RWF-2000: An Open Large Scale Video Database for Violence Detection

Ming Cheng[1], Kunjing Cai[2], Ming Li[1]

[1] Data Science Research Center, Duke Kunshan University
[2]School of Data and Computer Science, Sun Yat-sen University

昆山杜克大学
DUKE KUNSHAN
UNIVERSITY

中山大学
Sun Yat-sen
University

OUTLINE

1. Introduction
2. Datasets
3. Methods
4. Experiments
5. Demonstration

1.1 Motivation

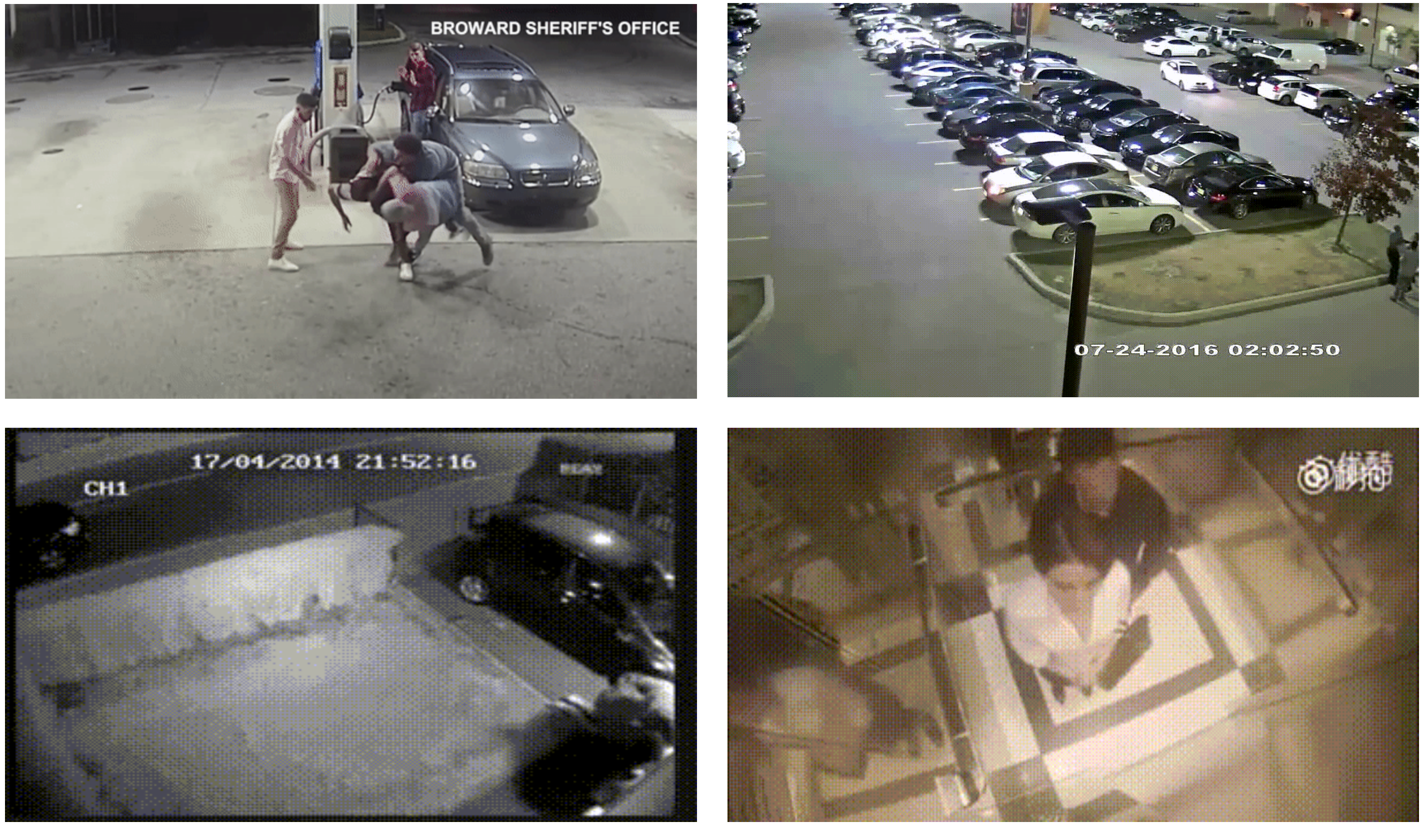


Fig 1. Violent Activities in cities

- Surveillance cameras just provide cues and evidences after crimes have been conducted.
- It is both time and labor consuming to manually monitor the large amount of video data.
- Automatically recognizing violence becomes important.

