# Towards Practical Compressed Video Action Recognition: A Temporal Enhanced Multi-Stream Network

Bing Li[1], Longteng Kong[1], Dongming Zhang[2], Xiuguo Bao[2], Di Huang[1] and Yunhong Wang[1]

[1]IRIP Lab, School of Computer Science and Engineering, Beihang University, Beijing 100191, China.
[2]National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China.

# Outline

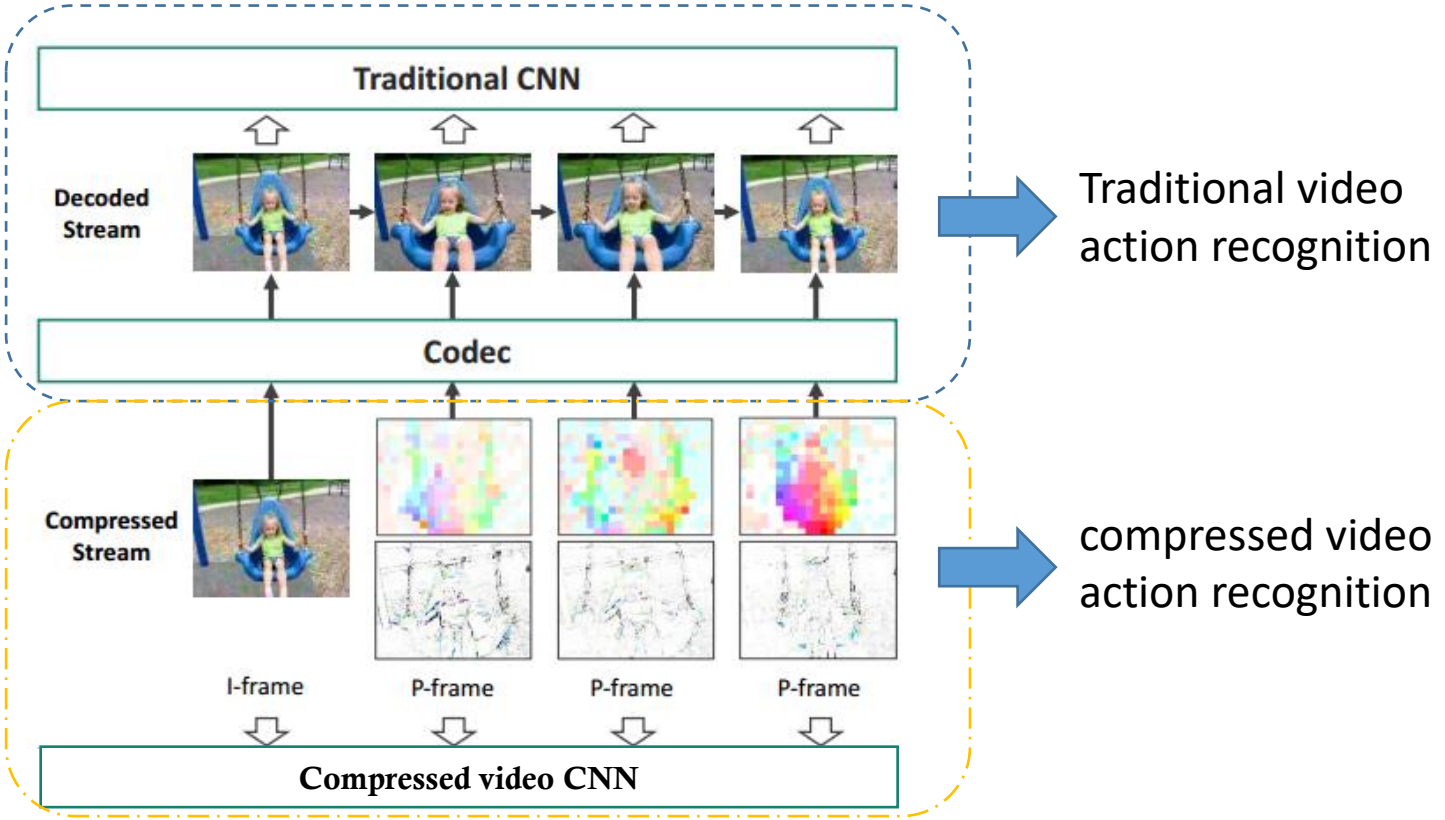
- Introduction
- Related Work
- The Proposed Approach
- Experimental Results
- Conclusions

# Outline

- Introduction
- Related Work
- The Proposed Approach
- Experimental Results
- Conclusions

# Introduction

Recently, many studies attempt to recognize actions in compressed videos rather than regular ones, aiming to avoid the resource overhead of decoding.




