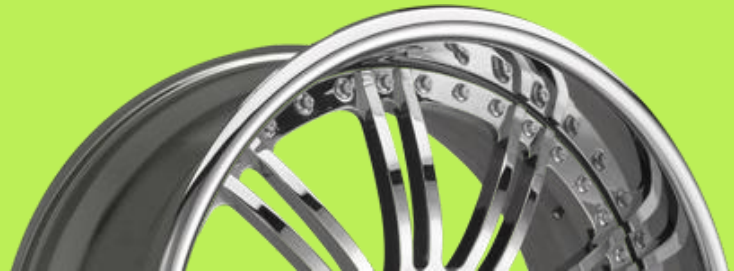
Multimodal Gesture Recognition Based on the ResC3D Network

Qiguang Miao Yunan Li Wanli Ouyang Zhenxin Ma Xin Xu Weikang Shi



CONTENTS

CONTENTS

INTRODUCTION

ChaLearn LAP IsoGD

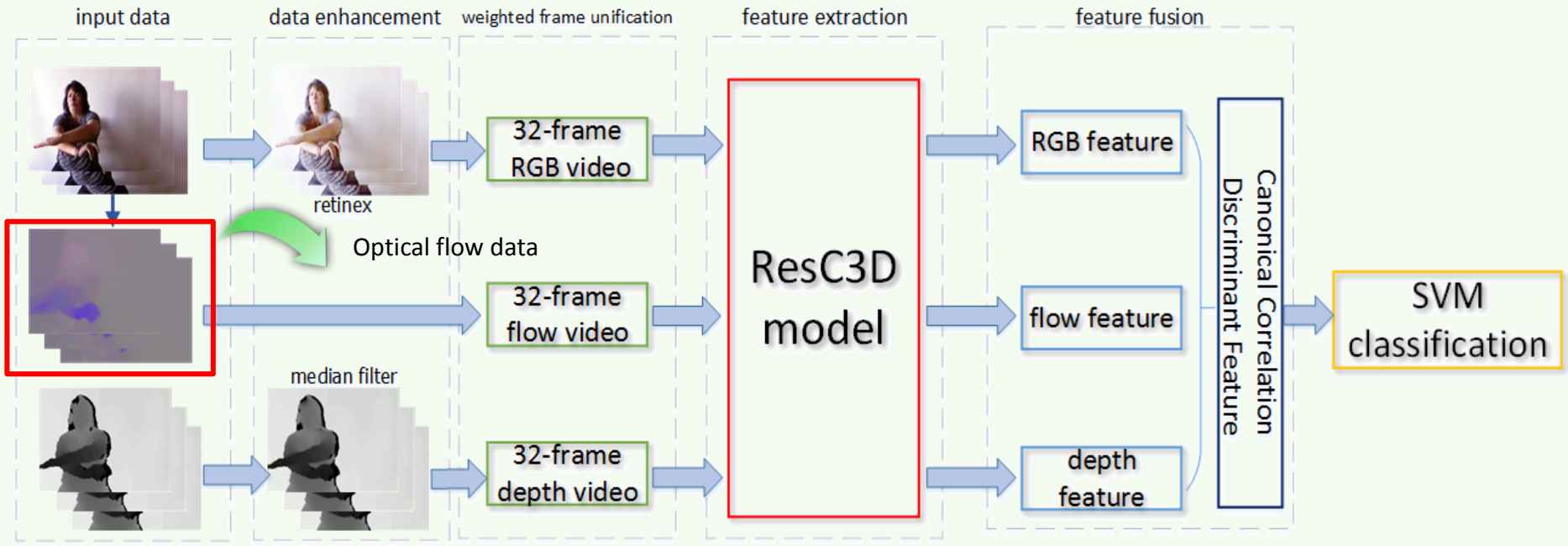