1.2 Related Work

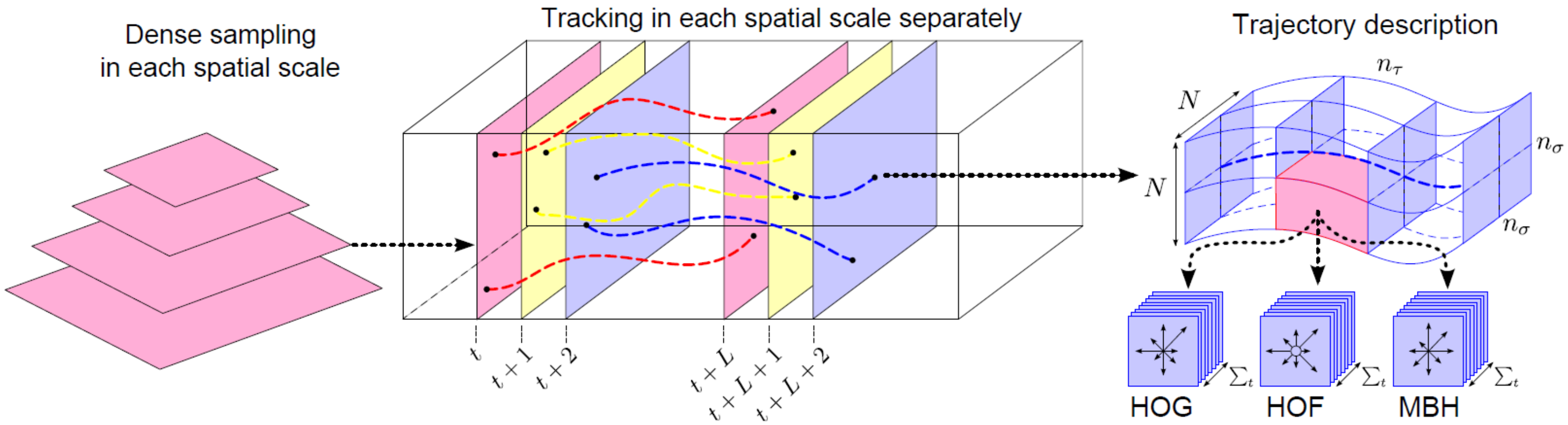


Fig 2. Traditional Method

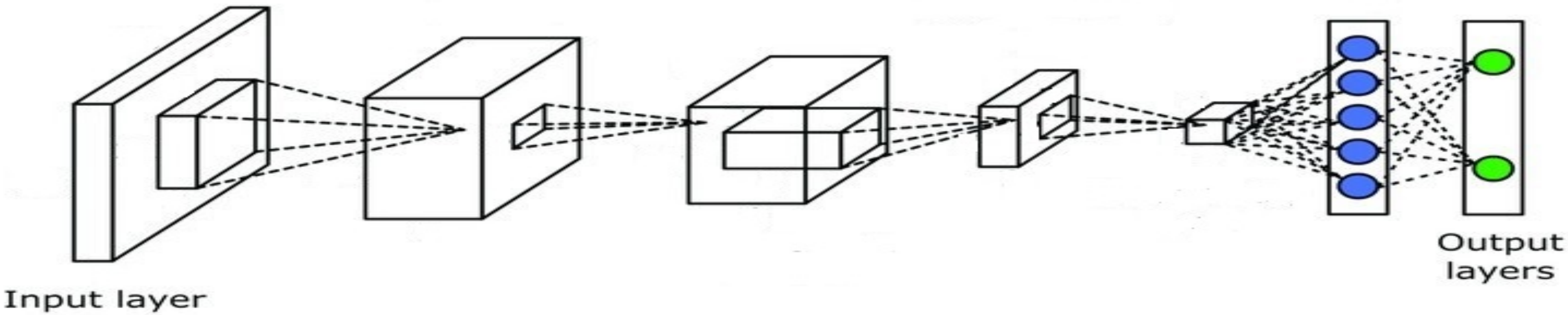
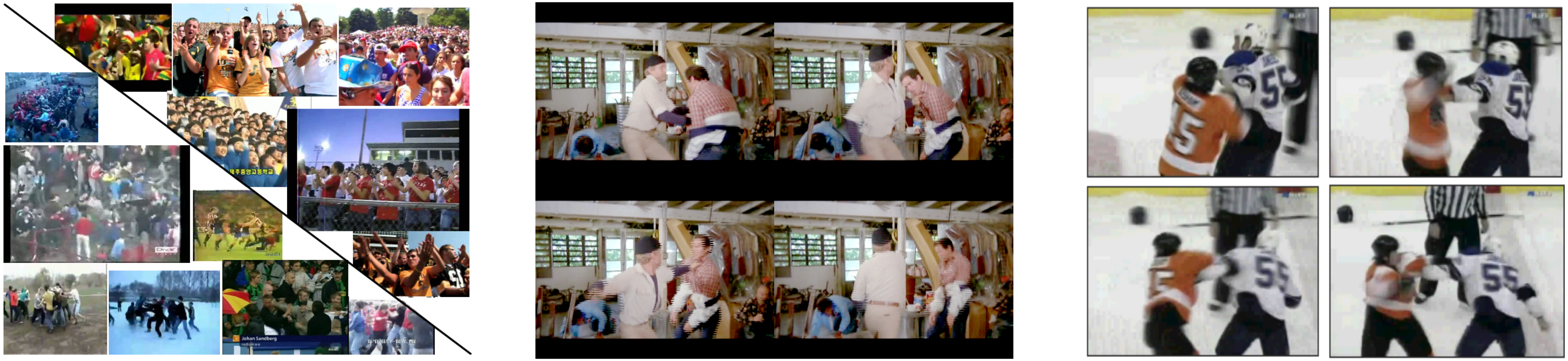


Fig 3. Deep Learning based Method

2.1.1 Previous Datasets



Crowd Violence

246 videos captured in crowded places

Movies Fight

200 videos extracted from action movies

Hockey Fight

1k videos extracted from hockey games

Fig 4. Previous Datasets

2.1.2 Previous Datasets

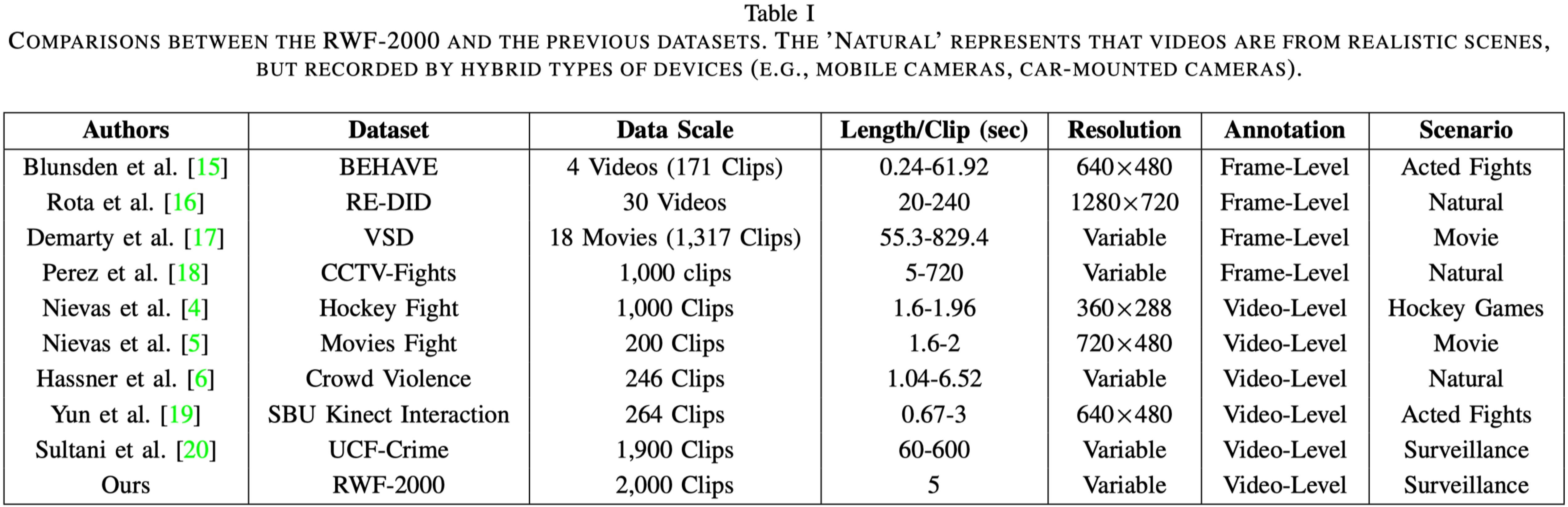


Table I
COMPARISONS BETWEEN THE RWF-2000 AND THE PREVIOUS DATASETS. THE 'NATURAL' REPRESENTS THAT VIDEOS ARE FROM REALISTIC SCENES, BUT RECORDED BY HYBRID TYPES OF DEVICES (E.G., MOBILE CAMERAS, CAR-MOUNTED CAMERAS).

