Shanghai Jiao Tong University

UNSUPERVISED MOTION REPRESENTATION ENHANCED NETWORK FOR ACTION RECOGNITION

Xiaohang Yang, Lingtong Kong and Jie Yang

Paper #4982

2021 TORONTO Canada
June 6-11, 2021
Metro Toronto Convention Centre

Motivation

Reliable motion representation, such as optical flow, has proven to have great promotion in action recognition task.

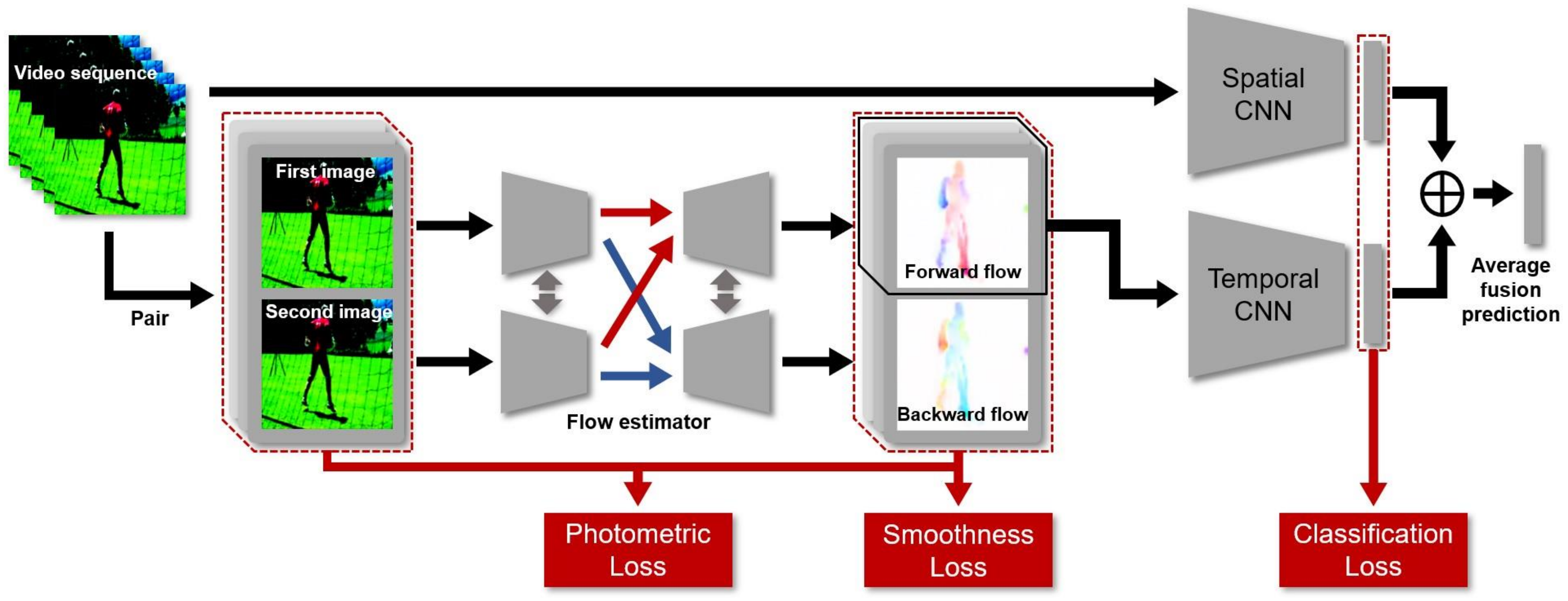
However, present methods have their drawbacks.