- large-scale
- video-based

C3D model

- 3D ConvNets
- spatiotemporal feature learning
- Auto feature extraction

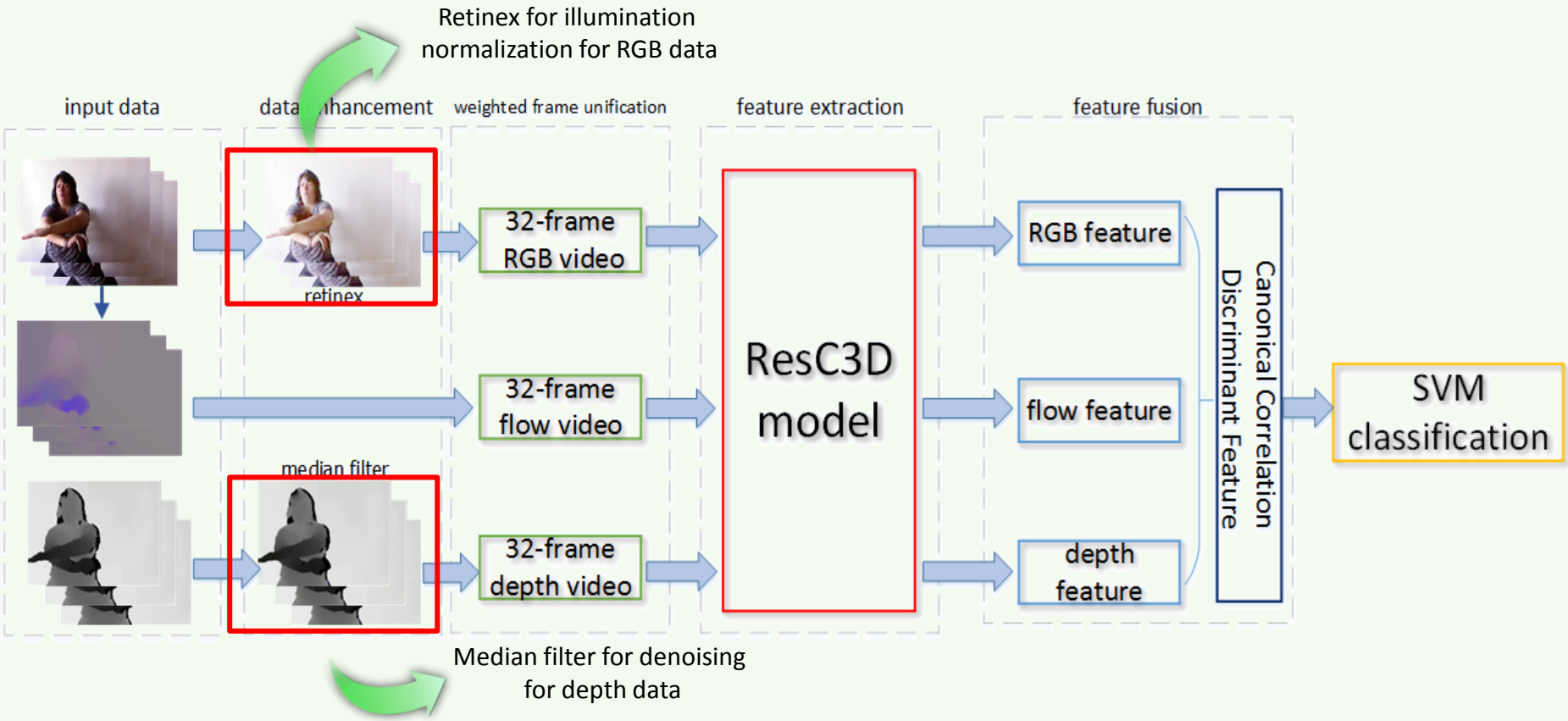
CONTENTS

Our Scheme



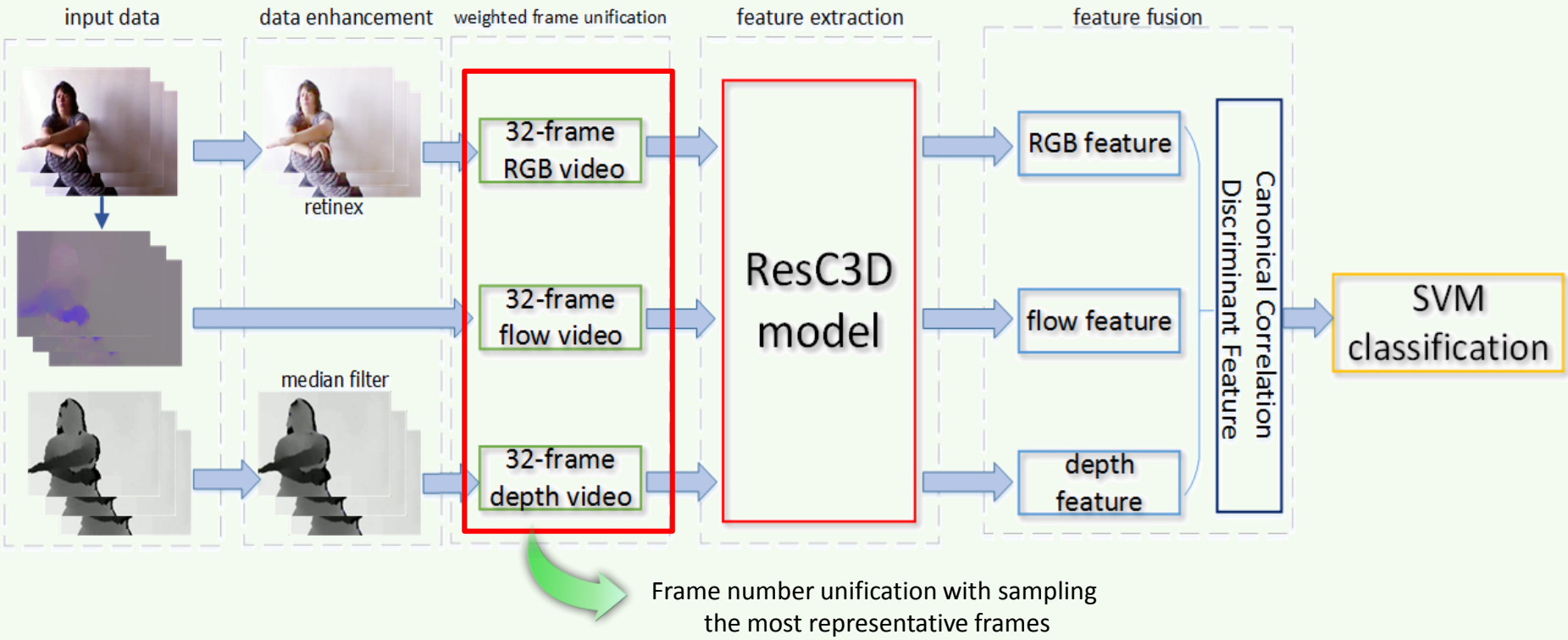
- ✓ Generating optical flow data from the RGB one