Authors	**Dataset**	**Data Scale**	**Length/Clip (sec)**	**Resolution**	**Annotation**	**Scenario**
Blunsden et al. [15]	BEHAVE	4 Videos (171 Clips)	0.24-61.92	640×480	Frame-Level	Acted Fights
Rota et al. [16]	RE-DID	30 Videos	20-240	1280×720	Frame-Level	Natural
Demarty et al. [17]	VSD	18 Movies (1,317 Clips)	55.3-829.4	Variable	Frame-Level	Movie
Perez et al. [18]	CCTV-Fights	1,000 clips	5-720	Variable	Frame-Level	Natural
Nievas et al. [4]	Hockey Fight	1,000 Clips	1.6-1.96	360×288	Video-Level	Hockey Games
Nievas et al. [5]	Movies Fight	200 Clips	1.6-2	720×480	Video-Level	Movie
Hassner et al. [6]	Crowd Violence	246 Clips	1.04-6.52	Variable	Video-Level	Natural
Yun et al. [19]	SBU Kinect Interaction	264 Clips	0.67-3	640×480	Video-Level	Acted Fights
Sultani et al. [20]	UCF-Crime	1,900 Clips	60-600	Variable	Video-Level	Surveillance
Ours	RWF-2000	2,000 Clips	5	Variable	Video-Level	Surveillance