Structure comparison between compressed video action recognition and traditional video action recognition.

# Outline

- Introduction
- Related Work
- The Proposed Approach
- Experimental Results
- Conclusions

# Popular methods

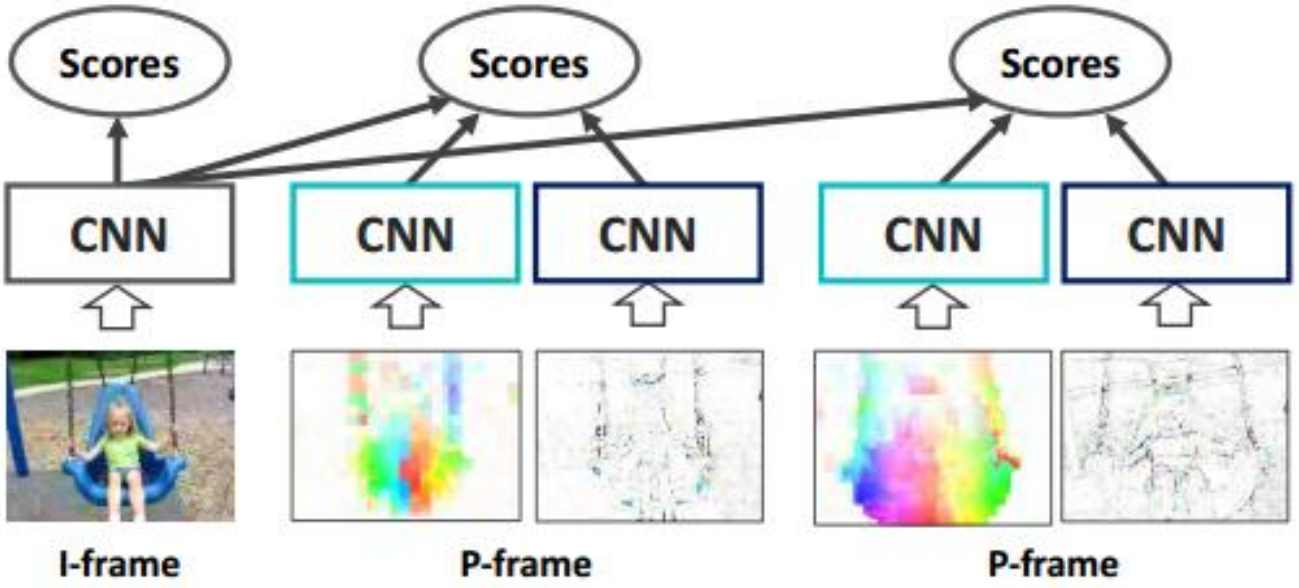
- EMV-CNN(2016)
  - Optical flow branch is replaced by motion vector
- CoViAR (2018)
  - Multi-Stream CNN
- DMC-Net (2019)
  - Using GAN to generate more refined motion vectors
- TTP (2019)
  - Temporal Trilinear Pooling strategy
- Others
  - Image denoising techniques, etc.