Our Scheme



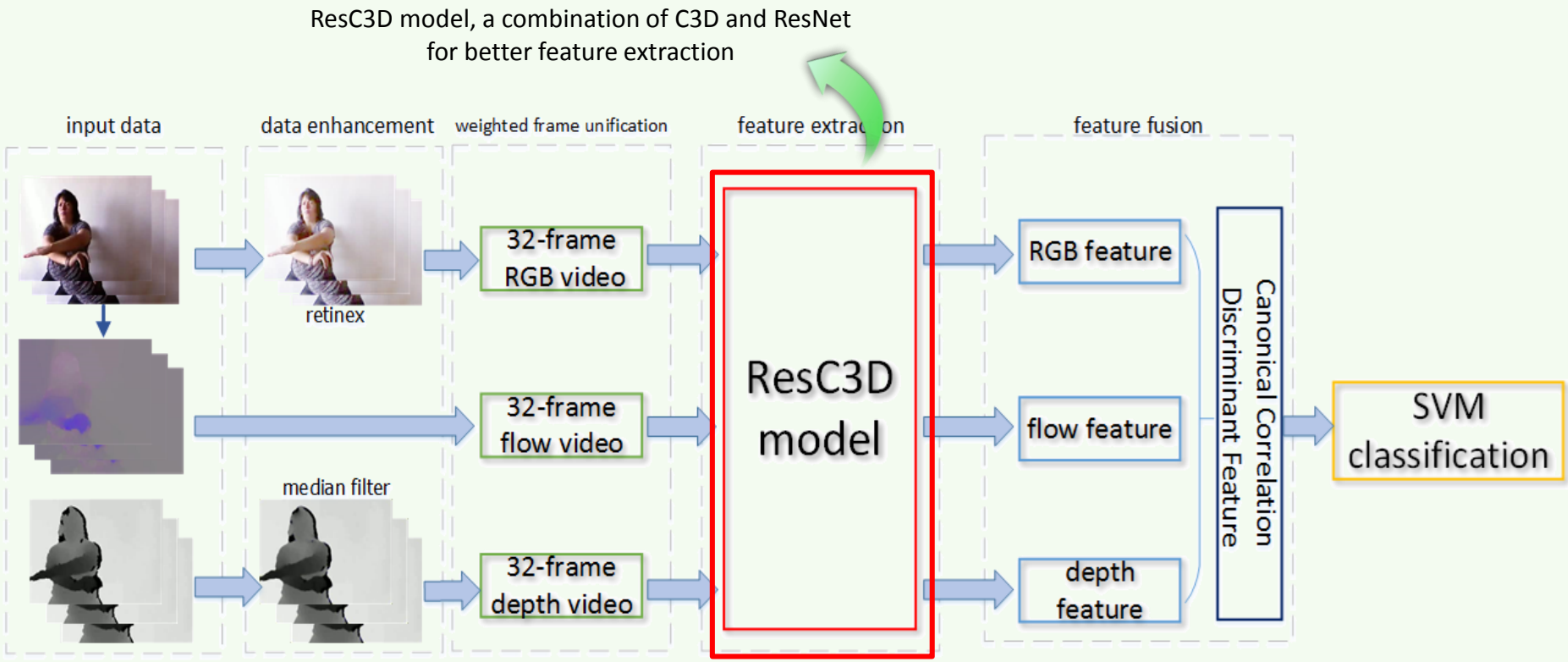
- ✓ Generating optical flow data from the RGB one
- ✓ Different strategies for video enhancement

Our Scheme



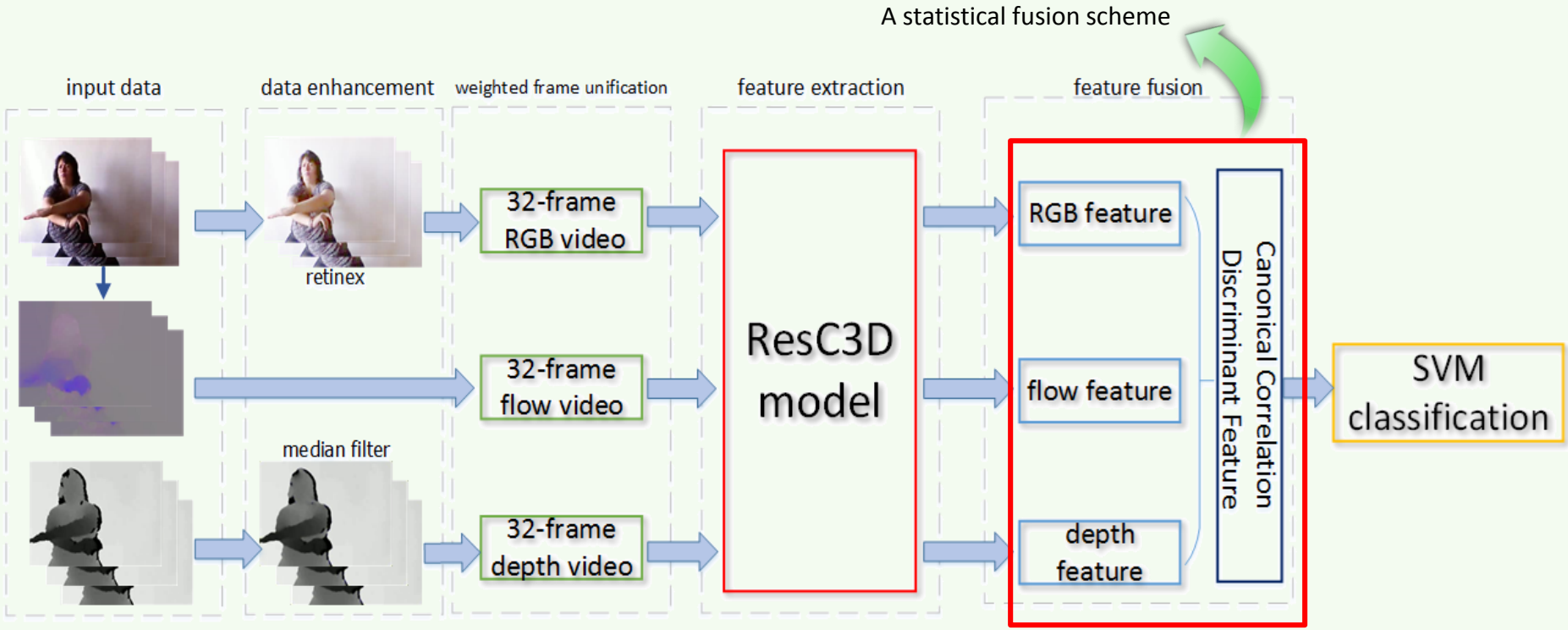
- ✓ Generating optical flow data from the RGB one
- ✓ Different strategies for video enhancement
- ✓ A weighted frame number unification strategy to sample the most representative frames

