Action Recognition for People Monitoring

INRIA Sophia Antipolis – **STARS team**

Institut National Recherche Informatique et Automatisme

Francois.Bremond@inria.fr

http://www-sop.inria.fr/members/Francois.Bremond/

CoBTeK,

Nice University Hospital

What is
Human Action Recognition?

Introduction

An example

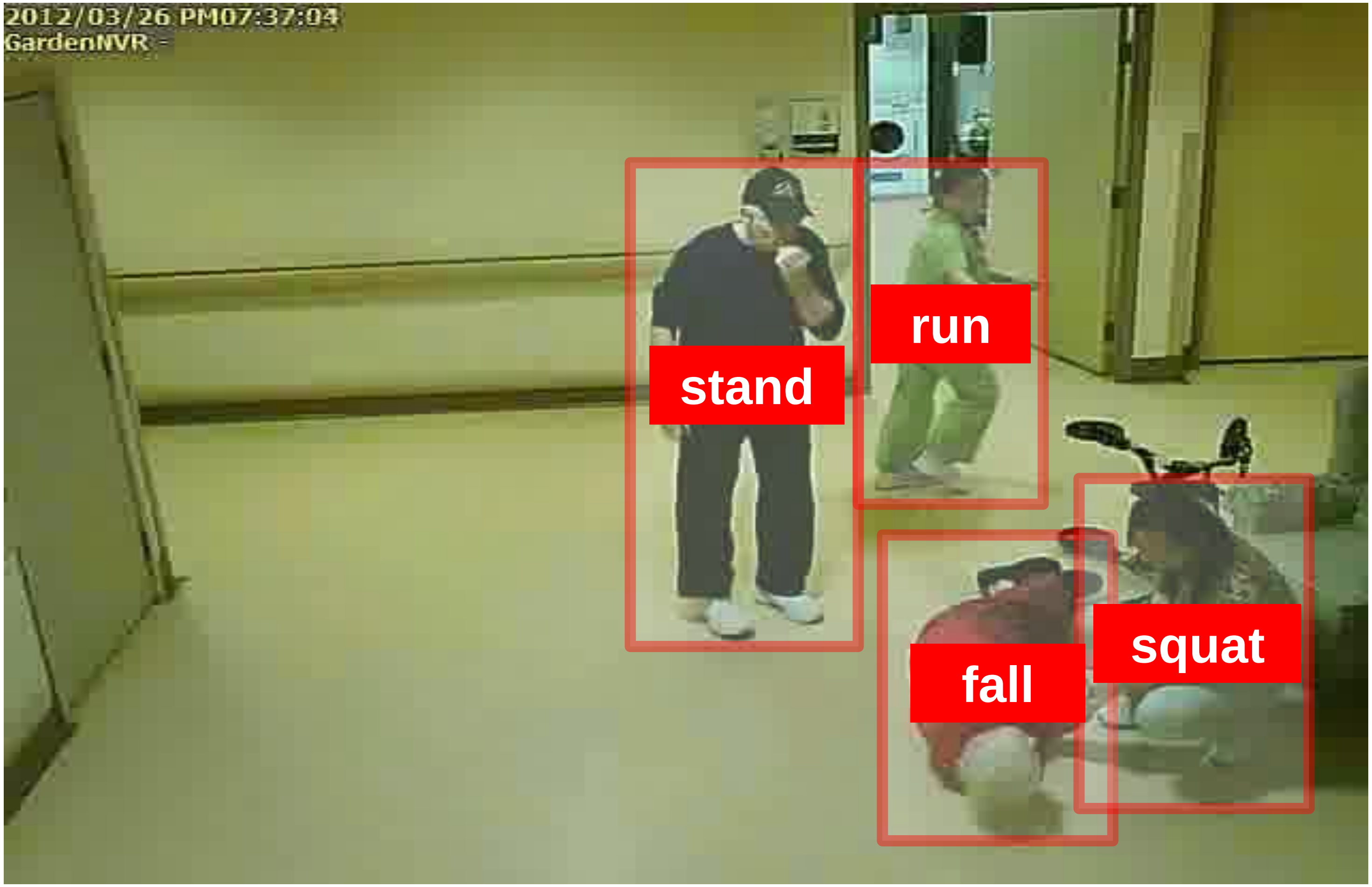
- What does action recognition involve?

An example

- Object Detection: Are they Human?