# Methods outlines


Overview of EMV-CNN

B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang, “Real-time action recognition with enhanced motion vector cnns,” in CVPR, 2016

# Methods outlines




Overview of CoViAR

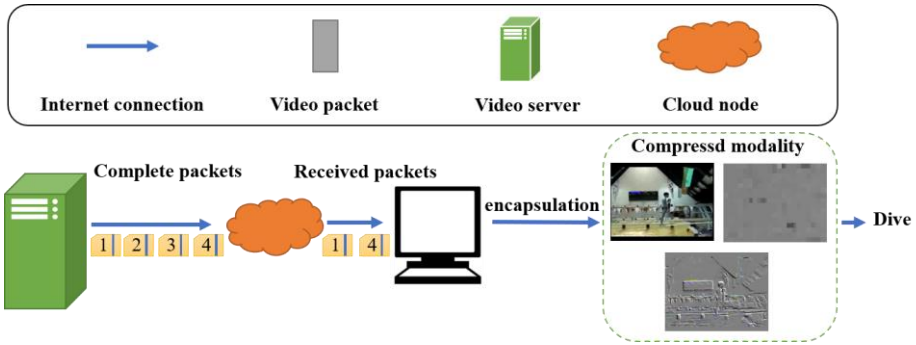
C. Wu, M. Zaheer, H. Hu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, “Compressed video action recognition,” in *CVPR*, 2018

# Outline

- Introduction
- Related Work
- The Proposed Approach
- Experimental Results
- Conclusions

# Motivation

- In the inference stage, the existing methods generally assume that all the observations of samples are available.
- In practical transmission, the compressed video packets are usually disorderly received and lost due to network jitters or congestions.
- In this work, we concentrate on practical compressed video action recognition, and consider to complement the missing video packets with partial received ones.

