NTT_CQUPT@TRECVID2019 ActEV

Yongqing Sun[1], Xu Chen[2], Chaoyu Li[2], Kiyohito Sawada[3], Takashi Hosono[1], Jun Zhu[2], Chengjuan Xie[2], Sixiang Huang[2], Lan Wang[2], Kai Hu[2], Qingsong Zhou[2], Chenqiang Gao[2], Jun Shimamura[1], Atsushi Sagata[1]

1: NTT, 2: Chongqing Univ. of Posts and Telecommunications, 3: National Police Academy

Unique points

1. **Classification by Conv-LSTM, which can preserve spatial-temporal information**
2. **Various post-processing to suppress false alarm**
3. **Proposal alignment to learn efficiently with few training data**

System overview

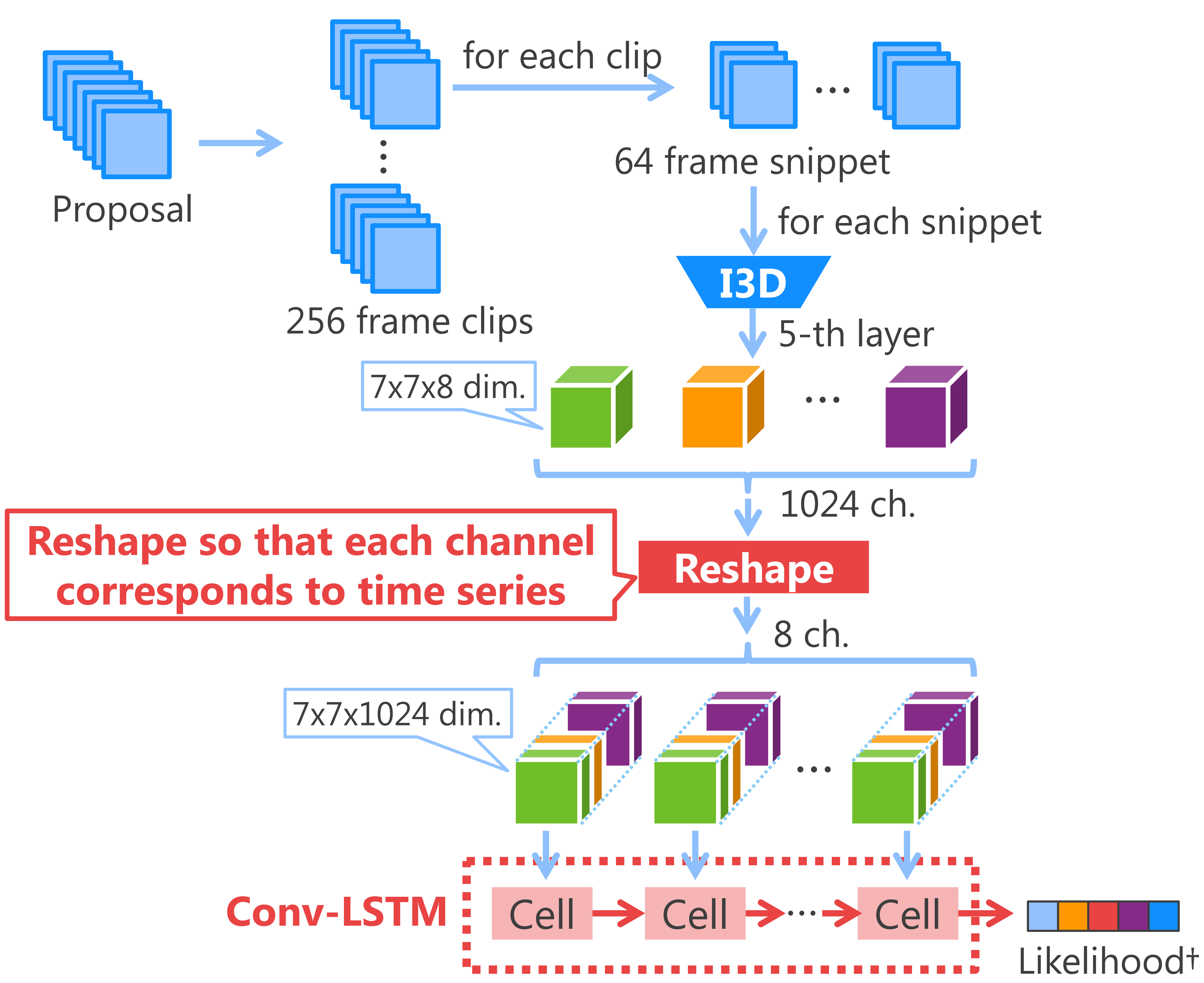
Input → **Object detection** → **Proposal generation** → **Feature extraction** → **Classification** → **Post-processing** → **Event category and location**