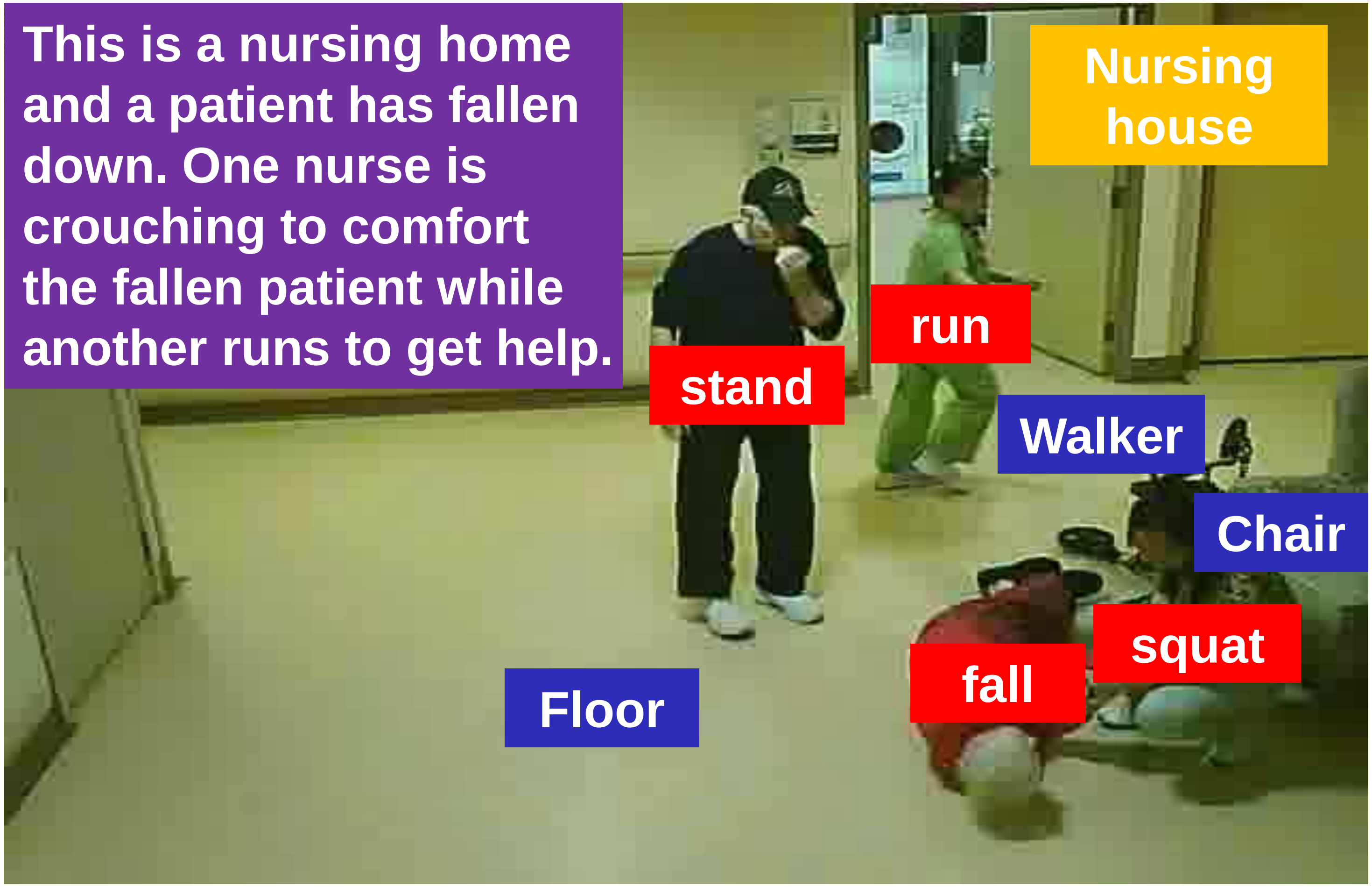
An example

- Action Recognition: What are they doing?



