DeepPEAR: Deep Pose Estimation and Action Recognition

Authors: You-Ying Jhuang, Wen-Jiin Tsai

National Chiao Tung University, TAIWAN

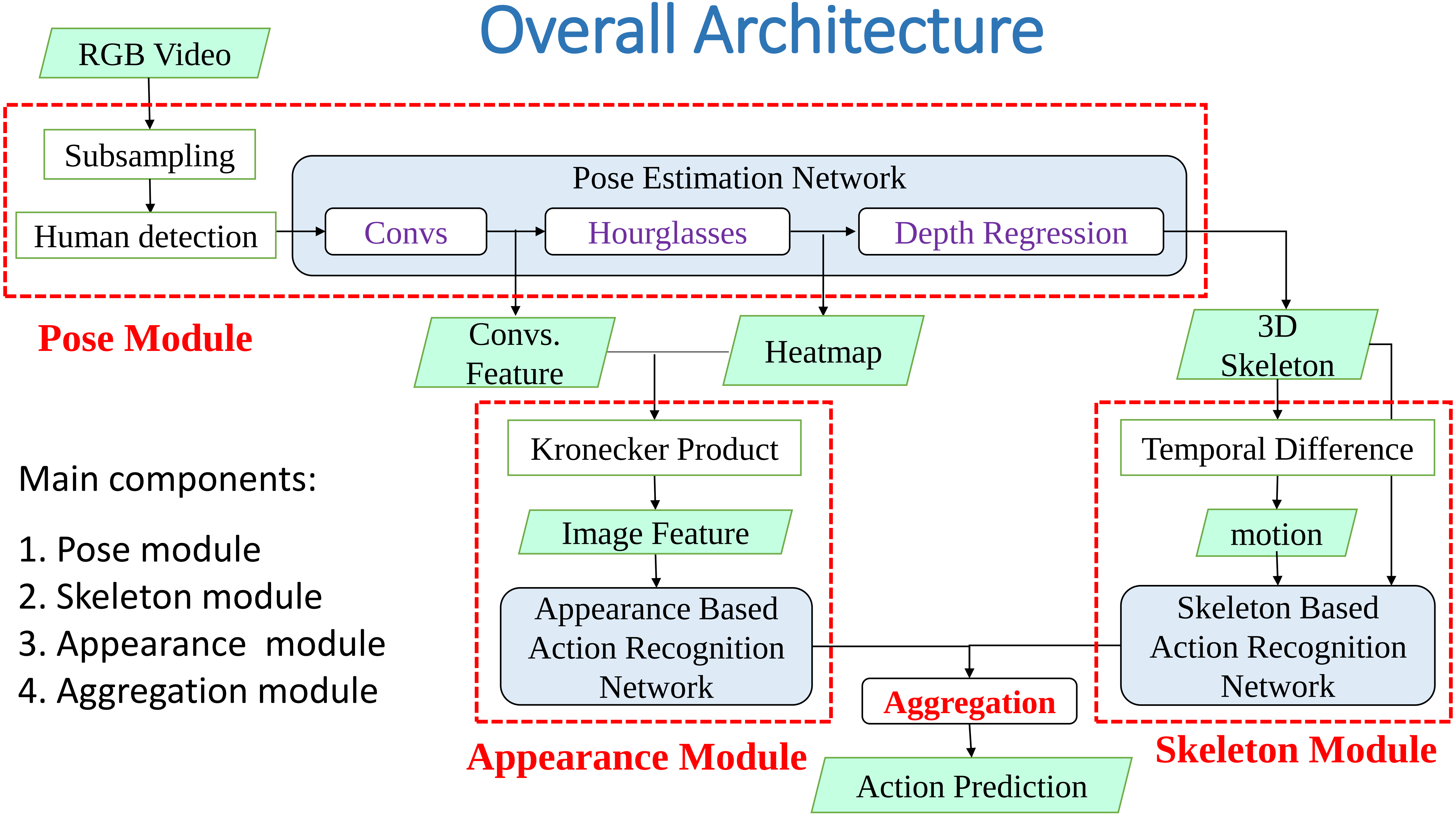
Idea

To use skeleton and appearance features to do action recognition, but use RGB video as the only input.

