1. Pseudo-Labeling Schemes

How to generate pseudo-labels for unlabeled data based on the model outputs is also an important question for CMPL. Many pseudo-labeling strategies become possible with the introduction of the auxiliary network. Besides the one described in our method, we list some other options:

1) **Self-First**: Each network first checks whether its own prediction is confident enough, and if it is not, then the label is obtained from its sibling. 2) **Opposite-First**: Each network instead prioritizes its companion over itself. 3) **Maximum**: The most confident prediction from the two networks is taken as the pseudo-label. 4) **Average**: The predictions from the two networks are averaged before deriving the pseudo-label.

Let the pseudo-label confidence produced by $F$ and $A$ be $l_F$ and $l_A$. The pseudo-label confidences for a video $u_i$ are thus $l_F(u_i) = p_F^i$, $l_A(u_i) = p_A^i$. And now the corresponding mathematical formulations of different pseudo-labeling schemes are presented as following, where $u_i$ is removed for clarity.

1. **Self-First**:

   \[
   \begin{align*}
   l_F &= \mathbb{1}(\text{max}(p_F^i) \geq \tau)p_F + (1 - \mathbb{1}(\text{max}(p_F^i) \geq \tau))p_A^i, \\
   l_A &= \mathbb{1}(\text{max}(p_A^i) \geq \tau)p_A + (1 - \mathbb{1}(\text{max}(p_A^i) \geq \tau))p_F. 
   \end{align*}
   \]

2. **Opposite-First**:

   \[
   \begin{align*}
   l_F &= \mathbb{1}(\text{max}(p_A^i) \geq \tau)p_A + (1 - \mathbb{1}(\text{max}(p_A^i) \geq \tau))p_F, \\
   l_A &= \mathbb{1}(\text{max}(p_F^i) \geq \tau)p_F + (1 - \mathbb{1}(\text{max}(p_F^i) \geq \tau))p_A. 
   \end{align*}
   \]

3. **Maximum**:

   \[
   l_F = l_A = \mathbb{1}(\text{max}(p_F^i) \geq \text{max}(p_A^i))p_F + (1 - \mathbb{1}(\text{max}(p_F^i) \geq \text{max}(p_A^i)))p_A. 
   \]

4. **Average**:

   \[
   l_F = l_A = \frac{p_F + p_A}{2}. 
   \]

Table 1. Comparison across different pseudo-labeling schemes.

<table>
<thead>
<tr>
<th>Pseudo-Labeling</th>
<th>Top-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>FixMatch</td>
<td>6.78</td>
</tr>
<tr>
<td>Self-Confident</td>
<td>10.80</td>
</tr>
<tr>
<td>Opposite-Confident</td>
<td>11.13</td>
</tr>
<tr>
<td>Maximum</td>
<td>11.85</td>
</tr>
<tr>
<td>Average</td>
<td>12.16</td>
</tr>
<tr>
<td>Cross</td>
<td>12.90</td>
</tr>
</tbody>
</table>

Experimental Results. Tab. 1 presents the results of different pseudo-labeling schemes. The baseline strategy (FixMatch [8]) performs the worst. Due to the lack of the auxiliary networks, the unlabeled data mainly distinguishable beyond the representation of the primary backbone rarely gets paired with confident pseudo labels since the scores of those unseen videos are easily below the threshold. After introducing the temporal information derived from the auxiliary network, clear improvements are observed.

For the remaining strategies except the proposed cross-model scheme, there is a chance that the primary network will dominate the pseudo labeling decisions, leading to the wrong decisions for unlabeled samples. In contrast, for our cross-model strategy, each network always receives pseudo labels from its companion and never from itself, and this is shown to be more effective.

2. Comparison to Self-Supervised Methods

In this section, we compare CMPL with state-of-the-art self-supervised learning approaches. We use UCF-101 as the labeled data and Kinetics-400 as the unlabeled data. A described in the main paper, CMPL jointly use labeled and unlabeled data in a semi-supervised manner. As for self-supervised learning methods, we follow the standard
protocol to use unlabeled data in Kinetics-400 for pre-
training, followed by a fine-tuning on the labeled data in
UCF-101.

As shown in Tab. 2, in comparison to the CoCLR [4],
our model provides a performance gain of 1.3% only with
8 frames input, indicating the effectiveness of CMPL. It is
a very encouraging result, suggesting that semi-supervised
learning is a promising solution for action recognition with
limited labeled data. We hope that our result can provide
a strong baseline for comparison with more self-supervised
learning methods.

3. 3D-ResNet Network Structure

Tab. 3 shows the architecture of 3D-ResNet50. It inherits
the 2D-ResNet [5] and inflates the 2D kernel at conv1
cross all stages. The other convolution blocks are still in
2D format, focusing on the spatial semantics. Moreover,
there exist no temporal downsampling layers, in order to
maintain long-temporal fidelity. Notably, we shrink the
width of 3D-ResNet to a factor of 1/4 to use the 3D-
ResNet50 × 1/4 as the default auxiliary pathway.

4. Effects of Sampling Schemes

As illustrated in Section 4.1 of the main paper, the
number of sampled videos is the same across different
categories. However, different from UCF-101, the distribu-
tion of videos across different categories is not balanced
in Kinetics-400. We re-sample a new video subset under the
Kinetics-400 distribution, called ‘category-wise sampling
scheme’. To be specific, we first compute the number of
each category and next randomly sample the videos from
each category with the corresponding ratio and the total
number. Tab. 4 presents the results of different sampling
schemes under the same setting of ablation study in Section
4.3 of the main paper. Even with the unbalanced distribu-
tion, CMPL obtains nearly the same performance with the
‘uniform sampling’ scheme, suggesting the robustness and
generality of our approach.
Table 4. Study on sampling Scheme.

<table>
<thead>
<tr>
<th></th>
<th>Uniform (Default)</th>
<th>Category-Wise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-1</td>
<td>12.90</td>
<td>12.68</td>
</tr>
</tbody>
</table>

References