An example

- Full semantic understanding

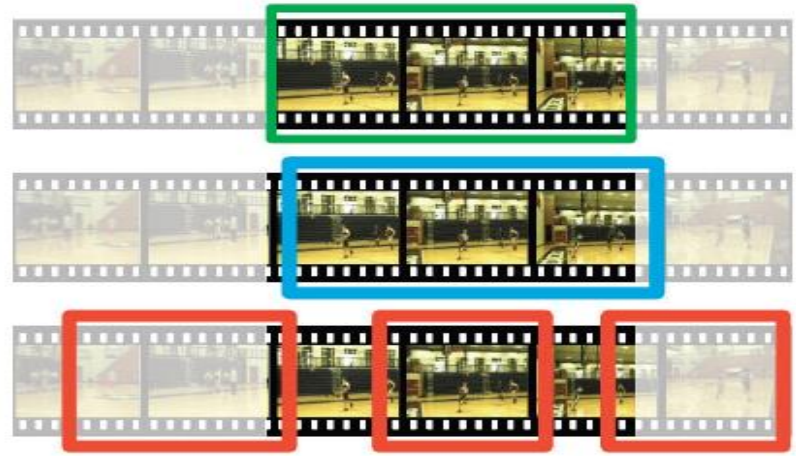


Introduction: Video Understanding

Activity Recognition

Classification of Video Clips into
Pre-defined Activity Categories

Activity Detection



Localizing Pre-defined Activities
Temporally in untrimmed Videos

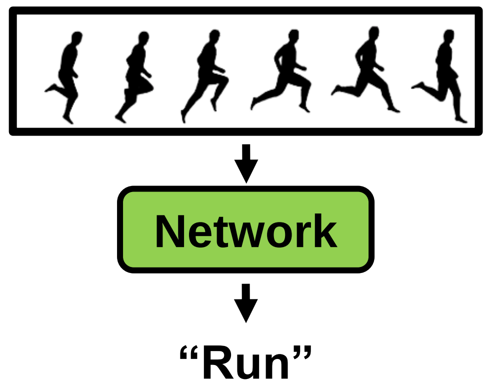
Understanding Activities in trimmed videos is thus important!

Srijan Das - Rui Dai

Action Recognition

Video classification task:

Input: A clipped video (a sequence of frames)



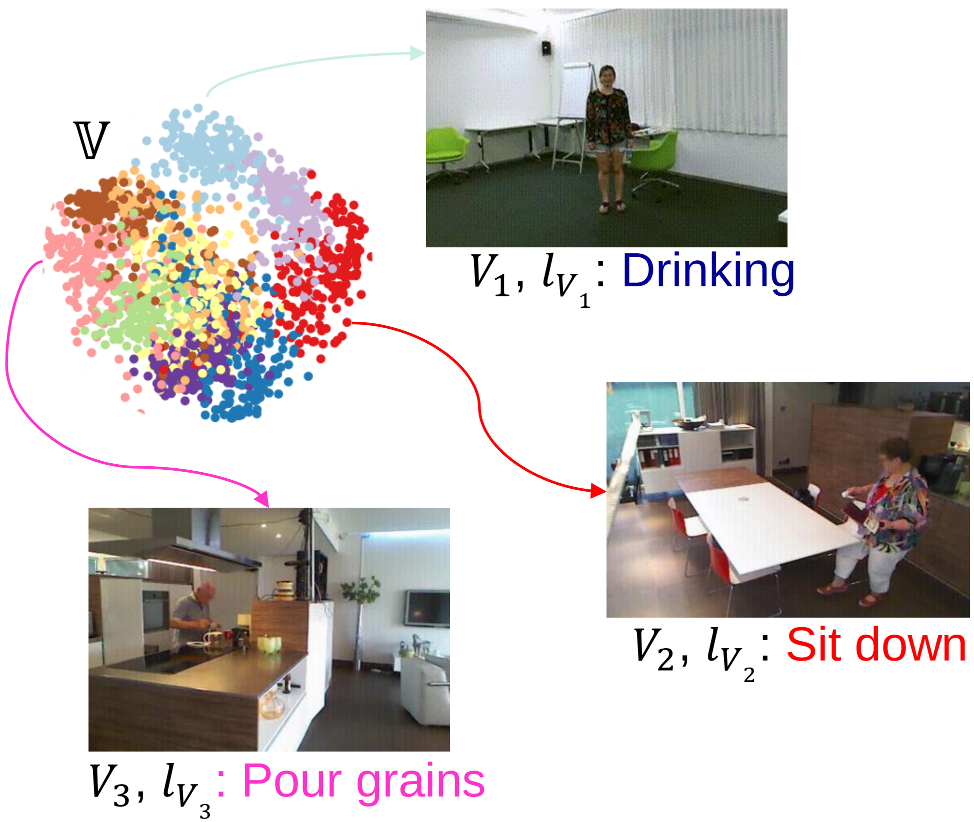
V_i

e_i $embedding$

l_{Vi}: Run

Output: An action label

Action Recognition



Problem statement:

- Let's assume that we have a set of videos $\mathbb{V}$ and a set of corresponding action labels L.
- We assume that each video $V \in \mathbb{V}$ contains only one action l_V .
- Thus the goal of action recognition problem is to predict the label l_V based on a video representation V.

To learn **compact** and **discriminative** video representations for the task of activity classification.

Outline : Human Action Recognition

- Introduction
  - Action Recognition Datasets
  - Toyota Smart-Home

- Different Modalities
  - RGB, Optical Flow, 3D Poses

- 4 Attention Mechanisms for Action Recognition
  - 3D Pose guided Attention : P-I3D, Separable STA, VPN, VPN++
  - Transformers

Section 1

Action Recognition Datasets

Activity Recognition for Daily-living activities

Web and movies datasets:

(Kinetics, UCF101, ActivityNet,…)

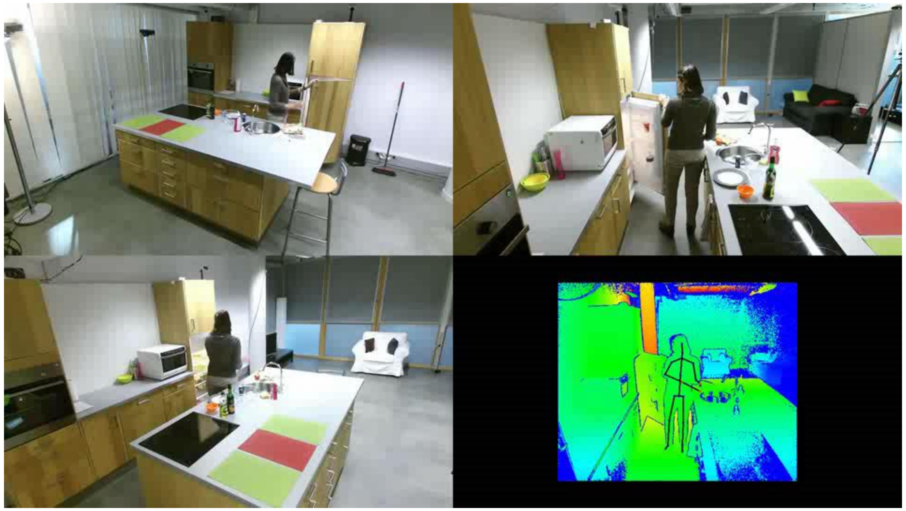
- Large number of classes
- High inter-class variation
- Camera motion
- Different environments
- Short actions

Different challenges compared to Fine-grained video datasets:

(Toyota smart home, Dahlia, NTU,…)

- **Real-time** recognition
- High intra-class variation
- **Low inter-class** variation
- Same environment, background
- Long and Composed actions