Main Contributions

- Propose a method to recognize actions from the predicted 3D pose and the appearance features generated by the pose estimation network
- Require less equipment, compared to the skeleton based action recognition
- Achieve state-of-the-art results on NTU RGB+D dataset

Overall Architecture



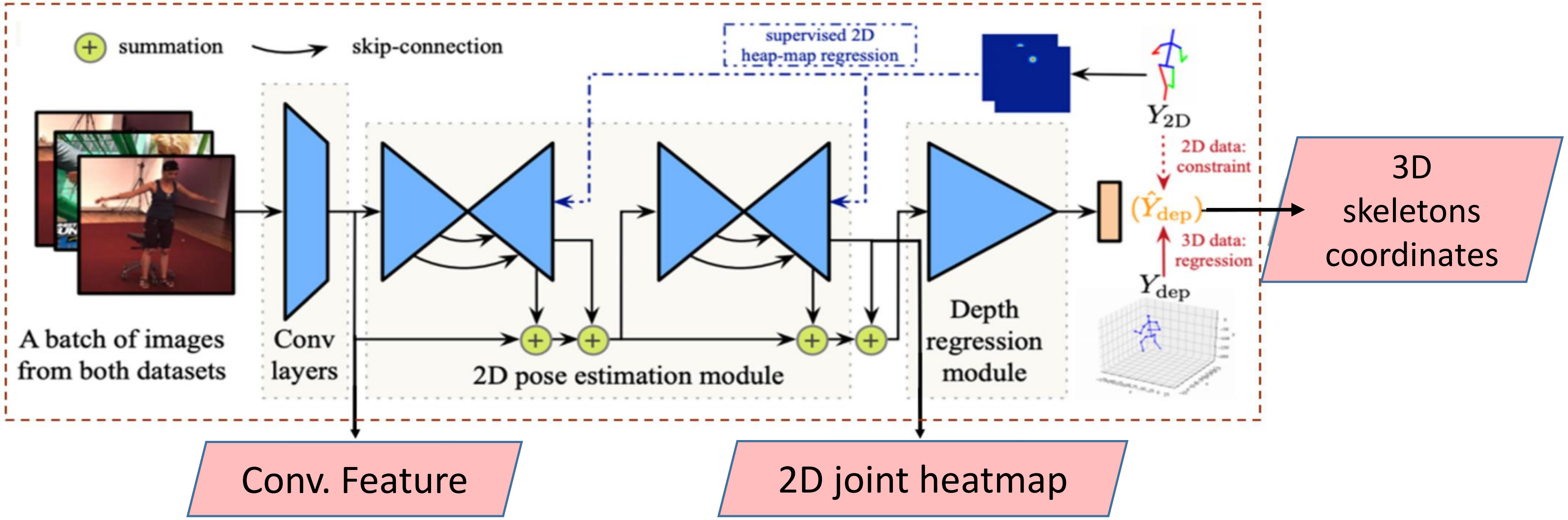
Main components:

1. Pose module
2. Skeleton module
3. Appearance module
4. Aggregation module

Pose module

- It uses stacked hourglasses networks to estimate 2D heatmaps and then uses conv. features and heatmaps to estimate depths for each body joint.
- It takes RGB video as input and outputs 3D skeleton coordinates, conv. features and 2D joint heatmap.