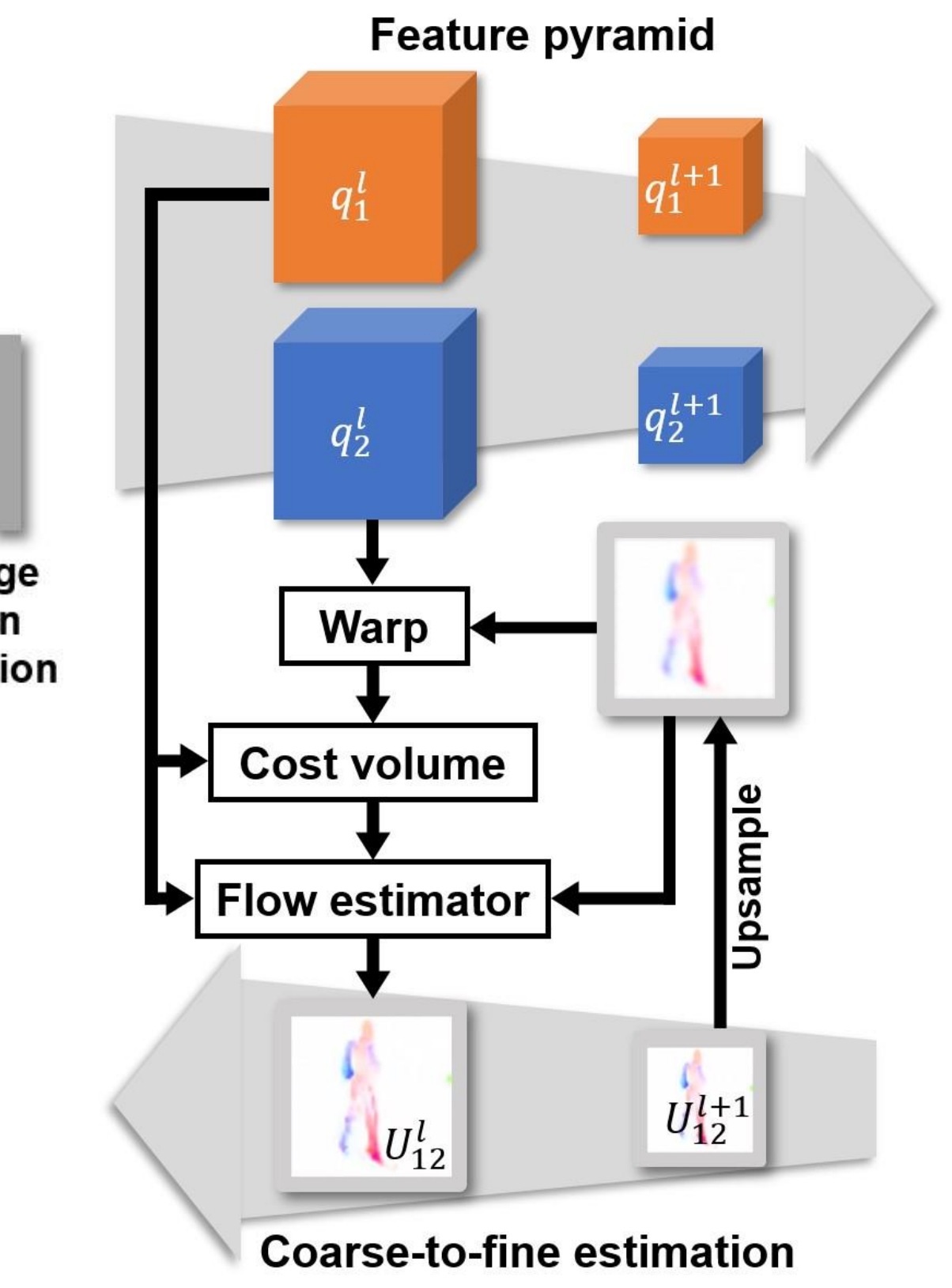
- Pre-processed methods, including TV-L1 optical flow and PWC-Net optical flow, are time-consuming and require a large amount of storage space. While the Pwc-Net method has domain gap for action datasets.
- Embedded methods, like PCL-Net, ActionFlowNet, Hidden two-stream, TV-Net, etc, are either inaccurate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

$$\hat{q}_2^l(\mathrm{x}) = q_2^l(\mathrm{x} + \mathrm{up}(U_{12}^{l+1})(\mathrm{x})) \qquad \mathrm{CostVolume}^l(\mathrm{x}_1, \mathrm{x}_2) = \frac{1}{N} q_1^l(\mathrm{x}_1)^{\mathsf{T}} \hat{q}_2^l(\mathrm{x}_2)$$

- We use photometric losses and smoothness loss to lead unsupervised motion training.
  - Photometric losses for unsupervised training [1].

$$\mathcal{L}_{\text{diff}} = \sum_{i,j}^{M} \left\| I_1^l(i,j) - \hat{I}_2^l(i,j) \right\|_1 \qquad \mathcal{L}_{\text{census}} = \sum_{i,j}^{M} d(c_1^l(i,j), c_2^l(i,j))$$

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

$$\mathcal{L}_{\text{smooth}} = \sum_{d=x,y} \sum_{i,j}^{M} \|\nabla_d U_{12}^l\|_1 \, e^{-(\|\nabla_d I_1^l\|_1)} \qquad \mathcal{L} = \frac{1}{M} (\lambda_1 \mathcal{L}_{\text{diff}} + \lambda_2 \mathcal{L}_{\text{census}} + \lambda_3 \mathcal{L}_{\text{smooth}})$$

  - 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 DMC-Net [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.

