\Orchestrating a brighter world

NEC

A Hybrid Two-stream Approach For Multi-person Action Recognition in Top-view 360º Videos

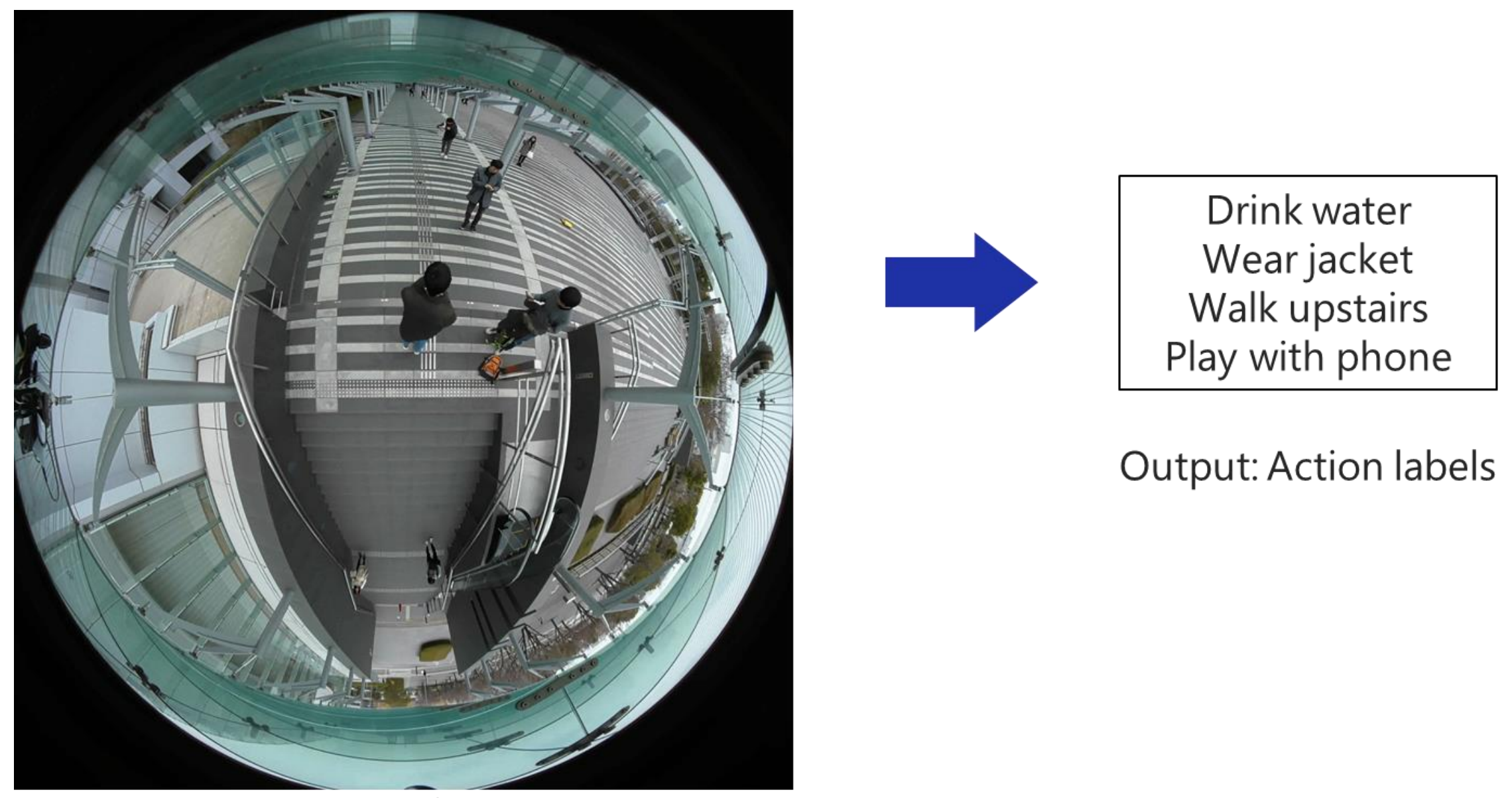
Karen Stephen[1], Jianquan Liu[1], Vivek Barsopia[2]
[1]Biometrics Research Laboratories [2]Data Science Research Laboratories
NEC Corporation, Japan