Our Scheme



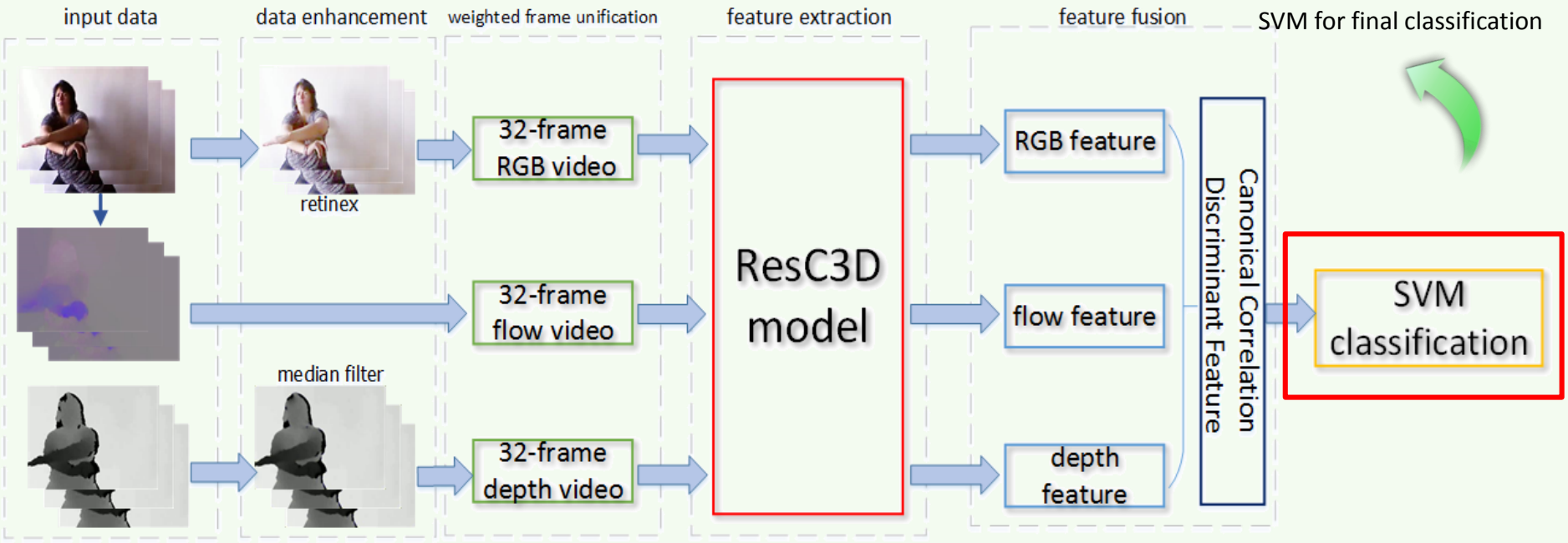
- ✓ Generating optical flow data from the RGB one
- ✓ Different strategies for video enhancement
- ✓ A weighted frame number unification strategy to sample the most representative frames
- ✓ A ResC3D model for feature extraction

Our Scheme



- ✓ Generating optical flow data from the RGB one
- ✓ Different strategies for video enhancement
- ✓ A weighted frame number unification strategy to sample the most representative frames
- ✓ A ResC3D model for feature extraction
- ✓ Using Canonical Correlation Analysis for feature fusion

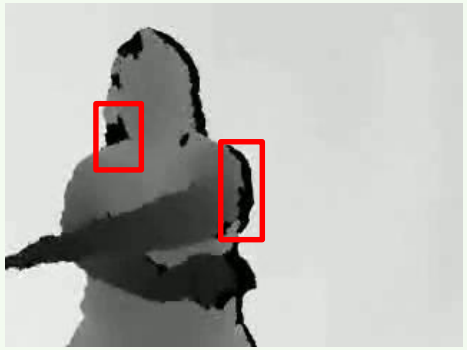
Our Scheme



- ✓ Generating optical flow data from the RGB one
- ✓ Different strategies for video enhancement
- ✓ A weighted frame number unification strategy to sample the most representative frames
- ✓ A ResC3D model for feature extraction
- ✓ Using Canonical Correlation Analysis for feature fusion
- ✓ SVM classifier for the final score

