Rethinking Semantic Segmentation: A Prototype View

Tianfei Zhou\textsuperscript{1}, Wenguan Wang\textsuperscript{2,1*}, Ender Konukoglu\textsuperscript{1}, Luc Van Gool\textsuperscript{1}

\textsuperscript{1} Computer Vision Lab, ETH Zurich \textsuperscript{2} ReLER, AAII, University of Technology Sydney

CVPR 2022, Oral
Parametric & Non-Parametric

(a) FCN/attention network → dense feature → parametric softmax → dense prediction

(b) test image → attention network → dense feature → dense prediction

(c) test image → FCN/attention network → dense feature → nearest prototype decision → dense prediction

training images → embedding space → prototype selection → non-learnable prototypes
Introduction

Question:

1. What are the relation and difference between FCN based and attention based mask decoding strategies?

2. If the learnable query vectors indeed implicitly capture some intrinsic properties of data, is there any better way to achieve this?
Parametric & Non-Parametric

\[ W = [\mathbf{w}_1, \ldots, \mathbf{w}_C] \in \mathbb{R}^{C \times D}, \mathbf{w}_c \in \mathbb{R}^D \]

\[ E = [\mathbf{e}_1, \ldots, \mathbf{e}_C] \in \mathbb{R}^{C \times D} \]
Introduction

Question:

1. What are the relation and difference between FCN based and attention based mask decoding strategies?

2. If the learnable query vectors indeed implicitly capture some intrinsic properties of data, is there any better way to achieve this?

3. What are the limitations of this learnable prototype based parametric paradigm?

4. How to address these limitations?
Parametric Prototype Learning

(a) FCN/attention network

\[ W = [w_1, \cdots, w_c] \in \mathbb{R}^{C \times D}; w_c \in \mathbb{R}^D \]

\[ p(c|i) = \frac{\exp(w_c^T i)}{\sum_{c'=1}^{C} \exp(w_{c'}^T i)} \]

(b) Parametric softmax

\[ E = [e_1, \cdots, e_c] \in \mathbb{R}^{C \times D} \]

\[ p(c|i) = \frac{\exp(e_c \ast i)}{\sum_{c'=1}^{C} \exp(e_{c'} \ast i)} \]
Parametric Prototype Learning

\[ p(c|i) = \frac{\exp(-\langle i, g_c \rangle)}{\sum_{c' = 1}^{C} \exp(-\langle i, g_{c'} \rangle)}, \quad \langle \cdot, \cdot \rangle \text{ distance measure} \]
Introduction

Limitations:

1. Single learned prototype per class, insufficient to rich intra-class variance.

2. Much parameters needed for prototype learning, hurting generalizability.

3. Ignoring known inductive biases, intra-class compactness about feature distribution.
Architecture illustration

\[ I \in \mathbb{R}^{H \times W \times D} \]
\[ p_{c,k} \in \mathbb{R}^{D} \]
Non-Learnable Prototype based pixel classification

Pixel features, \( I \in R^{H \times W \times D} \), \( CK \) non-learnable prototypes \( \{p_{c,k} \in R^D\}_{c,k=1}^{C,K} \).

The category prediction of each pixel \( i \in I \).

\[
\hat{c}_i = c^*, \text{ with } (c^*, k^*) = \arg \min_{(c, k)} \{ \langle i, p_{c,k} \rangle \}_{c,k=1}^{C,K},
\]

Probability of Pixel \( i \) over the \( C \) class,

\[
p(c|i) = \frac{\exp(-s_{i,c})}{\sum_{c'=1}^{C} \exp(-s_{i,c'})}, \text{ with } s_{i,c} = \min_{k=1}^{K} \{ \langle i, p_{c,k} \rangle \}_{k=1}^{K},
\]

update prototypes,

\[
p_{c,k} \leftarrow \mu p_{c,k} + (1 - \mu) \tilde{i}_{c,k},
\]
Within-Class Online Clustering

Given pixels $I^c = \{i_n\}_{n=1}^N$ in a training batch that belong to class $c$. $K$ prototypes $\{p_{c,k}\}_{k=1}^K$ of class $c$.

Pixel-Prototype mapping, $L^c = [l_{in}]_{n=1}^N \in \{0,1\}^{K \times N},$

$$l_{in} = [l_{in,k}]_{k=1}^K \in \{0,1\}^K$$

Pixel embedding $X^c$, Prototypes $P^c$

$$\max_{L^c} \text{Tr}(L^c \cdot P^c \cdot X^c),$$

s.t. $L^c \in \{0,1\}^{K \times N}, L^c \cdot 1^K = 1^N, L^c \cdot 1^N = \frac{N}{K} \cdot 1^K,$
Within-Class Online Clustering

$$\max_{L^c} \text{Tr}(L^c P^c X^c) + \kappa h(L^c),$$

s.t.  \(L^c \in \mathbb{R}_+^{K \times N}, L^c 1^K = 1^N, L^c 1^N = \frac{N}{K} 1^K,\)

Solution using Sinkhorn-Knopp iteration.