2021 IEEE International Conference on Image Processing (ICIP)

Introduction

Task:

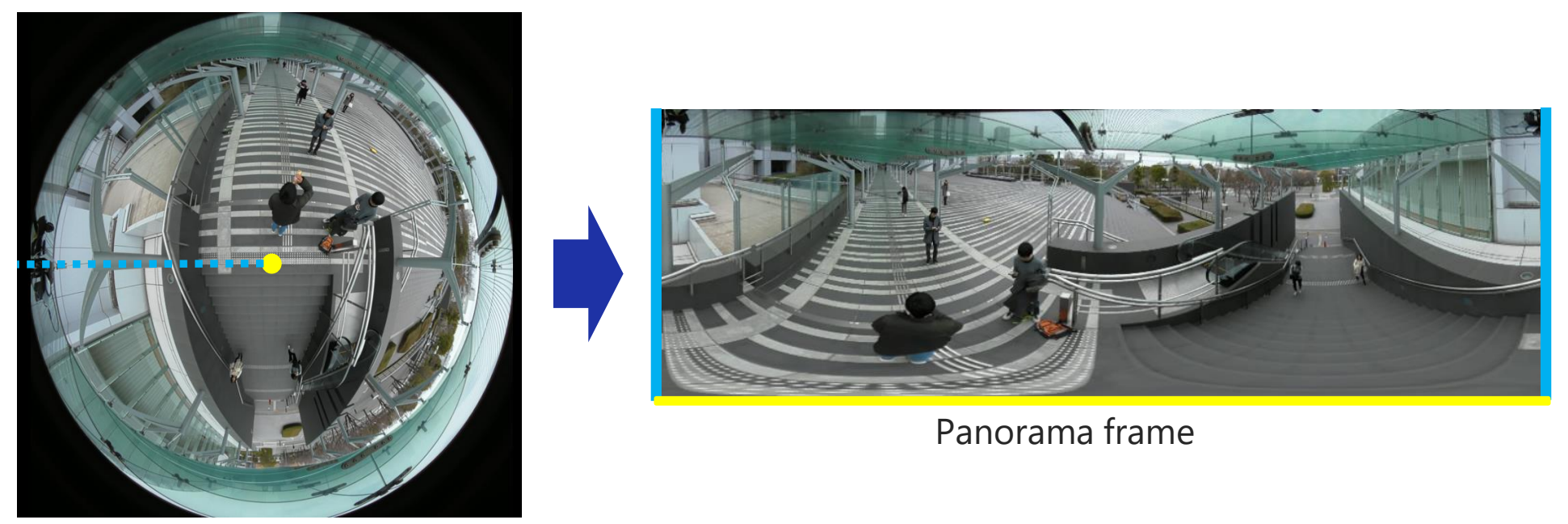
Multi-person action recognition from top-view 360^{o} videos