2.2.1 Proposed Dataset

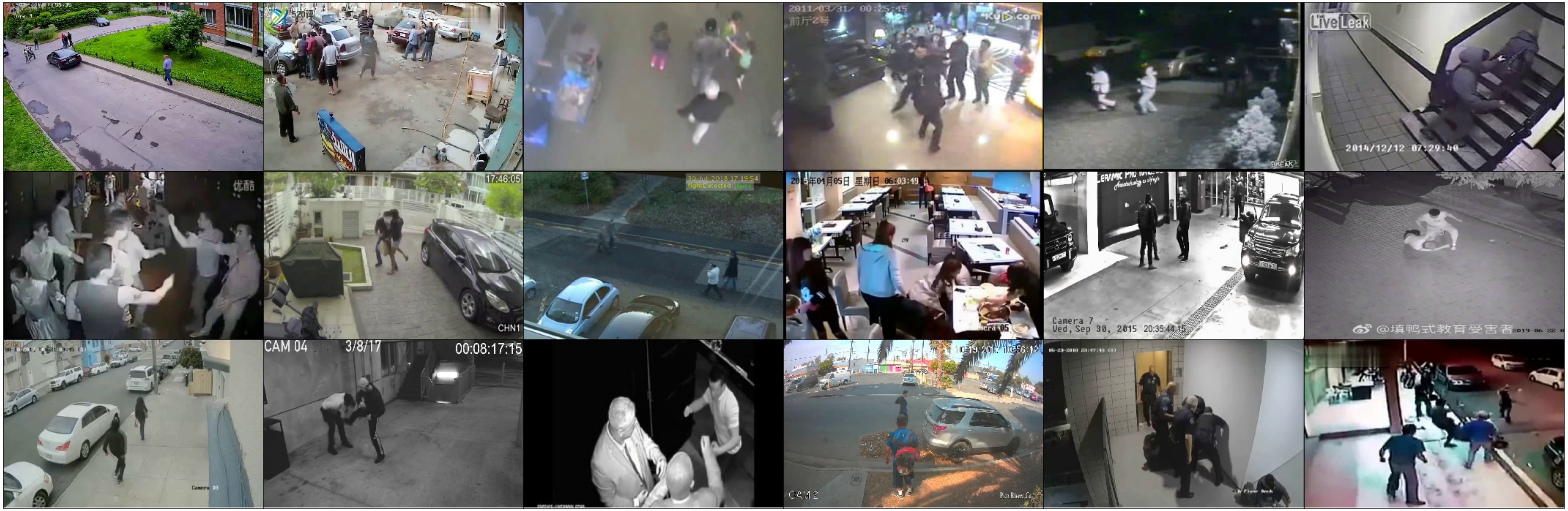


Fig 5. RWF-2000 Dataset

2000 real-world videos captured by surveillance cameras, with large diversity

2.2.2 Proposed Dataset

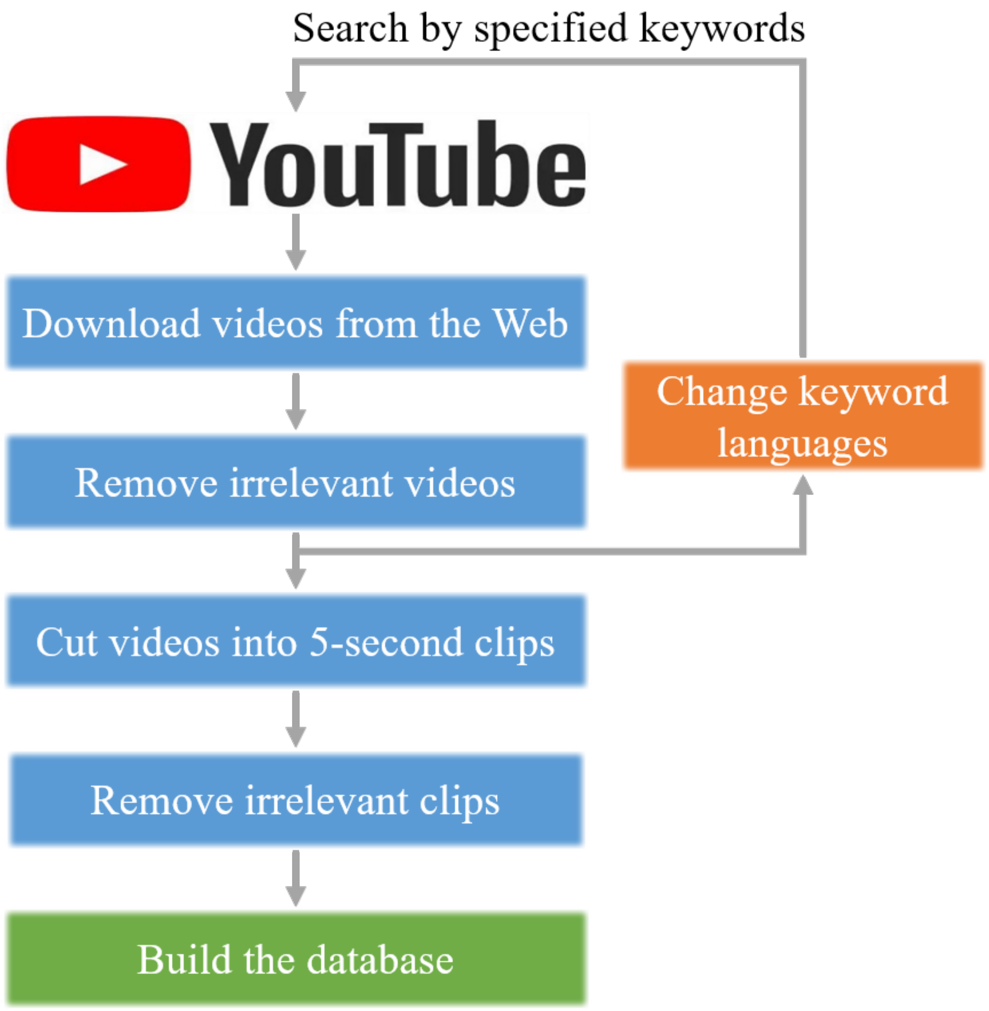


Fig 6. Pipeline of Data Collection

Collections of the RWF-2000

- Search and download videos from the YouTube website
- Remove irrelevant contents and cut videos into clips
- Repeat the above procedures by changing specified keywords
- Annotate collected clips manually to build the database

2.2.3 Proposed Dataset

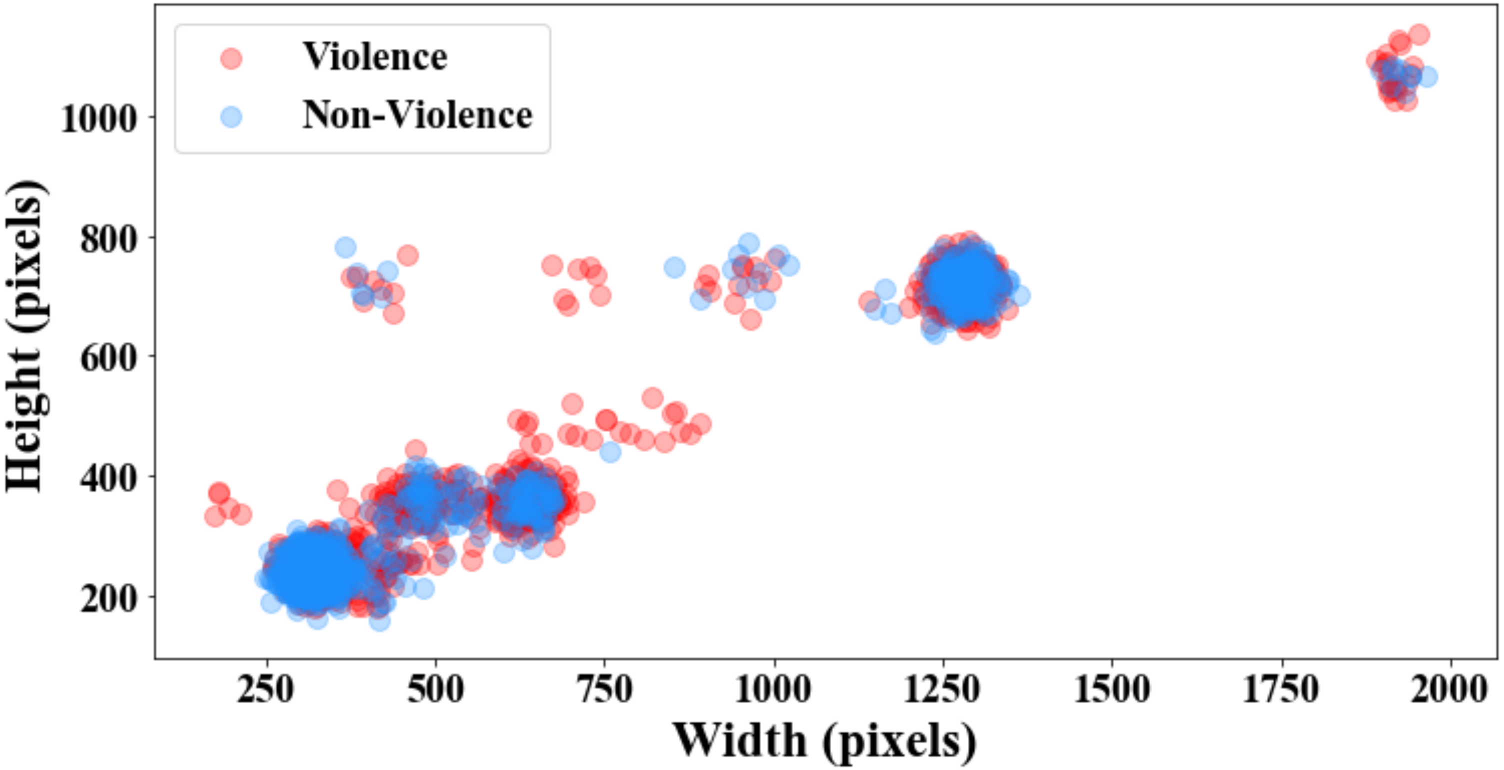


Fig 7. Resolution Distribution of the RWF-2000

Properties of the RWF-2000

- Large diversity
- Real-world scenes
- Adaptive to surveillance cameras

3.1 Proposed Method

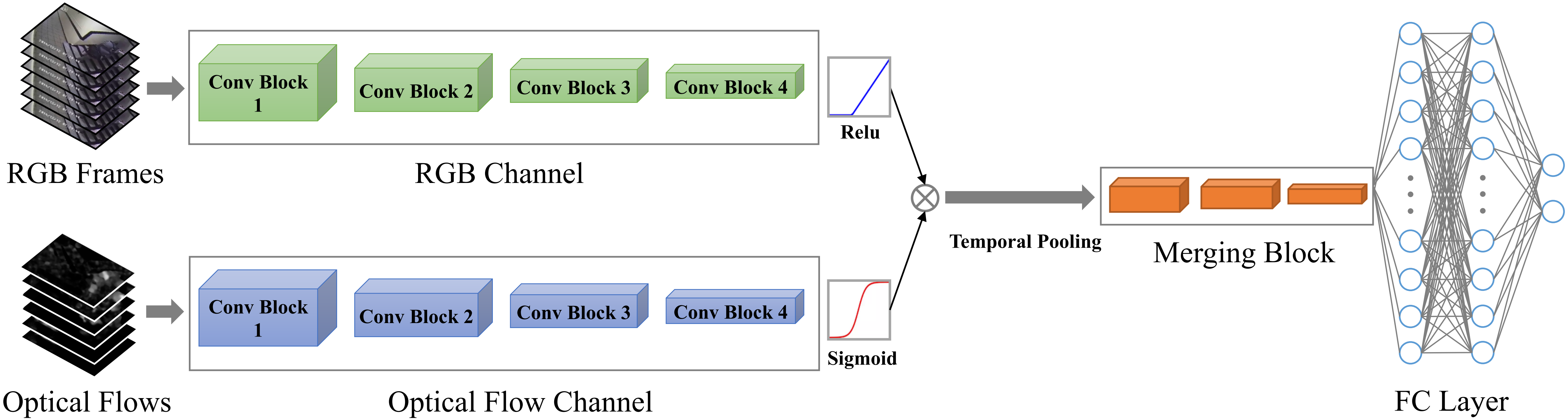


Fig 8. Structure of proposed method

3.2 Cropping Strategy

Optical flow is a field of 2D vector, we could calculate the norm of vector to represent magnitude of motion.