$$L^c = \text{diag}(u) \exp \left( \frac{P^c X^c}{\kappa} \right) \text{diag}(v),$$
Training Objects

CE loss

\[ \mathcal{L}_{CE} = -\log p(c_i|i) = -\log \frac{\exp(-s_{i,c_i})}{\exp(-s_{i,c_i}) + \sum_{c' \neq c_i} \exp(-s_{i,c'})}. \]

Pixel-Prototype Contrastive Learning

\[ \mathcal{L}_{PPC} = -\log \frac{\exp(i^\top p_{c_i,k_i}/\tau)}{\exp(i^\top p_{c_i,k_i}/\tau) + \sum_{p^- \in \mathcal{P}^-} \exp(i^\top p^-/\tau)}, \]

Pixel-Prototype Distance Optimization

\[ \mathcal{L}_{PPD} = (1 - i^\top p_{c_i,k_i})^2. \]

\[ \mathcal{L}_{SEG} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{PPC} + \lambda_2 \mathcal{L}_{PPD}. \]
## Experiments

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th># Param (M)</th>
<th>mIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCR [ECCV20]</td>
<td>HRNetV2-W48 [140]</td>
<td>70.3</td>
<td>45.6</td>
</tr>
<tr>
<td>MaskFormer [NeurIPS21]</td>
<td>ResNet-101 [46]</td>
<td>60.0</td>
<td>46.0</td>
</tr>
<tr>
<td>OCR [ECCV20]</td>
<td>HRFormer-B [132]</td>
<td>70.3</td>
<td>48.7</td>
</tr>
<tr>
<td>Setr [CVR21]</td>
<td>ViT-Large [31]</td>
<td>318.3</td>
<td>50.2</td>
</tr>
<tr>
<td>Segmenter [ICCV21]</td>
<td>ViT-Large [31]</td>
<td>334.0</td>
<td>51.8</td>
</tr>
<tr>
<td>†MaskFormer [NeurIPS21]</td>
<td>Swin-Base [79]</td>
<td>102.0</td>
<td>52.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th># Param (M)</th>
<th>mIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSPNet [CVPR17]</td>
<td>ResNet-101 [46]</td>
<td>65.9</td>
<td>78.4</td>
</tr>
<tr>
<td>Segmenter [ICCV21]</td>
<td>ViT-Large [31]</td>
<td>322.0</td>
<td>79.1</td>
</tr>
<tr>
<td>OCR [ECCV20]</td>
<td>HRNetV2-W48 [110]</td>
<td>70.3</td>
<td>81.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th># Param (M)</th>
<th>mIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>ResNet-101 [46]</td>
<td>68.5</td>
<td>41.1↑1.2</td>
</tr>
<tr>
<td>HRNet [PAM20]</td>
<td>HRNetV2-W48 [110]</td>
<td>65.9</td>
<td>42.0</td>
</tr>
<tr>
<td>Ours</td>
<td>HRNetV2-W48 [110]</td>
<td>65.8</td>
<td>43.0↑1.0</td>
</tr>
<tr>
<td>Swin [ICCV21]</td>
<td>Swin-Base [79]</td>
<td>90.6</td>
<td>48.0</td>
</tr>
<tr>
<td>Ours</td>
<td>Swin-Base [79]</td>
<td>90.5</td>
<td>48.6↑0.6</td>
</tr>
<tr>
<td>SegFormer [NeurIPS21]</td>
<td>MiT-B4 [120]</td>
<td>64.1</td>
<td>50.9</td>
</tr>
<tr>
<td>Ours</td>
<td>MiT-B4 [120]</td>
<td>64.0</td>
<td>51.7↑0.8</td>
</tr>
</tbody>
</table>

†: backbone is pre-trained on ImageNet-22K.

Table 1. Quantitative results (§5.2) on ADE20K [142] val.

Table 2. Quantitative results (§5.2) on Cityscapes [23] val.
## Experiments

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th># Param (M)</th>
<th>mIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCR [ECCV20]</td>
<td>HRNetV2-W48 [110]</td>
<td>70.3</td>
<td>40.5</td>
</tr>
<tr>
<td><strong>FCN [CVPR15]</strong></td>
<td><strong>ResNet-101 [46]</strong></td>
<td><strong>68.6</strong></td>
<td><strong>32.5</strong></td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>HRNet [PAMI21]</strong></td>
<td><strong>HRNetV2-W48 [110]</strong></td>
<td><strong>65.9</strong></td>
<td><strong>38.7</strong></td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Swin [ICCV21]</strong></td>
<td><strong>Swin-Base [79]</strong></td>
<td><strong>90.6</strong></td>
<td><strong>41.5</strong></td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SegFormer [NeurIPS21]</strong></td>
<td><strong>MiT-B4 [120]</strong></td>
<td><strong>64.0</strong></td>
<td><strong>43.3</strong></td>
</tr>
</tbody>
</table>

Table 3. **Quantitative results** (§5.2) on COCO-Stuff [10] test.