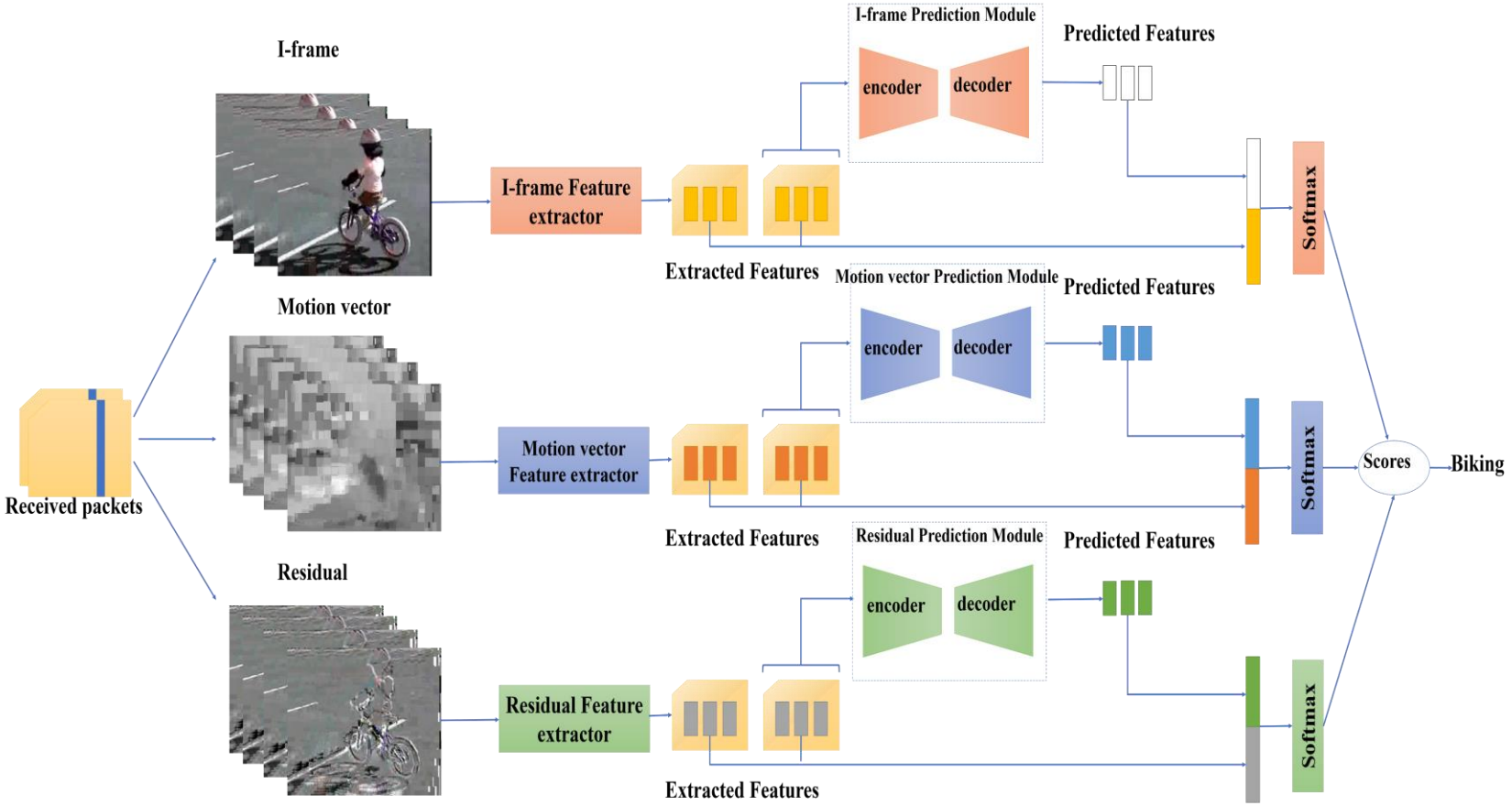



Compressed video action recognition in practical scenarios. Number 2, 3 packets are lost during video transmission.

# Contribution

- We proposed a Temporal Enhanced Multi-Stream Network (TEMSN) for practical compressed video action recognition.
- We use three modalities in compressed domain as complementary cues to capture richer information from compressed video packets.
- We design a temporal enhanced module based on Encoder-Decoder structure to generate more complete action dynamics.
- The proposed approach is evaluated on the HMDB-51 and UCF-101 datasets and state-of-the-art results are reached.

# Methodology




Framework of the proposed Temporal Enhanced Multi-Stream Network (TEMSN).

北京航空航天大學 BEIHANG UNIVERSITY

# Methodology

Given limited compressed video packets, TEMSN takes three phases to make action recognition.

- Multi-modal Encapsulation:
- We first decode the compressed video into I-frame (intra-coded frame), P-frame (predictive frame).
- We exploit the relation between the I-frame and P-frame to decouple the input, as Eq. 1 and Eq. 2, resulting residual, motion vector, and intra-frame.

$$I_i^{(t)} = I_{i\,-v^{(t)}}^{(t-1)} + \Delta^{(t)} \quad (1)$$

$$I_i^{(t)} = I_{i\,-D_i^{(t)}}^{(0)} + R_i^{(t)} \quad (2)$$

北京航空航天大學 BEIHANG UNIVERSITY

# Methodology

- Feature Extraction:

We then transform the modalities into the feature spaces by the Multi-Stream Network which consists of three independent CNNs.

- Temporal Enhancement:

- The temporal enhanced module takes the packet features as input, and predicts the contiguous packets as Eq. 3.
- The original and synthesized features are concatenated to form global representation for final recognition.

$$
\begin{aligned}
f_{pre} &= \varphi(s_t) \\
f_{RGB}^{(p_t)} &= \emptyset_{RGB}(f_{RGB}^{(p_{t-1})}) \\
f_{M\_V}^{(p_t)} &= \emptyset_{M\_V}(f_{M\_V}^{(p_{t-1})}) \\
f_{Res}^{(p_t)} &= \emptyset_{Res}(f_{Res}^{(p_{t-1})})
\end{aligned}
\tag{3}
$$