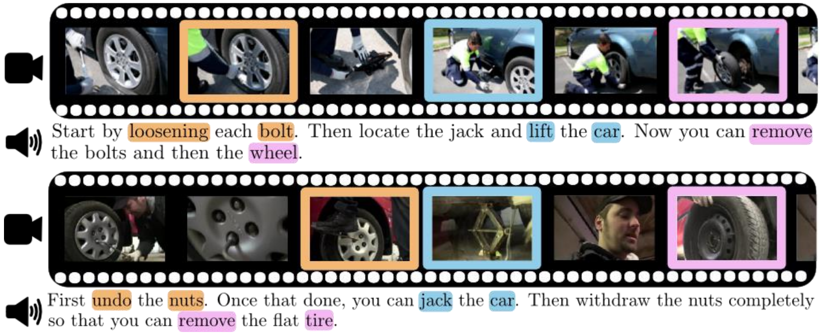
Need to model spatio-temporal relations



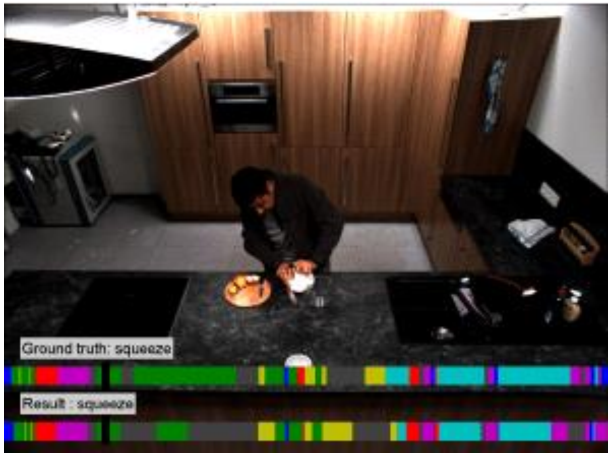
Categories of Fine-grained Action Videos

Specific Sports

Instruction videos



Cooking



Ego-centric

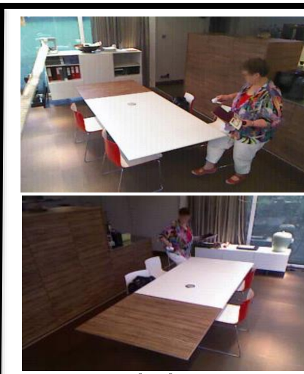
Challenges : Activities of Daily Living (ADL)

How to handle time?

Use microwave

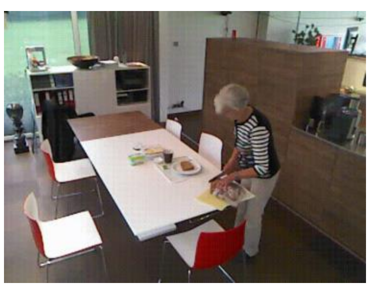
How to learn view-invariant representations?

pour grains

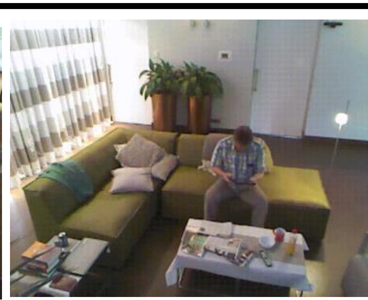


sitdown

How to learn representations for recognizing fine-grained actions? Interacting with objects?

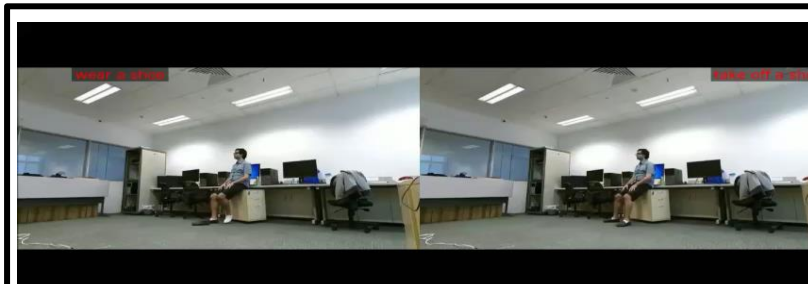


Cut bread



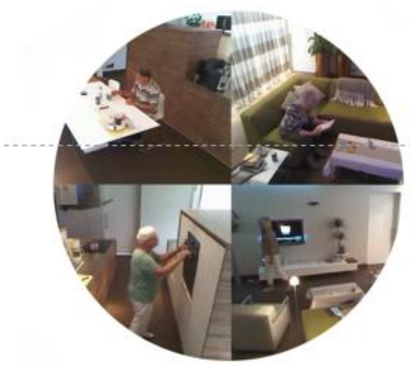
Use tablet

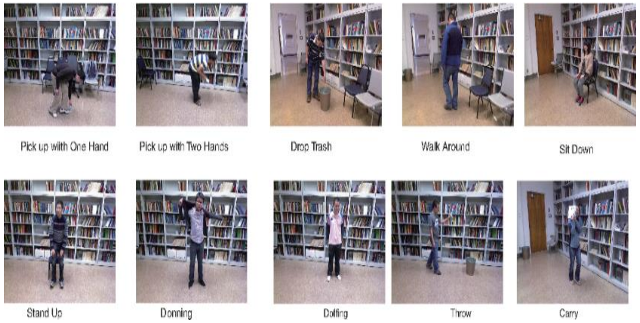
How to disambiguate similar appearance actions?



