In this paper, we propose a deep learning model based on deep multi-view representation learning, to address the video object segmentation task. The proposed model emphasizes the importance of the inherent correlation between video frames and incorporates a multi-view representation learning based on deep canonically correlated autoencoders. The multi-view representation learning in our model provides an efficient mechanism for jointly extracting useful features and learning better representation into a joint feature space, i.e., shared representation. To increase the training data and the learning capacity, we train the proposed model with pairs of video frames, i.e., \( F_i \) and \( F_j \).

During the segmentation phase, the deep canonically correlated autoencoders model encodes useful features by processing multiple reference frames together, which is used to detect the frequently reappearing. Our model enhances the state-of-the-art deep learning-based methods proposed model with pairs of video frames, i.e., \( F_i \) and \( F_j \). Experimental results over two large benchmarks demonstrate the ability of the proposed method to outperform competitive approaches and to reach good performances, in terms of semantic segmentation.

**Motivations and goals**

- **Motivations**
  - Existing models mainly focus on the intra-frame discrimination of primary objects in motion or appearance.
  - They ignore the valuable global-occurrence consistency across multiple video frames.
  - Recurrent neural networks (RNNs) fail to explore the rich relations, i.e., the high correlation between different video frames, hence do not attain a global perspective.

- **Goals**
  - Propose a video semantic segmentation model using deep multi-view representation learning to model video semantic segmentation task from a global view
  - Capture the rich inherent correlations between all frames
  - Improve the segmentation task

**Proposed methodology**

- **Multi-view deep representation learning**
  - Learn a better representation from pairs of frames, i.e, multimodal frames of a video by encoding their useful features in order to capture the inherent correlation between them

- **Multi-view representation learning**
  - Extract useful (relevant) features from multiple input modalities, i.e. pairs of video frames denoted by \( F_i \) and \( F_j \), which may be reconstructed.

**DecAE**

- **Encoder**: \( Enc_{AE} \)
- **Bottleneck layer**: \( z = Enc_{AE}(F_i) \)
- **Decoder**: \( Dec_{AE}(Enc_{AE}(F_i)) = F_i \)
- **Training (MSE)**:

\[
\frac{1}{n \times m} \sum_{i=1}^{n} \sum_{j=1}^{m} \| Dec_{AE}(Enc_{AE}(F_i)) - F_i \|_2^2
\]

**Multi-view representation learning**

- **Deep Canonical Correlation Analysis (DCCA)**

\[
\begin{align*}
\min_{u \in \mathbb{R}^{N_u}} & \sum_{i=1}^{N} \frac{1}{N_u} \left( (u_i^T (x_i - \mu_x))^2 + (u_i^T (y_i - \mu_y))^2 \right) \\
\text{s.t.} & \sum_{i=1}^{N} u_i = 0 \\
& u^T \Sigma_x u = 1 \quad \text{and} \quad u^T \Sigma_y u = 1
\end{align*}
\]

**Experimental results**

- **Data description**
  - **UAVid dataset**: consists of 30 video sequences. It is composed of 300 images and each of size \( 3840 \times 2160 \) or \( 4096 \times 2160 \). There are 8 classes that have been selected for semantic segmentation.
  - **DAVIS16 dataset**: which consists of 50 videos in total. We select 30 videos for training and 20 for testing.

**Evaluation on UAV dataset**

IoU scores for different deep learning models

<table>
<thead>
<tr>
<th>Model</th>
<th>Building</th>
<th>Tree</th>
<th>Car</th>
<th>Person</th>
<th>Moving car</th>
<th>Moving bus</th>
<th>Man</th>
<th>Woman</th>
<th>Caravan</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN</td>
<td>40</td>
<td>29.5</td>
<td>70.4</td>
<td>51.7</td>
<td>6.4</td>
<td>21.0</td>
<td>0.0</td>
<td>79.3</td>
<td>0.0</td>
</tr>
<tr>
<td>DCRN</td>
<td>79.5</td>
<td>90.8</td>
<td>90.8</td>
<td>83.2</td>
<td>11.2</td>
<td>47.5</td>
<td>10.0</td>
<td>40.9</td>
<td>10.0</td>
</tr>
<tr>
<td>DCRN+</td>
<td>79.8</td>
<td>90.9</td>
<td>90.7</td>
<td>83.0</td>
<td>11.1</td>
<td>47.3</td>
<td>10.0</td>
<td>40.8</td>
<td>10.0</td>
</tr>
<tr>
<td>Ours</td>
<td>76.1</td>
<td>71.3</td>
<td>43.2</td>
<td>65.5</td>
<td>12.1</td>
<td>45.8</td>
<td>10.0</td>
<td>45.8</td>
<td>10.0</td>
</tr>
</tbody>
</table>

**Evaluation on DAVIS16 dataset**

Results on the test set

- **Region similarity** \( \mathcal{J} \)
- **Boundary accuracy** \( \mathcal{F} \)
- **time stability** \( \tau \)

**Abstract**

In this paper, we propose a deep learning model based on deep multi-view representation learning, to address the video object segmentation task. The proposed model emphasizes the importance of the inherent correlation between video frames and incorporates a multi-view representation learning based on deep canonically correlated autoencoders. The multi-view representation learning in our model provides an efficient mechanism for jointly extracting useful features and learning better representation into a joint feature space, i.e., shared representation. To increase the training data and the learning capacity, we train the proposed model with pairs of video frames, i.e., \( F_i \) and \( F_j \). During the segmentation phase, the deep canonically correlated autoencoders model encodes useful features by processing multiple reference frames together, which is used to detect the frequently reappearing. Our model enhances the state-of-the-art deep learning-based methods proposed model with pairs of video frames, i.e., \( F_i \) and \( F_j \). Experimental results over two large benchmarks demonstrate the ability of the proposed method to outperform competitive approaches and to reach good performances, in terms of semantic segmentation.