$$|\boldsymbol{v}(x,y)| = \sqrt{v_x^2 + v_y^2}$$

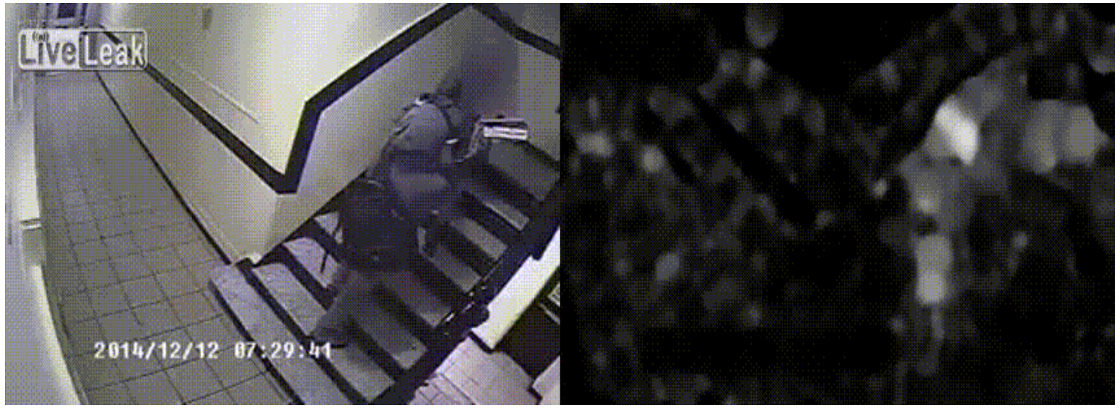
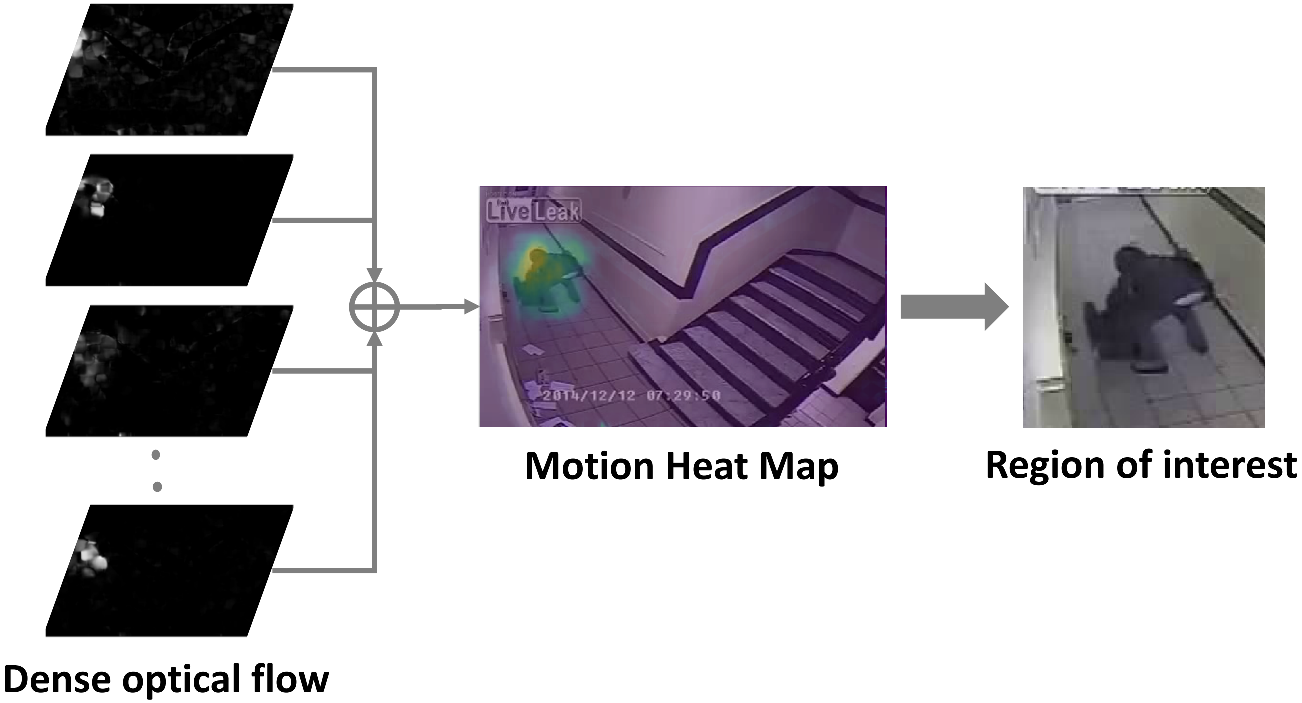


Fig 9. Motion estimation using optical flow



Dense optical flow

Motion Heat Map

Region of interest

Fig 10. Extracting the region of interest

3.3 Sampling Strategy

Video data has much redundant information between neighboring frames, a sparse sampling strategy is implemented to reduce the amount of computing cost.