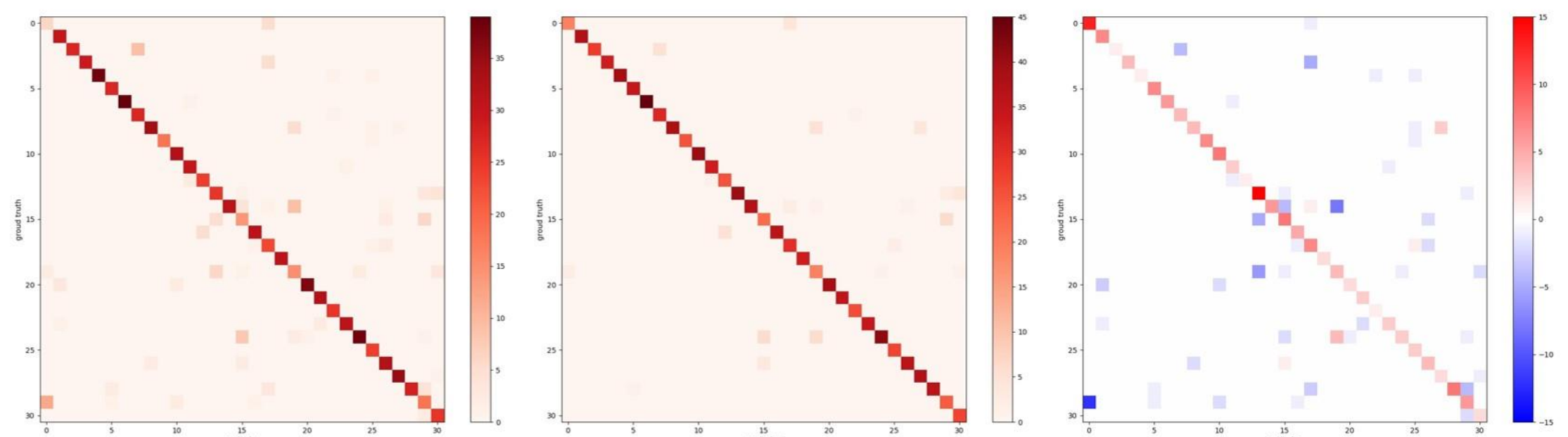
Methods	UCF101	HMDB51
Compressed Video		
CoViAR	90.4	59.1
DMC-Net	90.9	62.8
RGB only		
PCLNet	82.8	53.5
ActionFlowNet	83.9	56.4
Hidden two-stream (VGG16)	90.3	60.5
TV-Net	94.5	71.0
Our UF-TSN with ResNet-18	92.2	64.4
RGB + Optical Flow		
Two-stream	88.0	59.4
Two-stream Fusion	92.5	65.4

Methods	UCF101	HMDB51
TV-L1	92.2	64.5
PWC-Net	91.3	62.8
w/o cost-volume & multi-scale	90.4	60.9
w/o multi-scale	91.2	62.5