Ablation Study

parametric v.s. nonparametric

<table>
<thead>
<tr>
<th>Method</th>
<th># Proto</th>
<th>150 classes</th>
<th></th>
<th>300 classes</th>
<th></th>
<th>500 classes</th>
<th></th>
<th>700 classes</th>
<th></th>
<th>847 classes</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>mIoU (%)</td>
<td># Param (M)</td>
<td>mIoU (%)</td>
<td># Param (M)</td>
<td>mIoU (%)</td>
<td># Param (M)</td>
<td>mIoU (%)</td>
<td># Param (M)</td>
<td>mIoU (%)</td>
<td># Param (M)</td>
</tr>
<tr>
<td>parametric</td>
<td>1</td>
<td>45.1</td>
<td>27.48 (0.12)</td>
<td>36.5</td>
<td>27.62 (0.23)</td>
<td>25.7</td>
<td>27.80 (0.39)</td>
<td>19.8</td>
<td>27.98 (0.54)</td>
<td>16.5</td>
<td>28.11 (0.65)</td>
</tr>
<tr>
<td>nonparametric</td>
<td>1</td>
<td><strong>45.5 ↑ 0.4</strong></td>
<td>27.37 (0)</td>
<td><strong>37.2 ↑ 0.7</strong></td>
<td>27.37 (0)</td>
<td><strong>26.8 ↑ 1.1</strong></td>
<td>27.37 (0)</td>
<td><strong>21.2 ↑ 1.4</strong></td>
<td>27.37 (0)</td>
<td><strong>18.1 ↑ 1.6</strong></td>
<td>27.37 (0)</td>
</tr>
<tr>
<td>(Ours)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>parametric</td>
<td>10</td>
<td>45.7</td>
<td>28.56 (1.2)</td>
<td>37.0</td>
<td>29.66 (2.3)</td>
<td>26.6</td>
<td>31.26 (3.9)</td>
<td>20.8</td>
<td>32.86 (5.4)</td>
<td>17.7</td>
<td>33.96 (6.5)</td>
</tr>
<tr>
<td>nonparametric</td>
<td>10</td>
<td><strong>46.4 ↑ 0.7</strong></td>
<td>27.37 (0)</td>
<td><strong>37.8 ↑ 0.8</strong></td>
<td>27.37 (0)</td>
<td><strong>27.9 ↑ 1.3</strong></td>
<td>27.37 (0)</td>
<td><strong>22.1 ↑ 1.3</strong></td>
<td>27.37 (0)</td>
<td><strong>19.4 ↑ 1.7</strong></td>
<td>27.37 (0)</td>
</tr>
<tr>
<td>(Ours)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. **Scalability study** (.§5.3) of our nonparametric model against the parametric baseline (i.e., SegFormer [120]) on ADE20K [142]. For each model variant, we report its segmentation mIoU, parameter numbers of the entire model as well as the prototypes (in the bracket).
Ablation Study

Design

<table>
<thead>
<tr>
<th>$\mathcal{L}_{CE}$ (Eq. 7)</th>
<th>$\mathcal{L}_{PPC}$ (Eq. 11)</th>
<th>$\mathcal{L}_{PPD}$ (Eq. 12)</th>
<th>mIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>45.0</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>45.9</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>45.4</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>46.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th># Prototype</th>
<th>mIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K = 1$</td>
<td>45.5</td>
</tr>
<tr>
<td>$K = 5$</td>
<td>46.0</td>
</tr>
<tr>
<td>$K = 10$</td>
<td>46.4</td>
</tr>
<tr>
<td>$K = 20$</td>
<td>46.5</td>
</tr>
<tr>
<td>$K = 50$</td>
<td>46.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficient $\mu$</th>
<th>mIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu = 0$</td>
<td>44.9</td>
</tr>
<tr>
<td>$\mu = 0.9$</td>
<td>45.9</td>
</tr>
<tr>
<td>$\mu = 0.99$</td>
<td>46.0</td>
</tr>
<tr>
<td>$\mu = 0.999$</td>
<td>46.4</td>
</tr>
<tr>
<td>$\mu = 0.9999$</td>
<td>46.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distance Measure</th>
<th>mIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>45.7</td>
</tr>
<tr>
<td>Huberized</td>
<td>45.2</td>
</tr>
<tr>
<td>Cosine</td>
<td>46.4</td>
</tr>
</tbody>
</table>

(a) Training Objective $\mathcal{L}$
(b) Prototype Number $K$
(c) Momentum Coefficient $\mu$
(d) Distance Measure

Table 5. A set of ablative studies (§5.4) on ADE20K [142] val. All model variants use MiT-B2 [120] as the backbone.
Visualization

Segformer  Ours
Visualization

Segformer  Ours
Visualization

Segformer

Ours
Prototype Meaning

Figure 3. **Visualization of pixel-prototype similarity** for *person* (top) and *car* (bottom) classes. Please refer to §3 for details.
Embedding Space

Figure 5. **Embedding spaces** learned by (left) parametric model [120], and (right) our nonparametric model. For better visualization, we show five classes of Cityscapes [23] with two prototypes per class.