A. Data enhancement

RGB data
Suffering from different illumination condition



depth data
The noise exists around the edges

A. Data enhancment

- The results of enhancement with Retinex

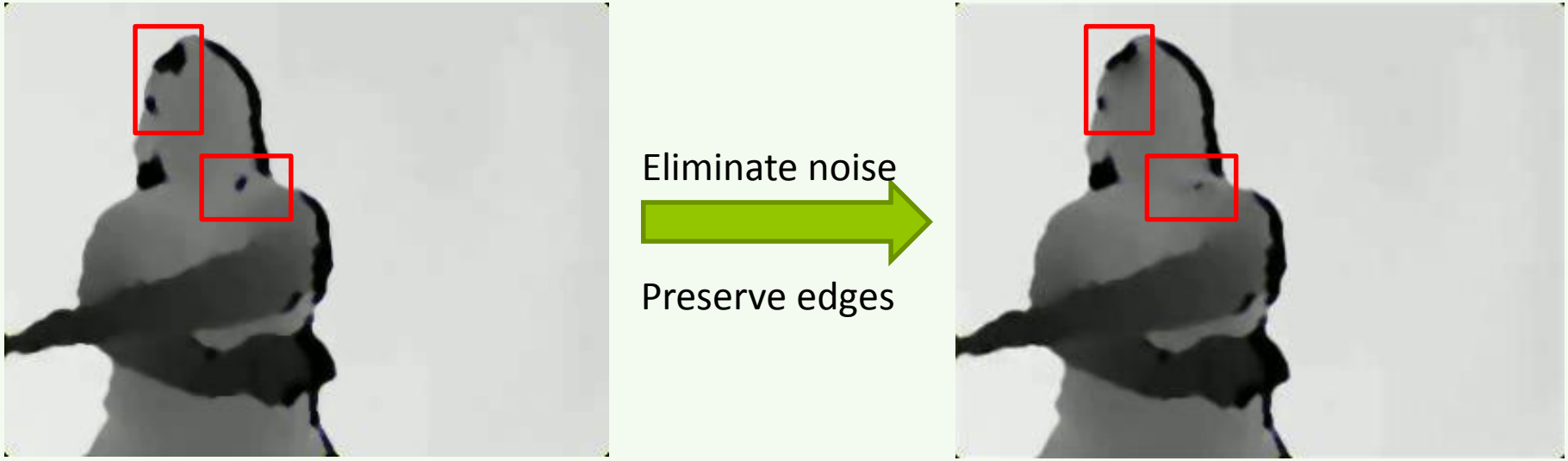




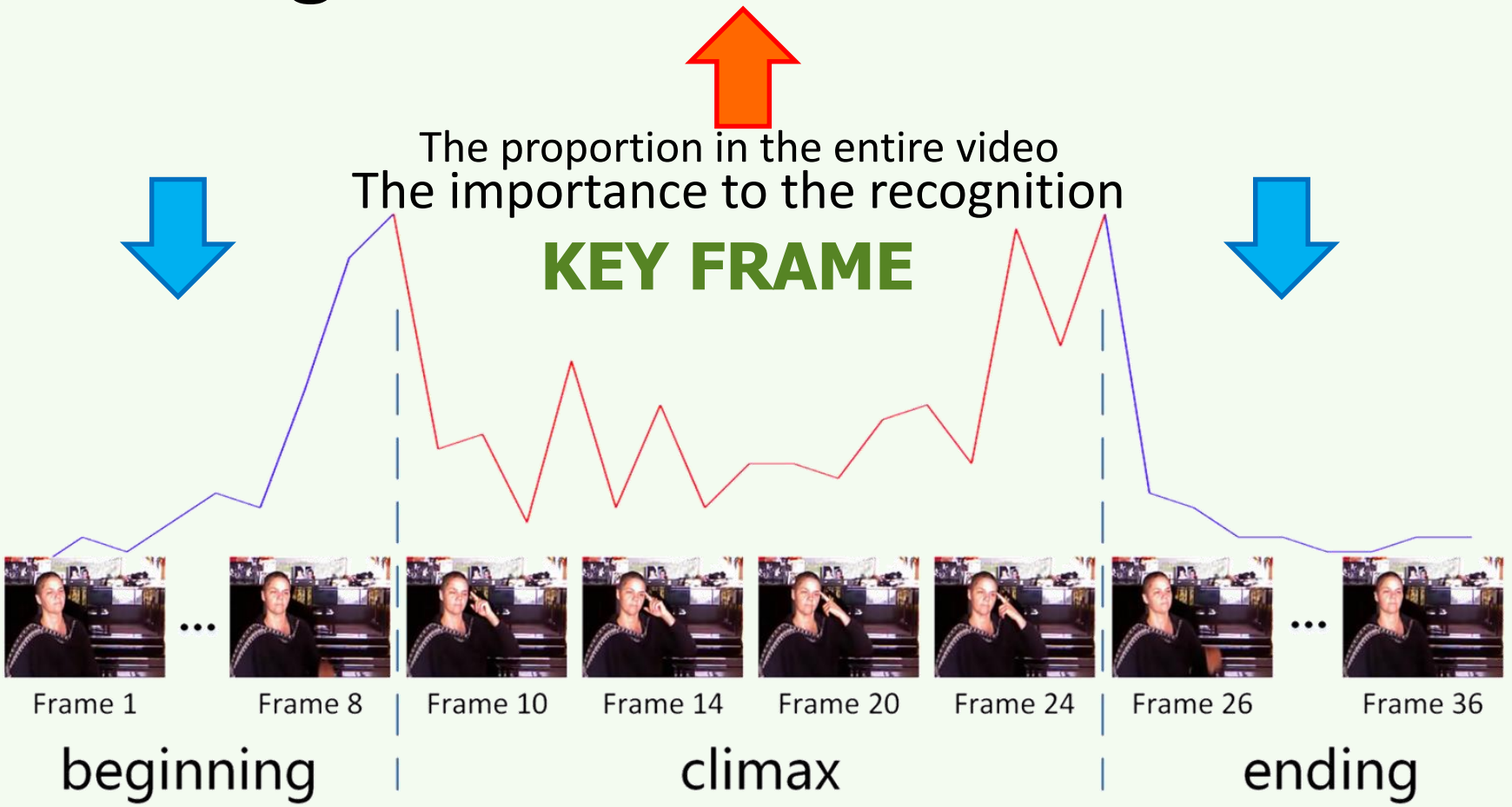



A. Data enhancment

- Denoising with median filter



B. Weighted frame unification

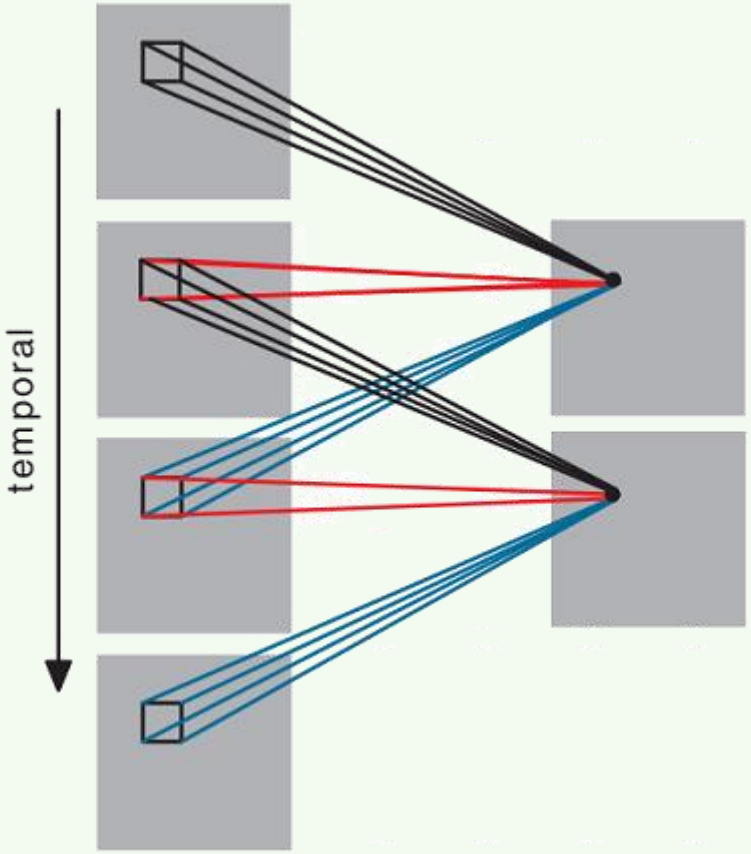


B. Weighted frame unification

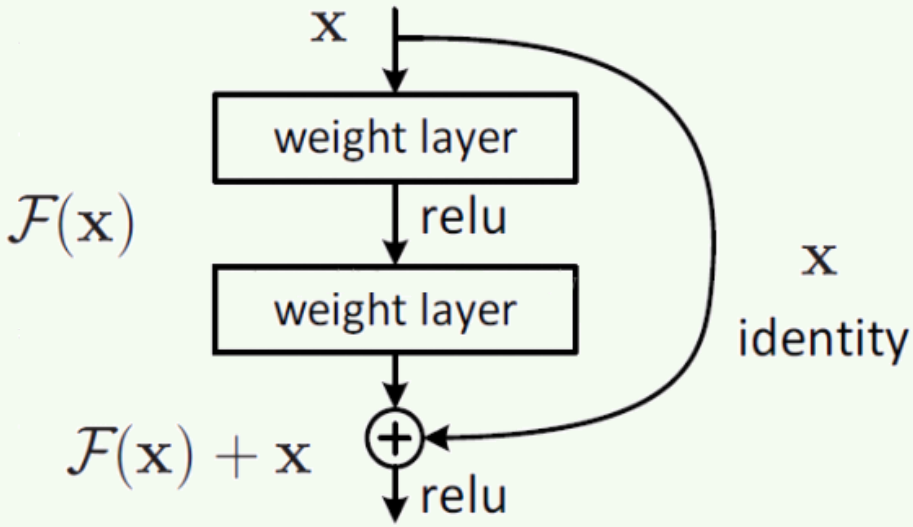
- Key frame
  - Divide the video into *n* sections