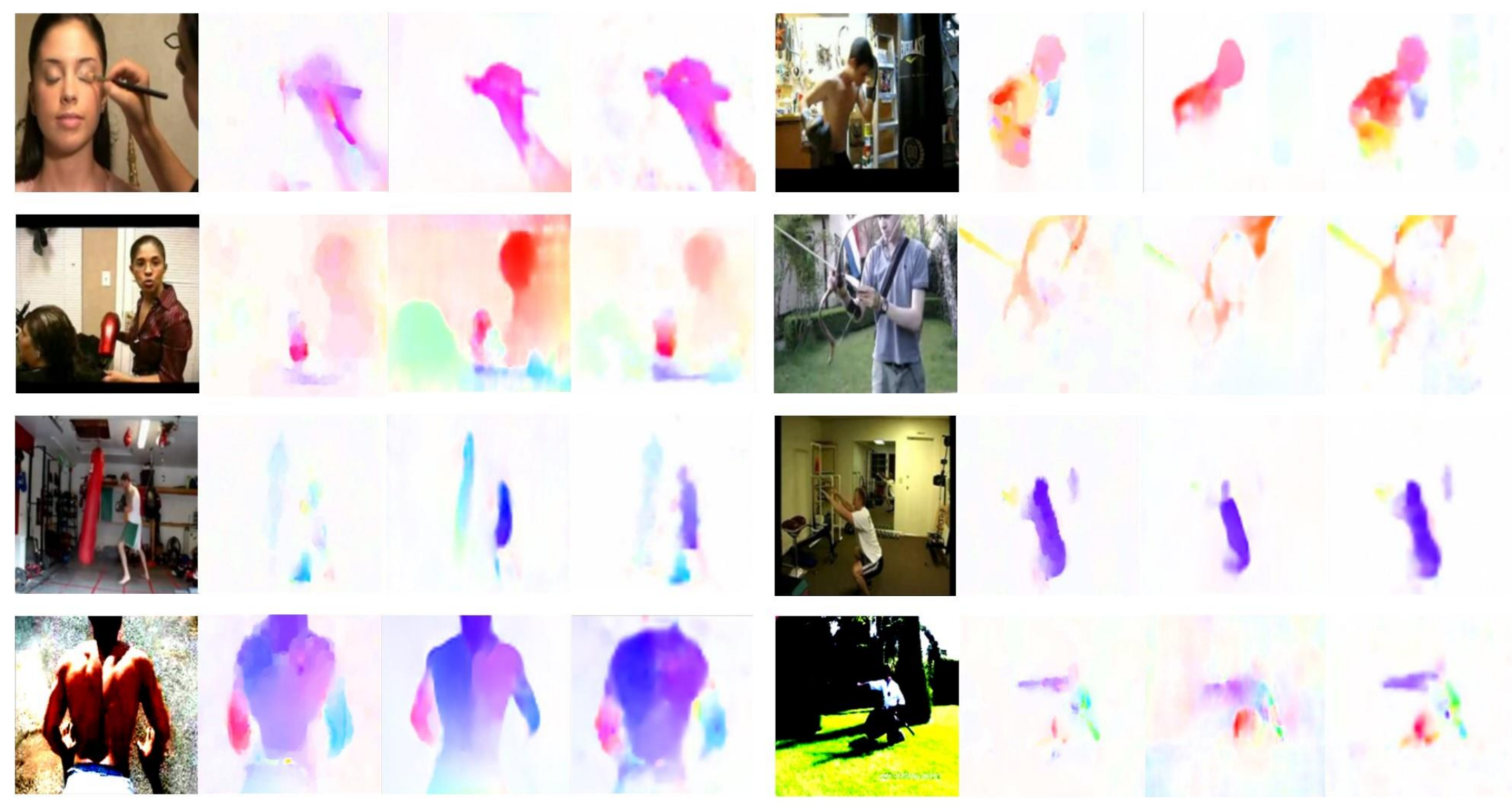
	Params (M)	FPS
PCLNet	11.9	19.2
ActionFlowNet	-	200
Hidden Two-stream	-	48.5
TV-Net	0.1	12.0
UF-TSN motion stream	5.7	131.6

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.



Contribution

We propose UF-TSN, a novel unsupervised motion representation learning framework for action recognition. UF-TSN applies feature pyramid and warping operation to reduce large displacement and estimates flow based on cost volume from coarse to fine, and then we constrain the prediction with image reconstruction and edge-aware smoothness losses in a multi-scale manner. Classification accuracy and visual instances of UF-TSN on two benchmark datasets have quantitatively and qualitatively demonstrated the competitive performance with TV-L1, which maintains the efficiency at the same time.

Reference

[1] Simon Meister, Junhwa Hur, and Stefan Roth, "Unflow: Unsupervised learning of optical flow with a bidirectional census loss," arXiv preprint arXiv:1711.07837, 2017.
[2] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R Manmatha, Alexander J Smola, and Philipp Kr¨ahenb¨uhl, "Compressed video action recognition," in CVPR, 2018.
[3] Zheng Shou, Xudong Lin, Yannis Kalantidis, Laura Sevilla-Lara, Marcus Rohrbach, Shih-Fu Chang, and Zhicheng Yan, "Dmc-net: Generating discriminative motion cues for fast compressed video action recognition," in CVPR, 2019.