Pose Estimation network



3D skeletons coordinates

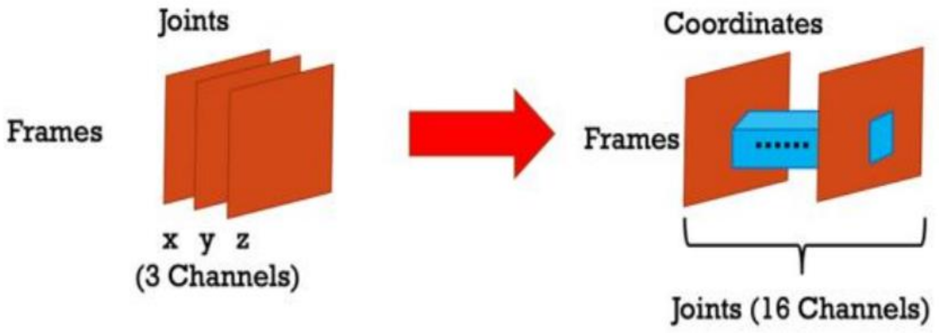
Conv. Feature

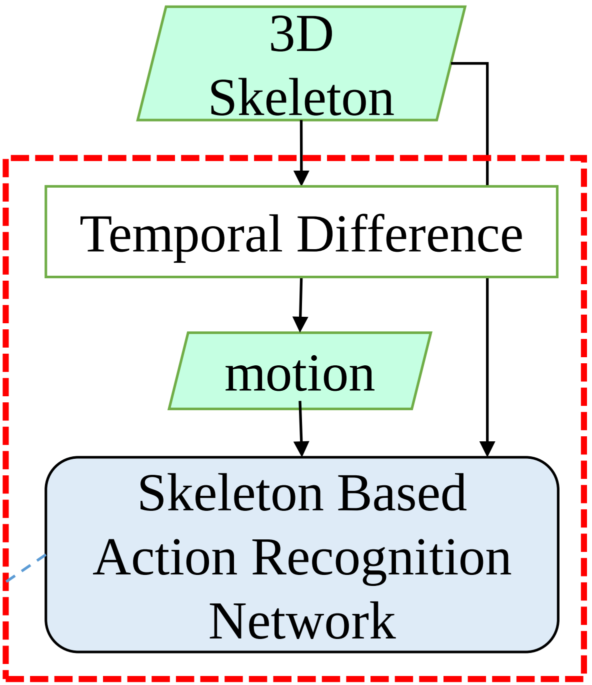
2D joint heatmap

Skeleton module

- Skeletal data has the benefits of insensitive to illumination changes and cluttered background, and is more correlated to human actions.