  - Calculate the average optical flow for each section
  - The frame numbers of each section are calculated by the proportion of optical flow value of the section and the whole video

Our Scheme

C. Feature extraction



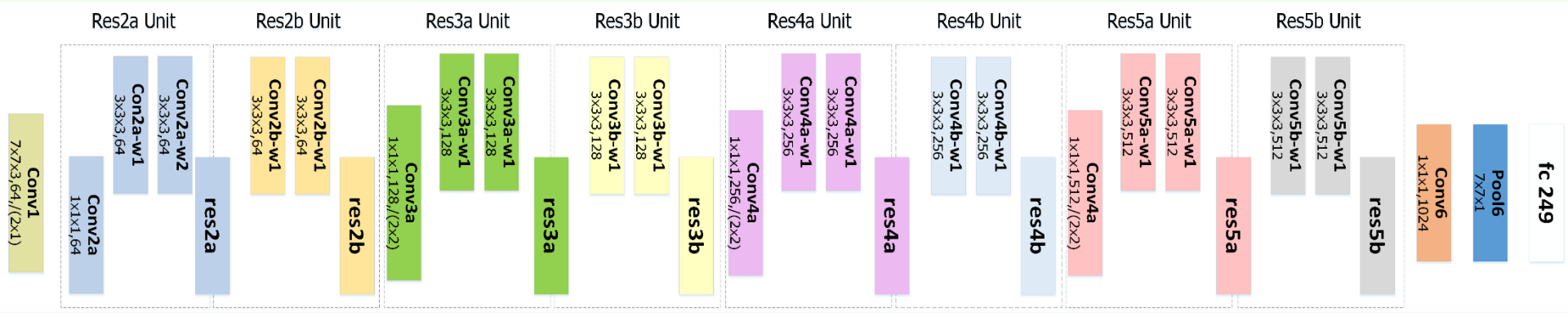
C3D



ResNet

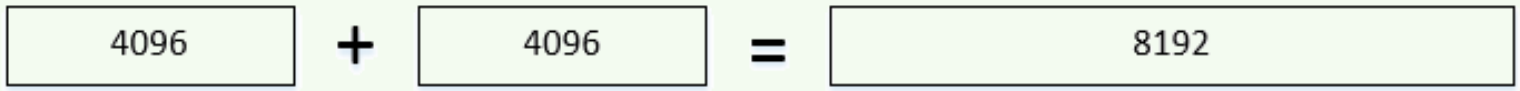
Our Scheme

C. Feature extraction



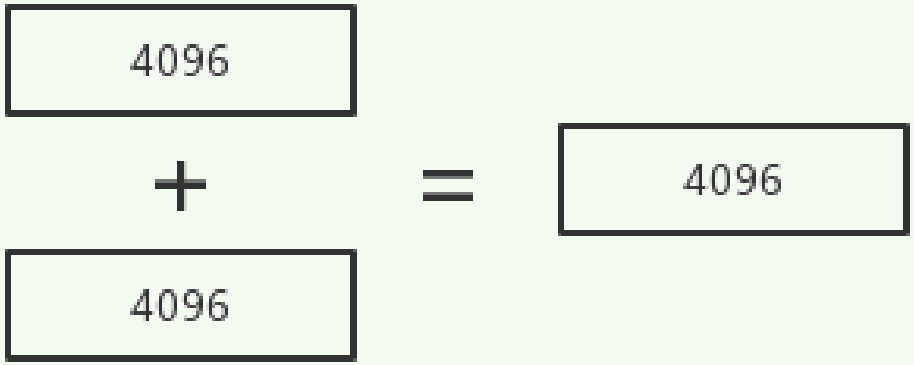
D. Feature fusion

- Traditional methods
  - Parallel (averaging)



D. Feature fusion

- Traditional methods
  - Parallel (averaging)
  - Serial (concatenating)



