R}_i| > 0$ **then**
      $(h_n, w_n) \leftarrow \mathcal{P}$.GetPreprocessShape$(H, W, 256)$, $\mathcal{L}_r^i \leftarrow \mathcal{F}_{\text{resize}}(\mathcal{L}_d[i], h_n, w_n)$
      $\mathcal{L}_r^i \leftarrow \mathcal{F}_{\text{pad}}(\mathcal{L}_r^i, 256 - w_n, 256 - h_n) \cdot \tau_m \cdot \mathbb{I}_{\{\mathcal{L}_r^i > \tau_c\}}$
      $\mathcal{B}_t^i \leftarrow \mathcal{P}$.TransformBoxes$(\mathcal{B}_i, h_d, w_d)$, $\mathcal{P}_t^i \leftarrow \mathcal{P}$.TransformPoints$(\mathcal{P}_i, h_d, w_d)$
      $\mathbb{I} \leftarrow \mathbb{I} \cup \{(i, \mathcal{B}_t^i, \mathcal{P}_t^i, \mathcal{L}_r^i, \mathcal{R}_i, \mathcal{B}_i, \mathcal{P}_i)\}$
    **end if**
  **end for**

  **for** $j \leftarrow 0$ **to** $|\mathbb{I}|$ **step** $b_{\text{size}}$ **do**
    $\mathbb{I}_j \leftarrow \mathbb{I}[j : j + b_{\text{size}}]$
    **for each** $(i, \mathcal{B}_t^i, \mathcal{P}_t^i, \mathcal{L}_r^i, \mathcal{R}_i, \mathcal{B}_i, \mathcal{P}_i) \in \mathbb{I}_j$ **do**
      $(\mathcal{M}_i, \mathcal{Q}_i, \mathcal{L}_i) \leftarrow \mathcal{P}$.Predict$(\mathcal{P}_t^i, \mathbf{1}, \mathcal{B}_t^i, \mathcal{L}_r^i)$, $\mathcal{V}_i \leftarrow \{\mathcal{Q}_i > \tau_{iou}\}$
      **if** $|\mathcal{V}_i| < |\mathcal{R}_i|$ **then** $\mathcal{R}_f \leftarrow \mathcal{R}_f \cup \{\mathcal{R}_i[\neg\mathcal{V}_i]\}$
      **end if**
      **if** $|\mathcal{V}_i| > 0$ **then** $\mathcal{M}_v \leftarrow \mathcal{M}_v \cup \{\mathcal{M}_i[\mathcal{V}_i]\}$, $\mathcal{Q}_v \leftarrow \mathcal{Q}_v \cup \{\mathcal{Q}_i[\mathcal{V}_i]\}$, $\mathcal{L}_v \leftarrow \mathcal{L}_v \cup \{\mathcal{L}_i[\mathcal{V}_i]\}$, $\mathcal{B}_c \leftarrow \mathcal{B}_c \cup \{\mathcal{B}_i[\mathcal{V}_i]\}$
      **end if**
    **end for**
  **end for**
    **if** $|\mathcal{B}_c| > 0$ **then**
      **with** mixed_precision($d$):
      $\mathcal{M}_{all} \leftarrow$ Concat$(\mathcal{M}_v)$, $\mathcal{Q}_{all} \leftarrow$ Concat$(\mathcal{Q}_v)$, $\mathcal{L}_{all} \leftarrow \mathcal{P}$.PostprocessMasks(Concat$(\mathcal{L}_v), H, W)$
      $\mathcal{B}_{all} \leftarrow$ Concat$(\mathcal{B}_c) \cdot r$, $(\mathcal{S}', \mathcal{L}') \leftarrow$ MapRefinement$(\mathcal{M}_{all}, \sigma(\mathcal{L}_{all}), \mathcal{B}, \mathcal{L})$
      $\mathcal{S}' \leftarrow \mathcal{S}' \cdot \mathbb{I}_{\{\sum_c \mathcal{L}'_c > 0\}} + \mathcal{S} \cdot \mathbb{I}_{\{\sum_c \mathcal{L}'_c = 0\}}$
      **if** $|\mathcal{R}_f| > 0$ **then** $\mathcal{R}_f \leftarrow$ Resize(Concat$(\mathcal{R}_f), H, W)$, $(\mathcal{S}', \mathcal{L}') \leftarrow$ MapFailedRegions$(\mathcal{S}', \mathcal{L}', \mathcal{R}_f, \mathcal{L}, \mathcal{S})$
      **end if**
      **return** $(\mathcal{S}', \mathcal{Q}_{all}, \mathcal{L}', \mathcal{B}_{all})$
    **else**
      **return** $(\mathcal{S}, \emptyset, \mathcal{L}, \emptyset)$
    **end if**