Change channel dimension from *coordinates* *to* *joints*





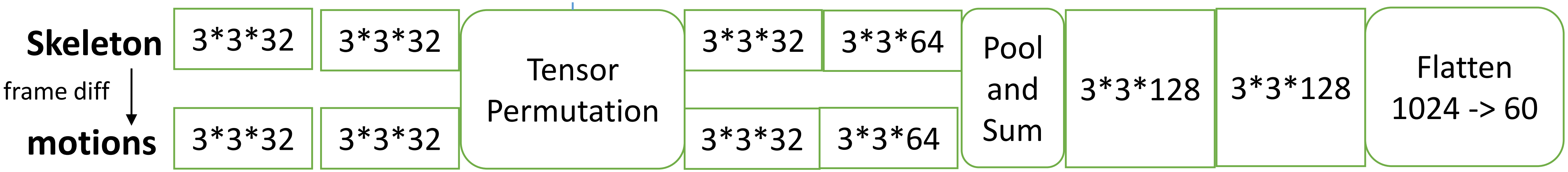
3D Skeleton

Temporal Difference

motion

Skeleton Based Action Recognition Network

Skeleton Module



Skeleton

frame diff

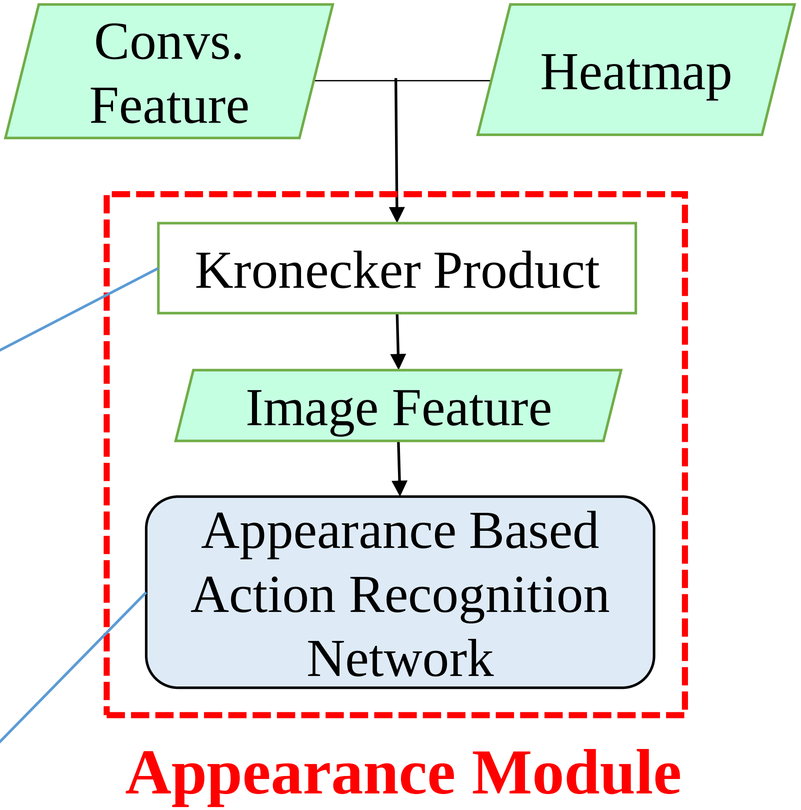
motions

3*3*32 | 3*3*32 | Tensor Permutation | 3*3*32 | 3*3*64 | Pool and Sum | 3*3*128 | 3*3*128 | Flatten 1024 -> 60

3*3*32 | 3*3*32 | | 3*3*32 | 3*3*64

Appearance module

- It takes convolutional features and bodyjoint heatmaps as the input to predict actions



Convs. Feature

Heatmap

Kronecker Product

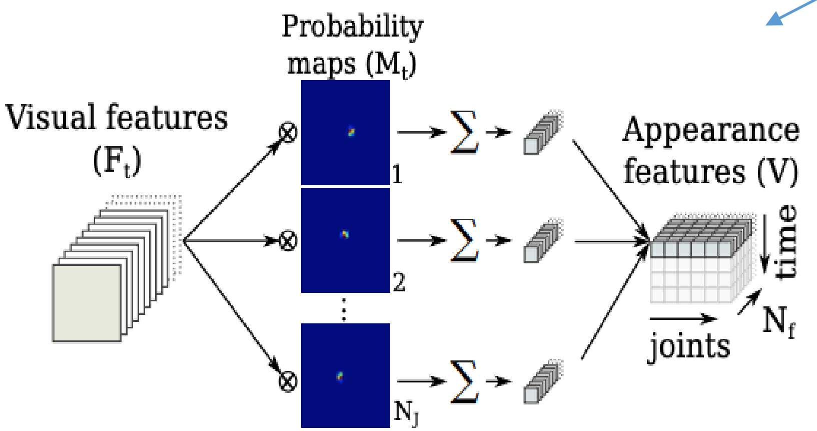
Image Feature

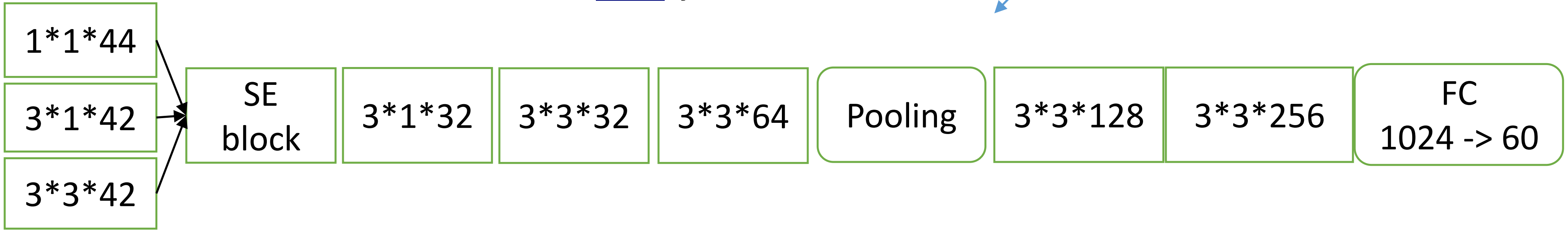
Appearance Based Action Recognition Network

