Reliable motion representation, such as optical flow, has proven to have great promotion in action recognition task. However, present methods have their drawbacks.

- Pre-processed methods, including TV-L1 optical flow and PWC-Net optical flow, are time-consuming and require a large amount of storage space. While the Pwc-Net method has domain gap for action datasets.
- Embedded methods, like PCL-Net, ActionFlowNet, Hidden two-stream, TV-Net, etc, are either inaccurate or inefficient.

We aim to design a lightweight and effective temporal action recognition framework with unsupervised optical flow estimation embedded.

---

**Model and Approach**

Our UF-TSN consists of two forward stream, spatial stream and temporal stream, whose outputs are fusion to produce the final prediction, which is displayed in the left image below.

- To obtain reliable optical flow for action datasets, we first apply a shallow network for feature extraction.
- Then the motion is estimated in a coarse-to-fine manner, as shown in the right image above.
- For each level, the coarse flow is predicted based on cost volume, which is the correlation of two feature maps.
- Photometric loss and smoothness loss to lead unsupervised motion training.
- Photometric losses for unsupervised training [1].

\[
L_{\text{diff}} = \sum_{i,j} \left\| I_1^i(i,j) - I_2^j(i,j) \right\|_1 \quad L_{\text{census}} = \sum_{i,j} d(c_1^i(i,j), c_2^j(i,j))
\]

- Smoothness loss for clear bound and smooth non-boundary area.

\[
L_{\text{smooth}} = \sum_{d=1}^{2} \sum_{i,j} \left\| \nabla_d u_1^i(i,j) \right\|_2 e^{-\left\| \nabla_d l^i(i,j) \right\|_2} \quad L = \frac{1}{M} (\lambda_1 L_{\text{diff}} + \lambda_2 L_{\text{census}} + \lambda_3 L_{\text{smooth}})
\]

- The overall loss for unsupervised optical flow is the combination of the three terms (above right).

---

**Evaluation**

- Implementation: ResNet-152 for spatial stream, and ResNet-18 for spatial feature extraction in temporal stream, both ResNet pre-trained on ImageNet. Flow estimator initialized randomly. For temporal stream, we first train the flow estimator, then the classifier, finally fine-tune the network end-to-end.
- We evaluate our model on UCF-101 and HMDB-51, the test procedure follows CoViAR [2] and DMCNet [3]. The comparison with other models is displayed in the left table, the right two tables are ablation study of the unsupervised flow estimator, cost volume part and multi-scale network design, and the efficiency comparison of various similar methods.

---

**Visualization**

- Confusion matrix of part categories from UCF-101. From left to right: RGB stream, UF-TSN, promotion with our methods.
- Visualization of optical flow from both UCF-101 and HMDB-51. From left to right: original RGB image, TV-L1, PWC-Net and UF-TSN.

---

**Reference**