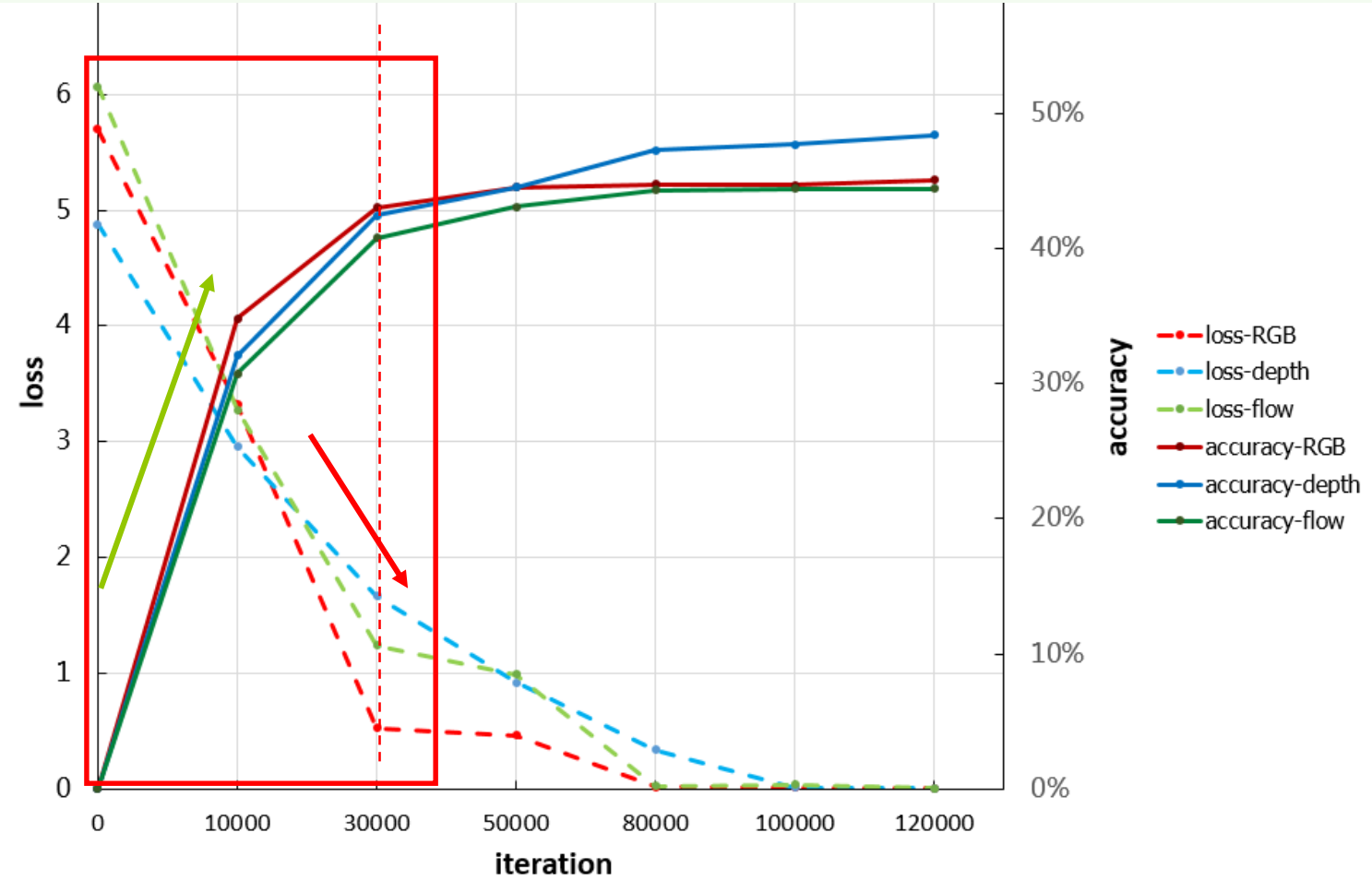
D. Feature fusion

- Canonical Correlation Analysis
  - a way of inferring information from cross-covariance matrices
  - CCA tries to maximize the pair-wise correlations across features with different modalities.

CONTENTS

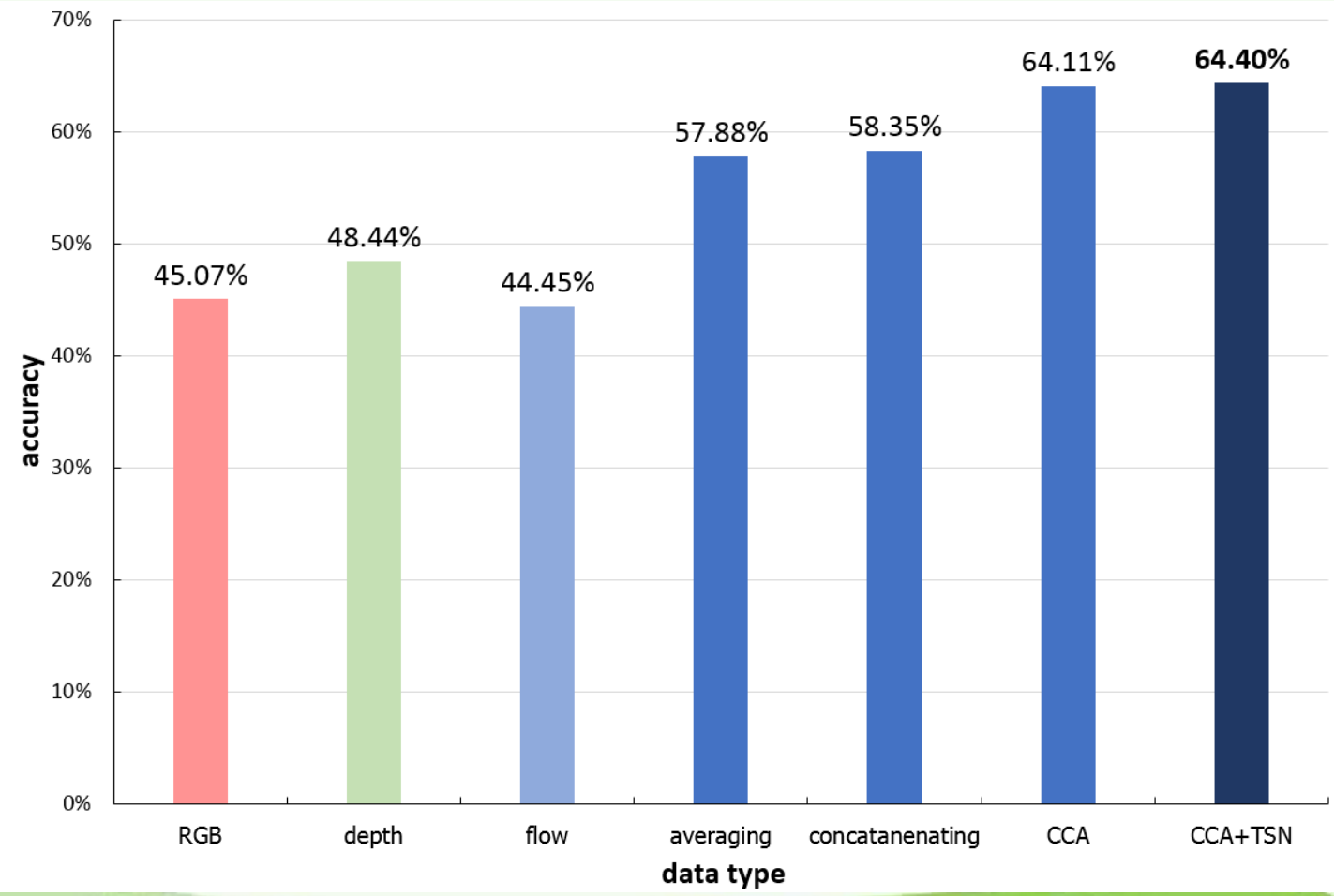
EXPERIMENTAL RESULTS

Iteration Times



EXPERIMENTAL RESULTS

Fusion



EXPERIMENTAL RESULTS

Comparison

Method	Accuracy
MFSK+BoVW [34]	18.65%
SFAM (Multi-score Fusion) [37]	36.27%
CNN+depth maps [38]	39.23%
Pyramidal C3D [45]	45.02%
2SCVN+3DDSN [5]	49.17%
32-frame C3D [19]	49.20%
C3D+LSTM [46]	51.02%
proposed method	**64.40 %**