Appearance Module

Kronecker product

To extract appearance features around joints





1*1*44

3*1*42

3*3*42

SE block

3*1*32

3*3*32

3*3*64

Pooling

3*3*128

3*3*256

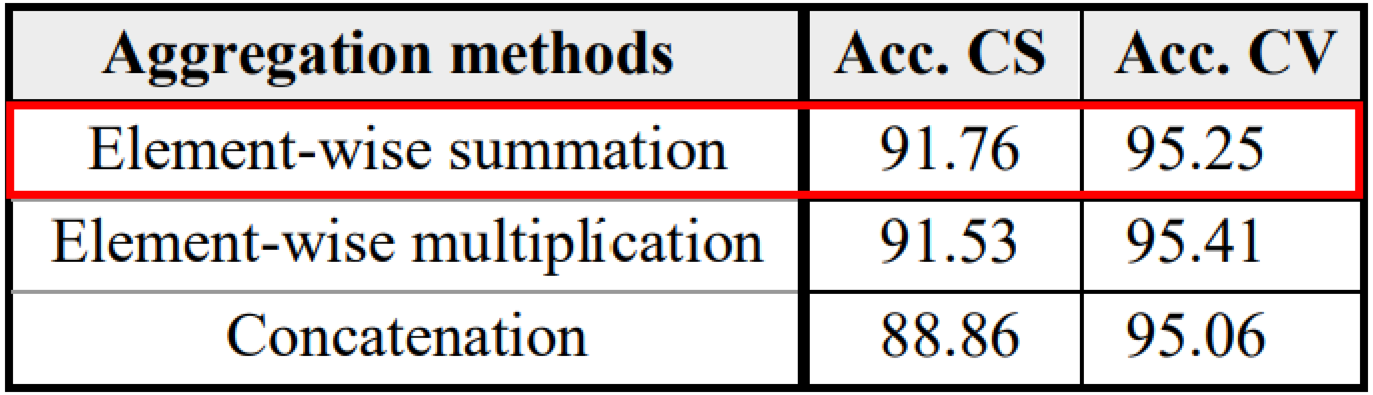
FC 1024 -> 60

SE block (Squeeze-and-Excitation): learn the weights of each channel.

Aggregation module

- It takes the prediction result from skeleton module and appearance module to produce the final prediction.
  - Element-wise summation
  - Element-wise multiplication
  - Concatenation
    - uses convolution layers to extract fused features and predict the final classification result by fully connected layers.