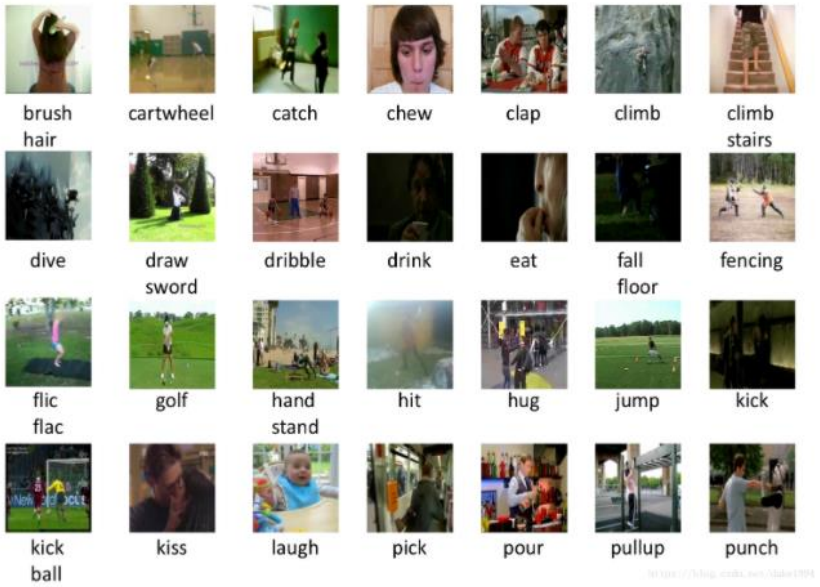
# Outline

- Introduction
- Related Work
- The Proposed Approach
- Experimental Results
- Conclusions

# Datasets

- **HMDB-51:** contains 6,766 trimmed videos from 51 action categories and provides 3 training/testing splits. Each training/test split has 3,570 training clips and 1,530 testing clips.

- **UCF-101:** contains 13,320 trimmed videos from 101 action categories. 3 training/testing splits are offered, each of which has approximately 9,600 clips for training and 3,700 clips for testing.




Example video class in HMDB-51

# Train

- Gradient descent: ADAM
- Hidden units: 2048 for I-frame and 512 for motion vector and residual
- Initial learning rate: 1e-4 for I-frame and 3e-4 for motion vector and residual
- Loss: cross-entropy loss for feature extraction network and L2 loss for Temporal Enhanced module

# Results

Table 1 accuracy averaged over three splits on HMDB-51 and UCF-101 for state-of-the-art compressed video-based methods

| Methods | HMDB-51 | UCF-101 | Ratio |
|---|---|---|---|
| EMV-CNN [10] | 51.2 | 86.4 | 100% |
| DTMV-CNN [11] | 55.3 | 87.5 | 100% |
| TTP [16] | 58.2 | 87.2 | 100% |
| CoViAR [28] | 59.1 | 90.4 | 100% |
| CoViAR‡ | 59.4 | 90.7 | 100% |
| CoViAR‡ | 57.3 | 88.5 | 50% |
| TEMSN (ours) | **61.1** | **91.8** | **100%** |
| TEMSN (ours) | **59.1** | **90.3** | **50%** |