- J. Wan, S. Z. Li, Y. Zhao, S. Zhou, I. Guyon, and S. Escalera. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In IEEE CVPR Workshops, pages 56–64. 2016.
- P.Wang,W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona. Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks.In IEEE CVPR, 2017.
- P. Wang, W. Li, S. Liu, Z. Gao, C. Tang, and P. Ogunbona. Large-scale isolated gesture recognition using convolutional neural networks. In IEEE ICPR Workshops, 2016.
- G. Zhu, L. Zhang, L. Mei, J. Shao, J. Song, and P. Shen. Large-scale isolated gesture recognition using pyramidal 3d convolutional networks. In IEEE ICPR Workshops, 2016.
- J. Duan, J. Wan, S. Zhou, X. Guo, and S. Li. A unified framework for multi-modal isolated gesture recognition. In ACM Transactions on Multimedia Computing, Communications, and Applications,2017
- Y. Li, Q. Miao, K. Tian, Y. Fan, X. Xu, R. Li, and J. Song. Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model. In IEEE ICPR Workshops. 2016.
- G. Zhu, L. Zhang, P. Shen, and J. Song. Multimodal gesture recognition using 3d convolution and convolutional lstm. IEEE Access, 2017.

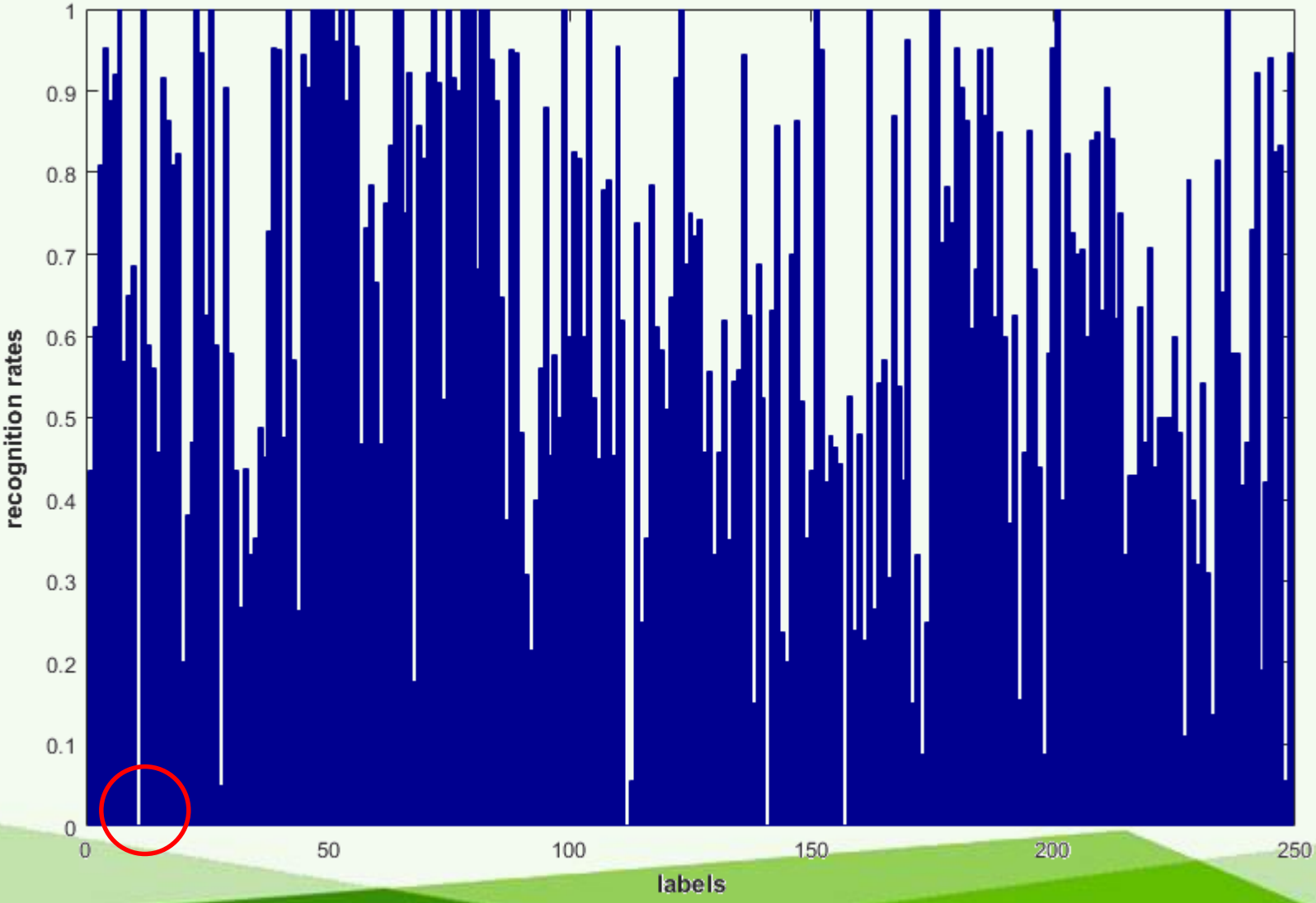
EXPERIMENTAL RESULTS

Comparison

Team	Accuracy (validation)	Accuracy (testing)
baseline [5]	49.17%	67.26 %
XDETVP	58.00%	60.47%
AMRL	60.81%	65.59%
Lostoy	62.02%	65.97%
SYSU_ISEE	59.70%	67.02%
ASU (our)	**64.40 %**	**67.71 %**

CONTENTS

FUTURE WORK


recognition rates
labels
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150
200
250

FUTURE WORK

Thank you !