Input: 360^{o} video

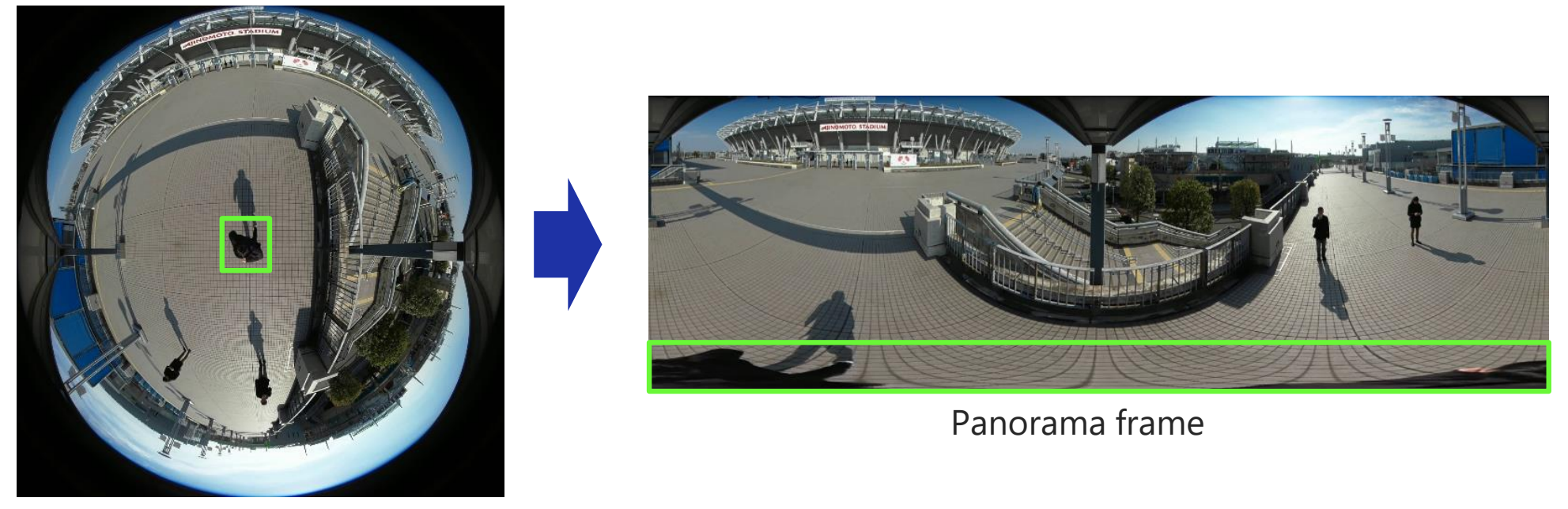
Challenges:

- **Unavailability of large-scale 360^{o} action datasets** to train existing deep learning models for action recognition in 360^{o} videos.
- Existing work utilizes a **global projection method** to transform 360^{o} video frames to panorama frames and uses a pre-trained network trained on perspective videos.



Input 360^{o} video frame

- This unwrapping suffers from **geometric distortion** i.e., people present near the center in the 360^{o} video frames appear highly stretched and distorted in the corresponding panorama frames, thereby **affecting the overall action recognition performance**.



Input 360^{o} video frame

- Other projection methods like cube-map or icosahedral projection reduces the amount of distortion in the projected image. But, they **introduce discontinuities** at the cube or icosahedral faces, causing the persons in the images to be cut into different parts.