- **Object detection:** detect objects by Mask R-CNN fine-tuned with our extra annotation
- **Proposal generation:** track each object by tracking method and merge them if person and vehicle are nearby
- **Feature extraction:** divide proposal into snippets and extract features using I3D network
- **Classification:**
  - ✓ Reshape I3D feature and predict likelihood for each action by Conv-LSTM
  - ✓ Fine-tune by hard negative mining
- **Post-processing:** merge estimation results for each proposal and delete those that seem to be false alarm

Feature extraction and classification

- X and y axis of I3D feature have spatial information, and z axis has temporal information
- **Problem: simple flattening algorithm loses spatial-temporal information**
- **Solution: reshape I3D features and use Conv-LSTM**

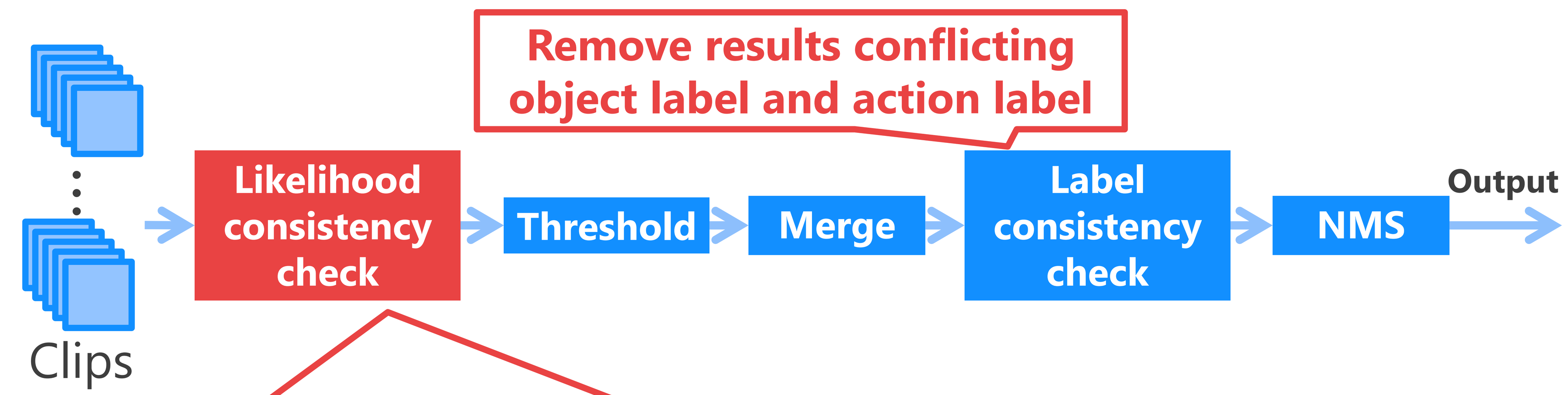
Flow of action likelihood prediction for each clip