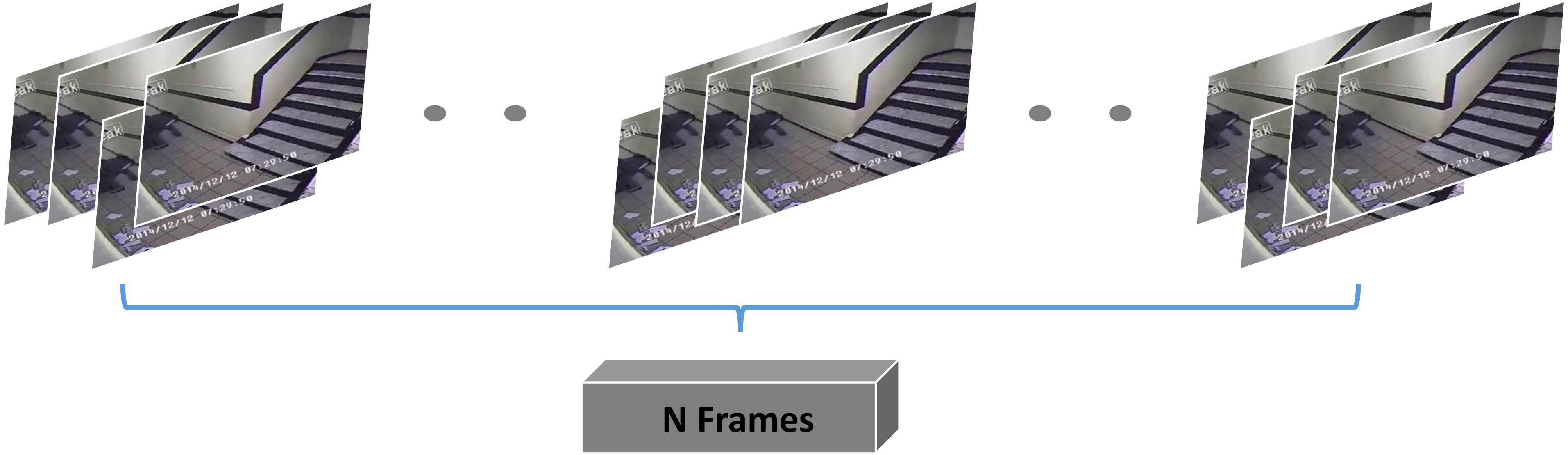
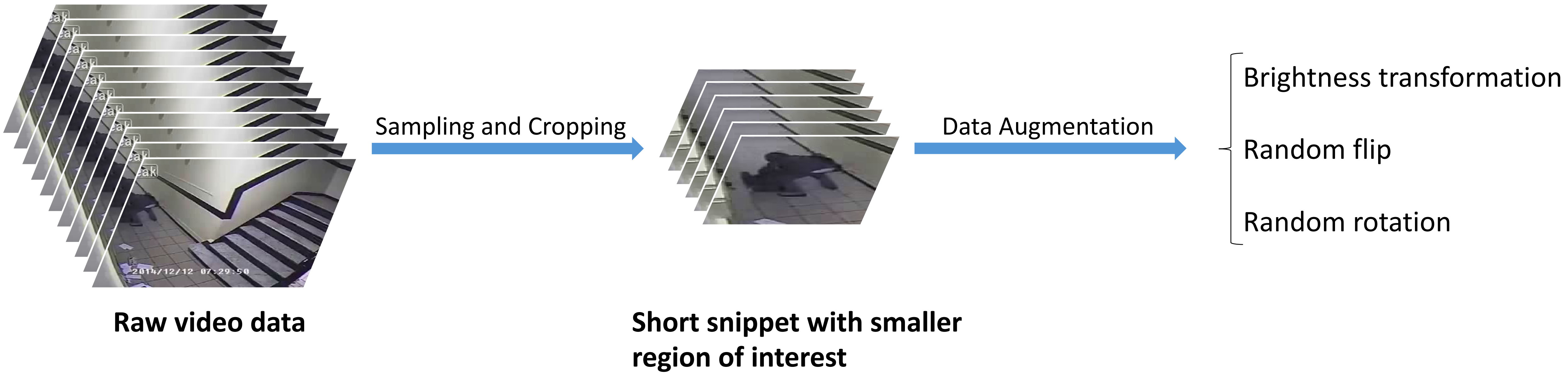


Fig 11. Sparse sampling

3.4 Combination



Sampling and Cropping

Data Augmentation

- Brightness transformation
- Random flip
- Random rotation

Raw video data

Short snippet with smaller region of interest

Fig 12. Combination of sampling and cropping

4.1 Experiments

In the training process, SGD optimizer with momentum (0.9) and learning rate decay (1e-6) were implemented. After 6,000 iterations of training, our model obtained the best accuracy of 87.25% on the test set (shown in Table III).

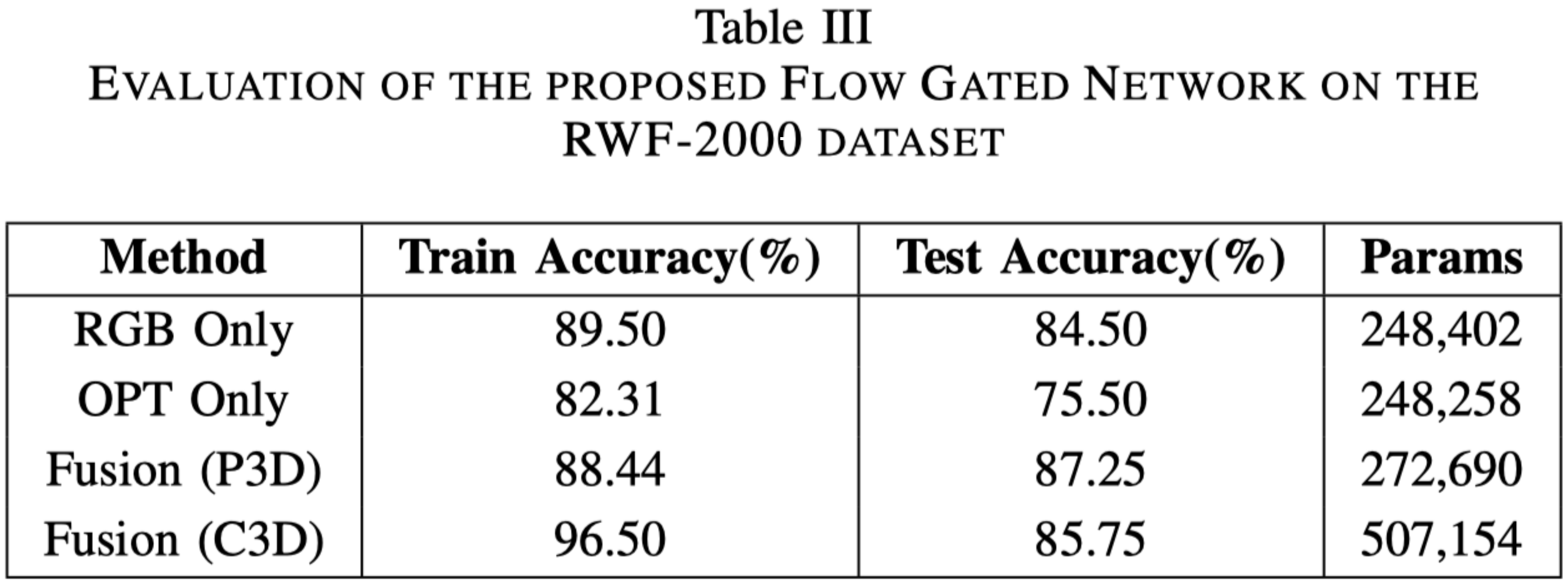


Table III
EVALUATION OF THE PROPOSED FLOW GATED NETWORK ON THE RWF-2000 DATASET

Method	Train Accuracy(%)	Test Accuracy(%)	Params
RGB Only	89.50	84.50	248,402
OPT Only	82.31	75.50	248,258
Fusion (P3D)	88.44	87.25	272,690
Fusion (C3D)	96.50	85.75	507,154