Acc. : accuracy
CS: cross subject
CV: cross view

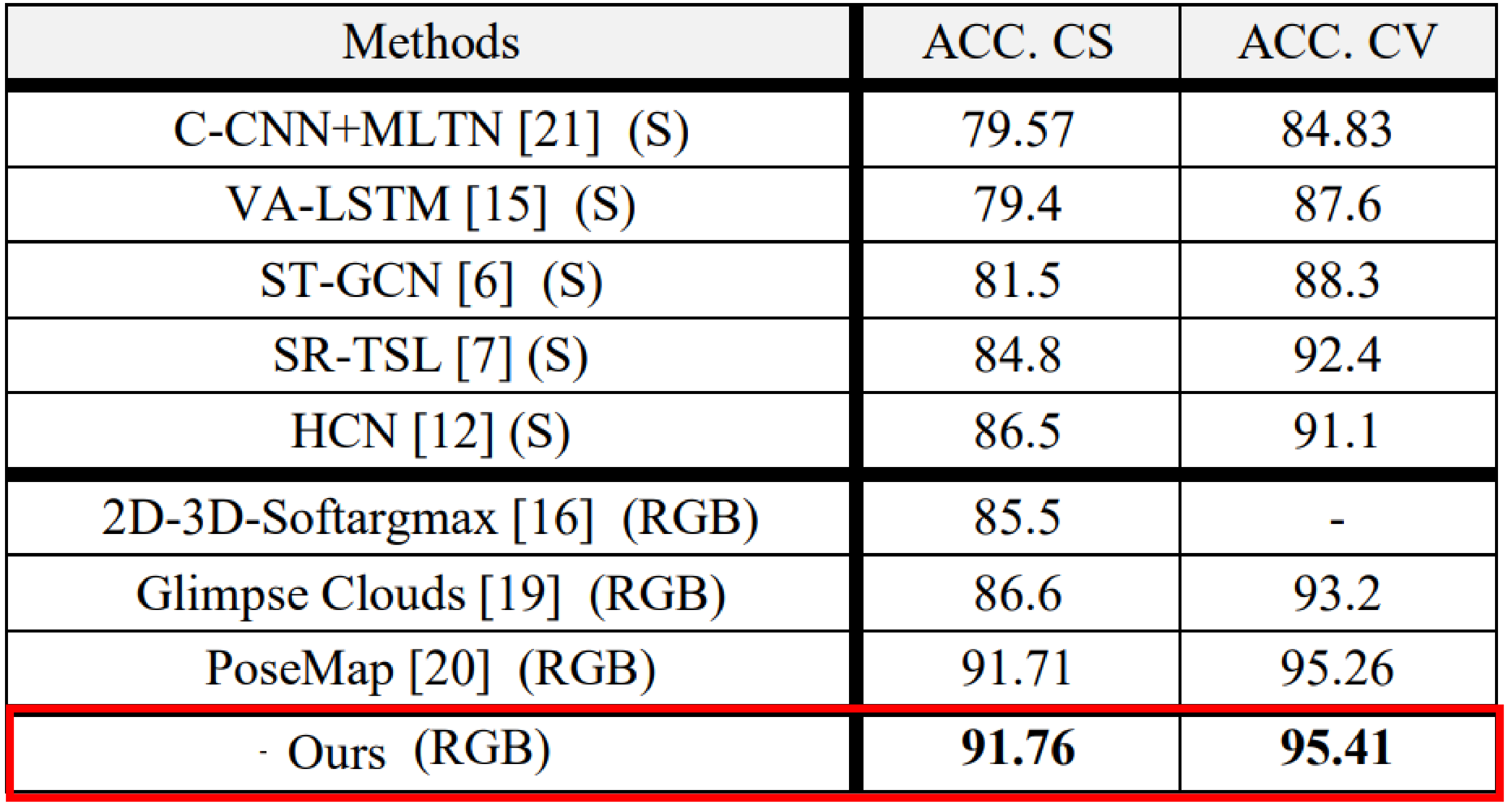


Aggregation methods	Acc. CS	Acc. CV
Element-wise summation	91.76	95.25
Element-wise multiplication	91.53	95.41
Concatenation	88.86	95.06

Experiment Result

Comparison with state-of-the art

NTU RGB+D dataset



Methods	ACC. CS	ACC. CV
C-CNN+MLTN [21] (S)	79.57	84.83
VA-LSTM [15] (S)	79.4	87.6
ST-GCN [6] (S)	81.5	88.3
SR-TSL [7] (S)	84.8	92.4
HCN [12] (S)	86.5	91.1
2D-3D-Softargmax [16] (RGB)	85.5	-
Glimpse Clouds [19] (RGB)	86.6	93.2
PoseMap [20] (RGB)	91.71	95.26
· Ours (RGB)	**91.76**	**95.41**

Acc. : accuracy
CS: cross subject
CV: cross view

Thanks for your attention