† Clip's likelihood is average of all snippets

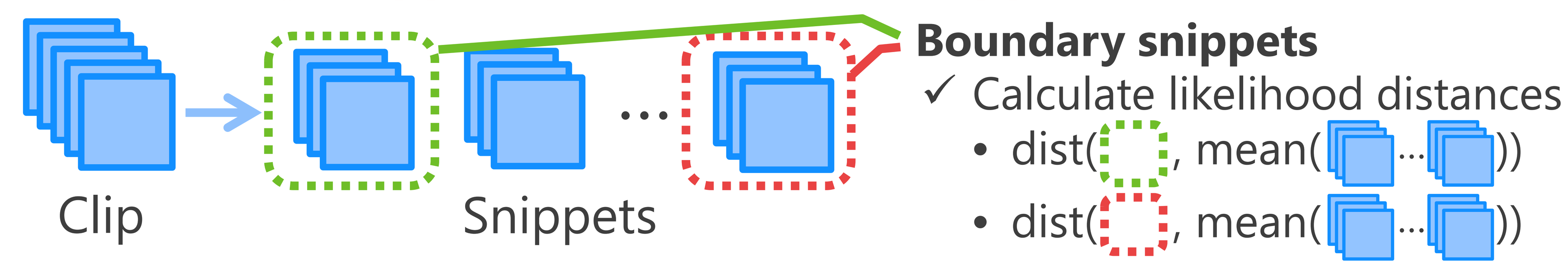
Post-processing

- **Problem: simply merging results of each clip produces extra estimation results**
- **Solution: remove extra results by likelihood consistency check, threshold, merge, label consistency check and NMS**



Likelihood consistency check

- **Assumption: clips that have large likelihood difference of boundary snippet and mean of all snippets are extra results**

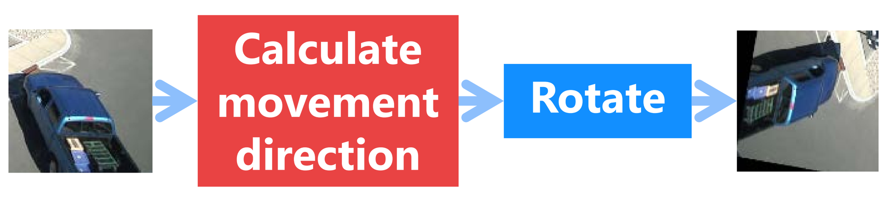


Proposal alignment*

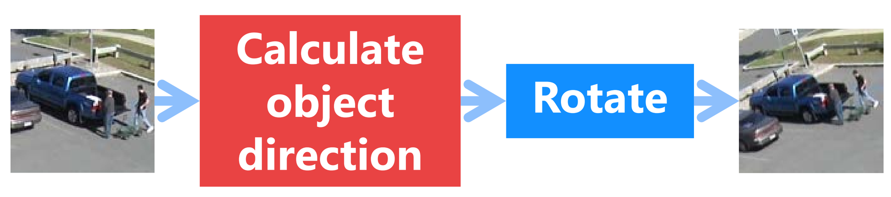
* it's not included in submitted system

- **Observation:** Each action has diversity of appearance due to various movement/object direction
- **Assumption: this diversity makes learning and predicting action recognition difficult**
- **Solution: rotate proposals to align movement/object direction**

Vehicle proposal alignment



Person-vehicle proposal alignment



Evaluation

Whole system results on test data

- Difference among our systems is post-processing threshold
- **5th accuracy at time of submission**

System	Partial AUCD	Mean-Pmiss @0.15TFA	Mean W Pmiss @0.15RFA
p-NTT-CQUPT	**0.60058**	**0.51122**	0.87254
p2_NTT_CUPT	0.60396	0.51677	**0.87168**
system2	0.60524	0.51755	0.87381
NIST-TEST (baseline)	0.85649	n/a	n/a

Each module results on validation data

Object detection results

Method	mAP
Original Mask R-CNN	19.6%
Fine-tuned Mask R-CNN	44.1%

Proposal generation result

Method	Number of proposals	Recall
Ours	4,151	85.6%

Classification results

✓ We used our proposals

Input	Method	mAP
RGB	Before fine-tuning	13.2%
	After fine-tuning	16.7%
Optical flow	Before fine-tuning	12.8%
	After fine-tuning	13.1%

Proposal alignment results

✓ We used GT proposals

Method	mAP
Without alignment	46.1%
With alignment	49.7%