4.2 Comparisons

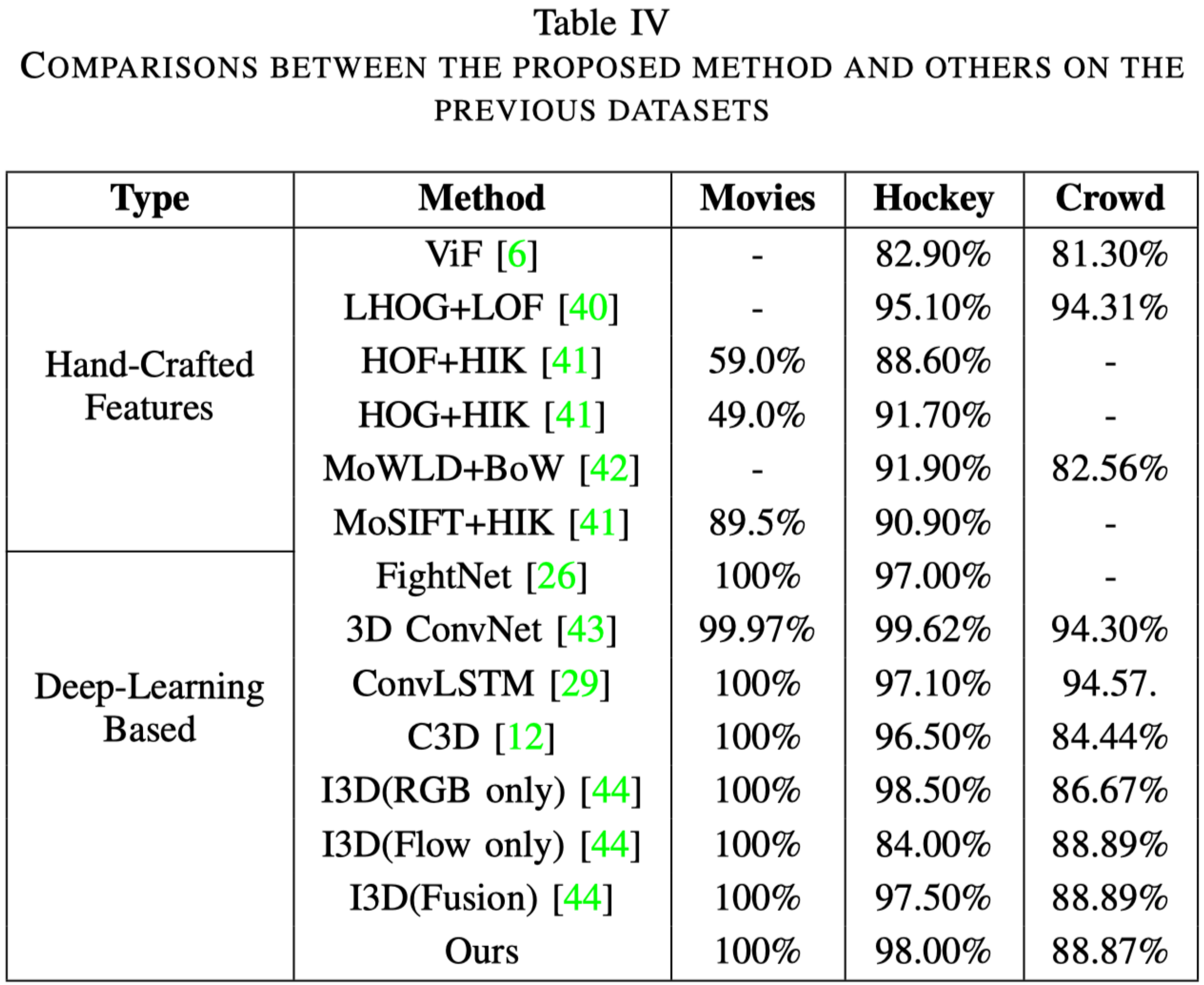


Table IV
COMPARISONS BETWEEN THE PROPOSED METHOD AND OTHERS ON THE PREVIOUS DATASETS

Type	Method	Movies	Hockey	Crowd
Hand-Crafted Features	ViF [6]	-	82.90%	81.30%
	LHOG+LOF [40]	-	95.10%	94.31%
	HOF+HIK [41]	59.0%	88.60%	-
	HOG+HIK [41]	49.0%	91.70%	-
	MoWLD+BoW [42]	-	91.90%	82.56%
	MoSIFT+HIK [41]	89.5%	90.90%	-
Deep-Learning Based	FightNet [26]	100%	97.00%	-
	3D ConvNet [43]	99.97%	99.62%	94.30%
	ConvLSTM [29]	100%	97.10%	94.57.
	C3D [12]	100%	96.50%	84.44%
	I3D(RGB only) [44]	100%	98.50%	86.67%
	I3D(Flow only) [44]	100%	84.00%	88.89%
	I3D(Fusion) [44]	100%	97.50%	88.89%
	Ours	100%	98.00%	88.87%