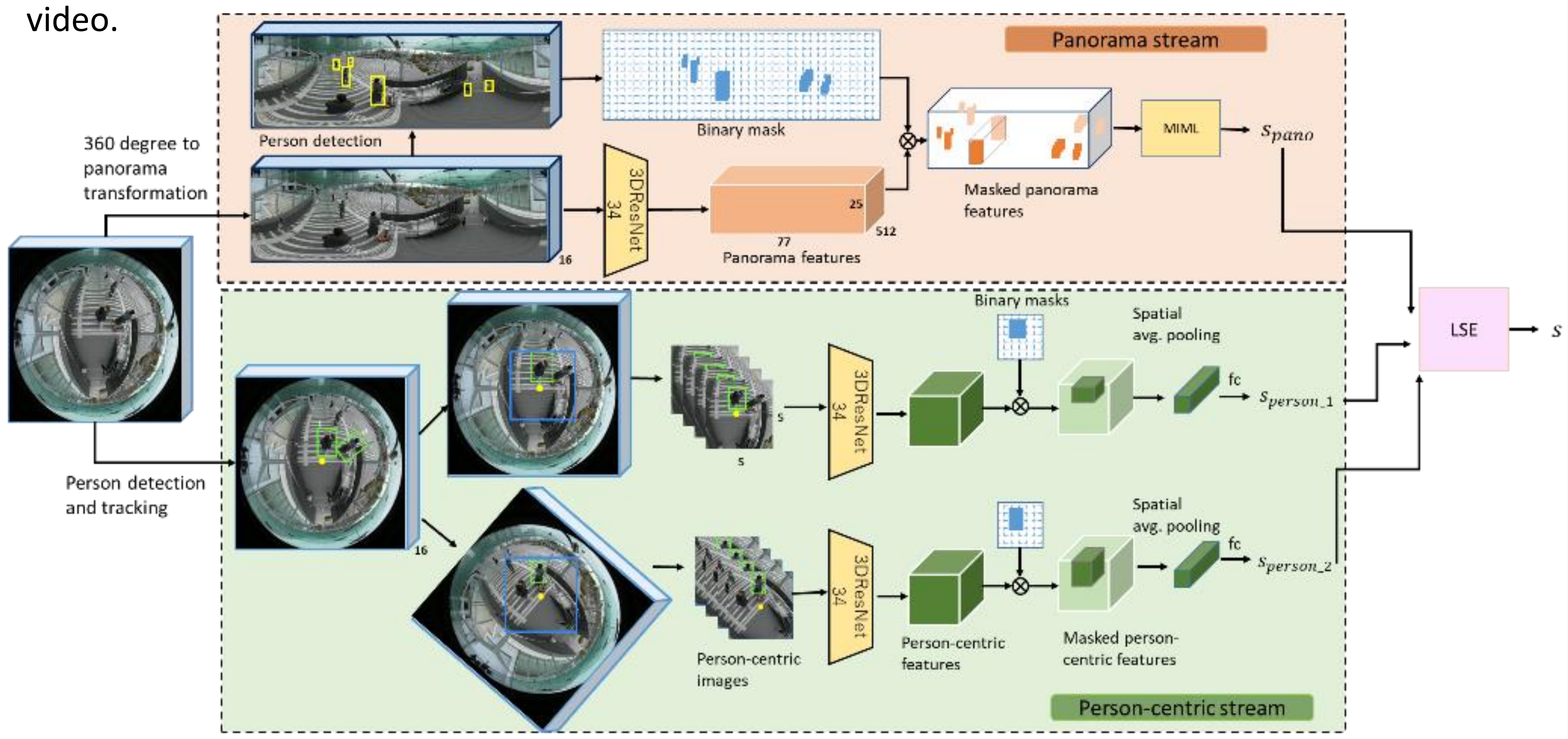
Our Approach:

In this work, we overcome the problem of distortion by utilizing **distortion-free person-centric images** of persons near the center (extracted directly from the input 360^{o} video frames), **along with panorama images, in a hybrid two-stream approach**.

Proposed Method

Network Architecture:

Hybrid two-stream architecture consisting of **a panorama stream and a person-centric stream**. Action scores output from both streams are combined together to detect the overall actions in a video.



Panorama stream:

- Multi-Instance Multi-Label learning module outputs a score vector $s_{pano} = \{s^{a}_{pano}\}\ \forall a \in C$

Person-centric stream:

Enables the recognition of actions of people present near the center in 360º video frames

- Firstly, persons are detected using Rotation-Aware People Detection method (RAPiD) trained on top-view 360º images.
- Secondly, every person within radius R from center person is uprightly aligned by rotating the frame by an angle α_p given by: $\alpha_p = \tan^{-1}\frac{(x_p - x_c)}{(y_c - y_p)}$ (x_c, y_c): co-ordinates of the center of the scene (x_p, y_p): centroid of person bounding box
- Finally, person-centric images are cropped out and input to a convolutional network that outputs action scores for each person.