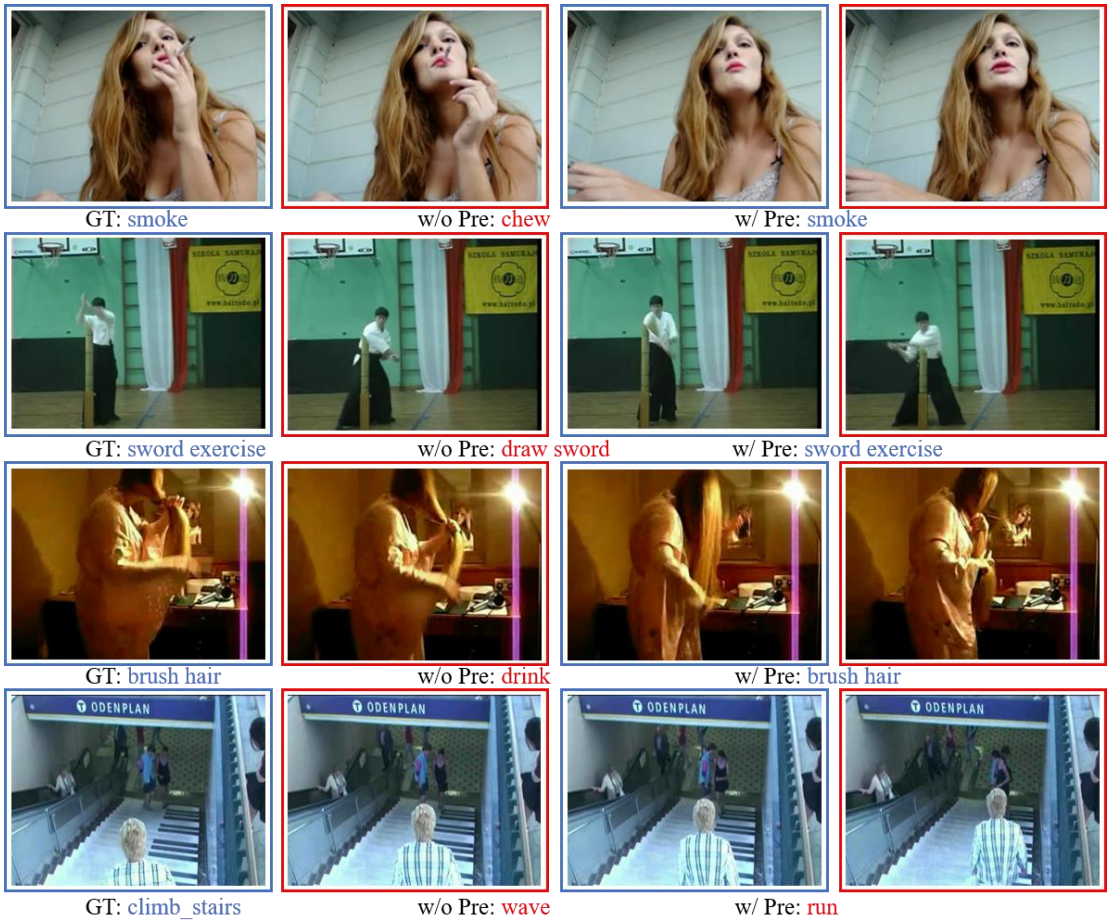
The proposed TEMSN shows superior performance compared to CoViAR and state-of-the-art, indicating the ability in dealing with practical conditions.

Table 2 results of three video packet prediction schemes on HMDB-51

| Baselines | Random | Uniform | Normal |
|---|---|---|---|
| w/o Pre. | 57.3 | 57.3 | 57.3 |
| +Pre. (20%) | 57.8 | 58.2 | 58.0 |
| +Pre. (50%) | 58.7 | 59.1 | 58.9 |
| +Pre. (100%) | 61.1 | 61.1 | 61.1 |

Longer temporal dynamics lead to better results and predicting all of the lost packets achieves the best performance.

# Results




Qualitative results of the proposed TEMSN on HMDB51, where each row belongs to an action. Frames in the blue box indicate the received ones, and in the red box indicate the lost ones.

北京航空航天大學 BEIHANG UNIVERSITY

# Outline

- Introduction
- Related Work
- The Proposed Approach
- Experimental Results
- Conclusions

# Conclusions

- This paper presents a Temporal Enhanced Multi-Stream Network towards practical compressed video action recognition.
- To obtain rich features from the compressed domain, we make use of three modalities to build a multi-stream network for action modeling.
- We further design a temporal enhanced module which is inserted into each stream to capture more complete motion dynamics.
- The experiments are conducted on the HMDB51 and UCF101 databases, and the results show the effectiveness of our proposed approach.

# References

[1].K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Advances on Neural Information Processing Systems, 2014, pp. 568–576.
[2].D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.
[3].C. Zach, T. Pock, and H. Bischof, “A duality based approach for realtime tv-L1 optical flow,” in Pattern Recognition, vol. 4713, 2007, pp. 214–223.
[4].B. Zhou, A. Andonian, A. Oliva, and A. Torralba, “Temporal relational reasoning in videos,” in European Conference on Computer Vision, vol.11205, 2018, pp. 831–846.
[5].C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” in IEEE Computer Vision and Pattern Recognition, 2016, pp. 1933–1941.
[6].R. D. Geest and T. Tuytelaars, “Modeling temporal structure with LSTM for online action detection,” in IEEE Conference on Applications of Computer Vision, 2018, pp. 1549–1557.


Thank You!