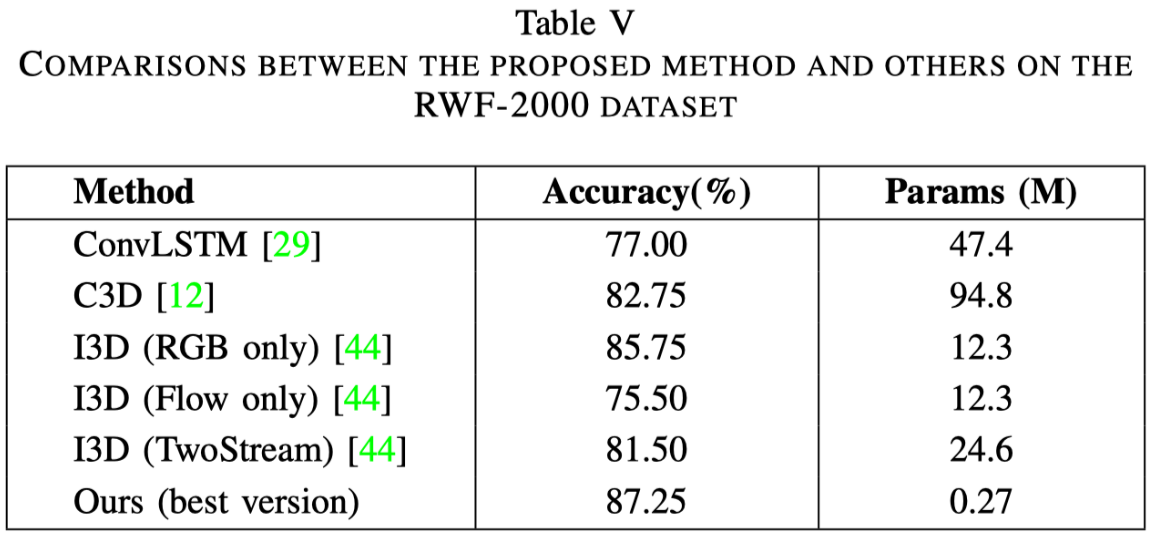
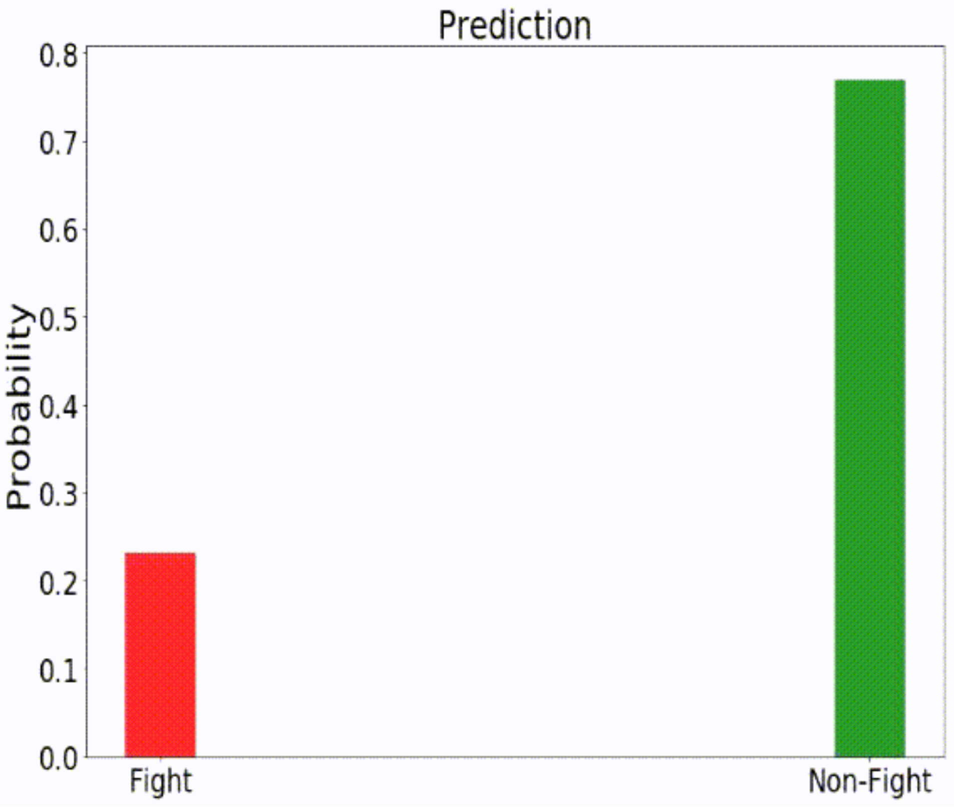


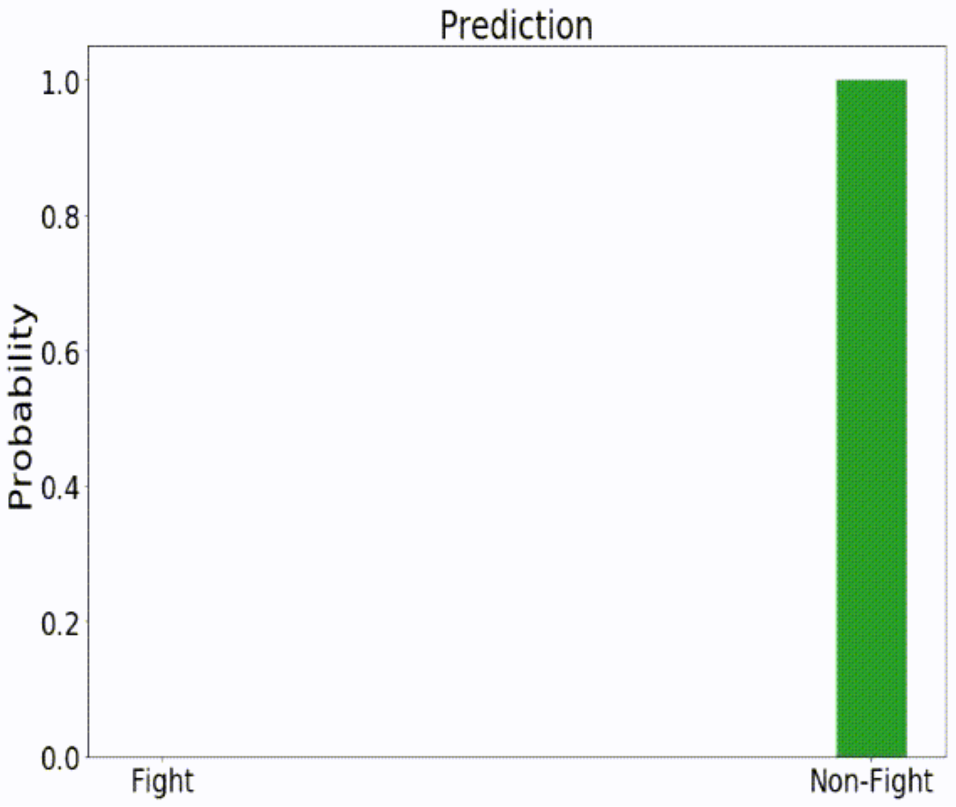
Table V
COMPARISONS BETWEEN THE PROPOSED METHOD AND OTHERS ON THE RWF-2000 DATASET

Method	Accuracy(%)	Params (M)
ConvLSTM [29]	77.00	47.4
C3D [12]	82.75	94.8
I3D (RGB only) [44]	85.75	12.3
I3D (Flow only) [44]	75.50	12.3
I3D (TwoStream) [44]	81.50	24.6
Ours (best version)	87.25	0.27

5.1 Demonstration



5.2 Demonstration



6 Future Work

- Videos from fixed cameras and mobile cameras could be treated differently.
- Dense optical flow is computationally expensive, an end-to-end model will be faster.
- The RWF-2000 dataset will be released as soon as possible, welcome to contact me for downloading it (ming.cheng@dukekunshan.edu.cn).