Combining the two streams:

Since only one set of action scores has to be output for a video, scores from both the streams are aggregated using a Log Sum Exponential (LSE) score aggregator

$$s^{a} = \log \sum_{i=1}^{N} \exp(s^{a}_{i}) \qquad N = N_{pano} + N_{person}$$

Total Loss: $L = L_{bce} + \lambda_1 L_{reg\_person} + \lambda_2 L_{reg\_instance}$

Multi-label Binary Cross Entropy Loss: $L_{bce} = -\sum_{a \in C}(y^{a}\log p^{a} + (1-y^{a})\log(1-p^{a}))$

Regularization Loss:

We penalize the model if it outputs high scores for multiple action classes for one person (in the person-centric stream) or one-instance (in the panorama stream)

$$L_{reg\_person} = \sum_{i=1}^{N_{person}} \frac{\sum_a p^{a}_{i} - \max_a p^{a}_{i}}{\max_a p^{a}_{i}} \qquad L_{reg\_instance} = \sum_{j=1}^{n \times N_{pano}} \frac{\sum_a p^{a}_{j} - \max_a p^{a}_{j}}{\max_a p^{a}_{j}}$$

Experiments and Results

Implementation details

- The fully connected layers of both the streams and last layer of the pre-trained 3DResNet-34 are trained while keeping rest of the network frozen.
- For the person-centric stream, **the central area radius was fixed to 750 pixels** and person crop size was fixed to 1504x1504 pixels
- The proposed method was experimentally validated on **360 Action dataset.**

Comparison with state of the art:

Method	mAP %
Collective [T. Bagautdinov et al. CVPR'17]	61.27
3D ResNet [K. Hara et al. ICCV'17]	61.95
R-C3D [H. Xu et al. ICCV'17]	58.74
MiCT [Y. Zhou et al. CVPR'18]	62.18
Panorama 3D-ResNet [J. Li et al. WACV'20]	70.12
Hybrid two-stream (Ours)	**72.40**

Ablation study:

- We performed experiments using both panorama and person-centric streams independently and combined together, to evaluate the effect of each stream on the overall network.

Per-class average precision for all 19 actions in the 360 Action dataset

Method	Eat snack	Phone call	Play with phone	Drink water	Drop sth.	Give sth.	Handshake	Pickup sth.	Wer jacket	Take off jacket	Push	Walk upstairs	Walk downstairs	Wave hand	Take sth.	Walk	Run	Tap in station	Tap out station
Panorama 3D-ResNet [J. Li et al. WACV'20]	39.6	**44.9**	48.0	**48.5**	**51.4**	**56.7**	63.7	65.4	67.0	69.4	73.5	77.9	79.6	81.2	**86.3**	89.8	93.8	**95.4**	**97.5**
Person-centric only (Ours)	**57.9**	44.5	**58.2**	47.5	47.0	50.2	63.2	79.6	77.7	70.9	68.4	**96.8**	81.5	**91.7**	74.1	89.0	86.5	53.1	35.9
Hybrid two-stream (Ours)	40.0	40.4	50.4	39.8	47.4	50.8	**66.2**	**88.1**	**83.3**	**73.8**	**85.8**	82.5	**89.4**	87.4	73.4	**92.8**	**100.0**	88.7	95.9

Ablation studies on the performance of different streams

Method	Panorama	Person-centric	mAP (%)
Panorama 3D-ResNet [J. Li et al. WACV'20]	✓	-	70.12
Person-centric only(Ours)	-	✓	67.0
Hybrid two-stream(Ours)	✓	✓	**72.40**

- Person-centric only (Ours) model uses only the person-centric stream for processing the entire input 360 video frame (not restricted to radius R) and performs better for actions with less motion (subtle actions)
- Using both panorama and person-centric streams together (hybrid two-stream) gives the best overall performance (72.4%)

Inference Speed Analysis:

- The Panorama-branch runs at 4.9 fps, while the person-centric branch runs at 2.66 fps (assuming 3 persons in the center). The hybrid two-stream method in total runs at around 1.7 fps.