Wear a shoe *Taking off a shoe*

Daily-living Dataset Description





- **NTU RGB-D (NTU-60 / NTU-120)** dataset, one of the largest available human activity dataset.
  - ❑ 58K / 114K **videos**
  - ❑ 60 / 120 **actions**
  - ❑ 40 / 106 **subjects**
  - ❑ 80 / 155 **views**

- A human activity dataset possessing real-world challenges: **Toyota Smarthome** dataset (TS/TSU).
  - ❑ 16.1K **videos**
  - ❑ 31 **actions**
  - ❑ 18 **subjects**
  - ❑ 7 **views**

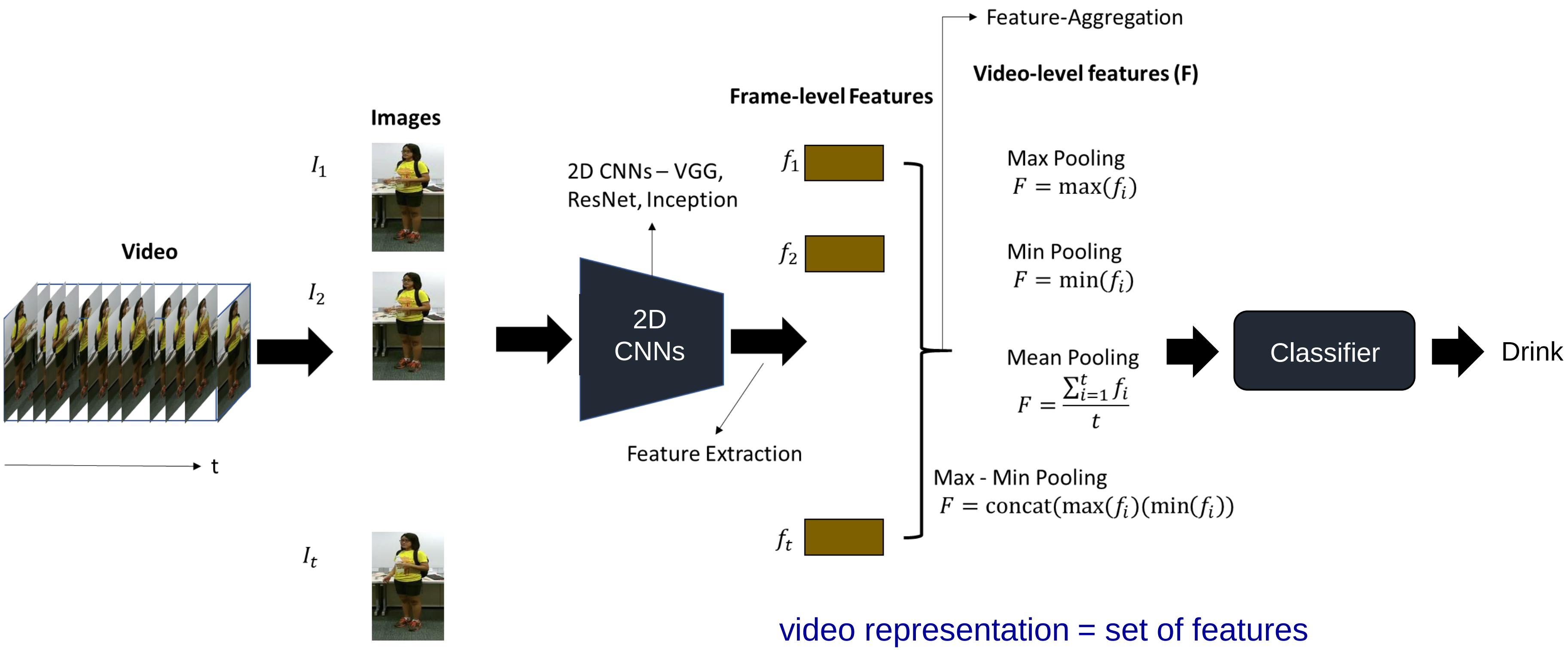
- A small-scale object-interaction human action recognition dataset: **Northwestern-UCLA Multiview Action 3D** dataset (NUCLA).
  - ❑ 1194 **videos**
  - ❑ 10 **actions**
  - ❑ 10 **subjects**
  - ❑ 3 **views**

Section 2

RGB based Deep Networks for Action Recognition

CNN: Video Classification

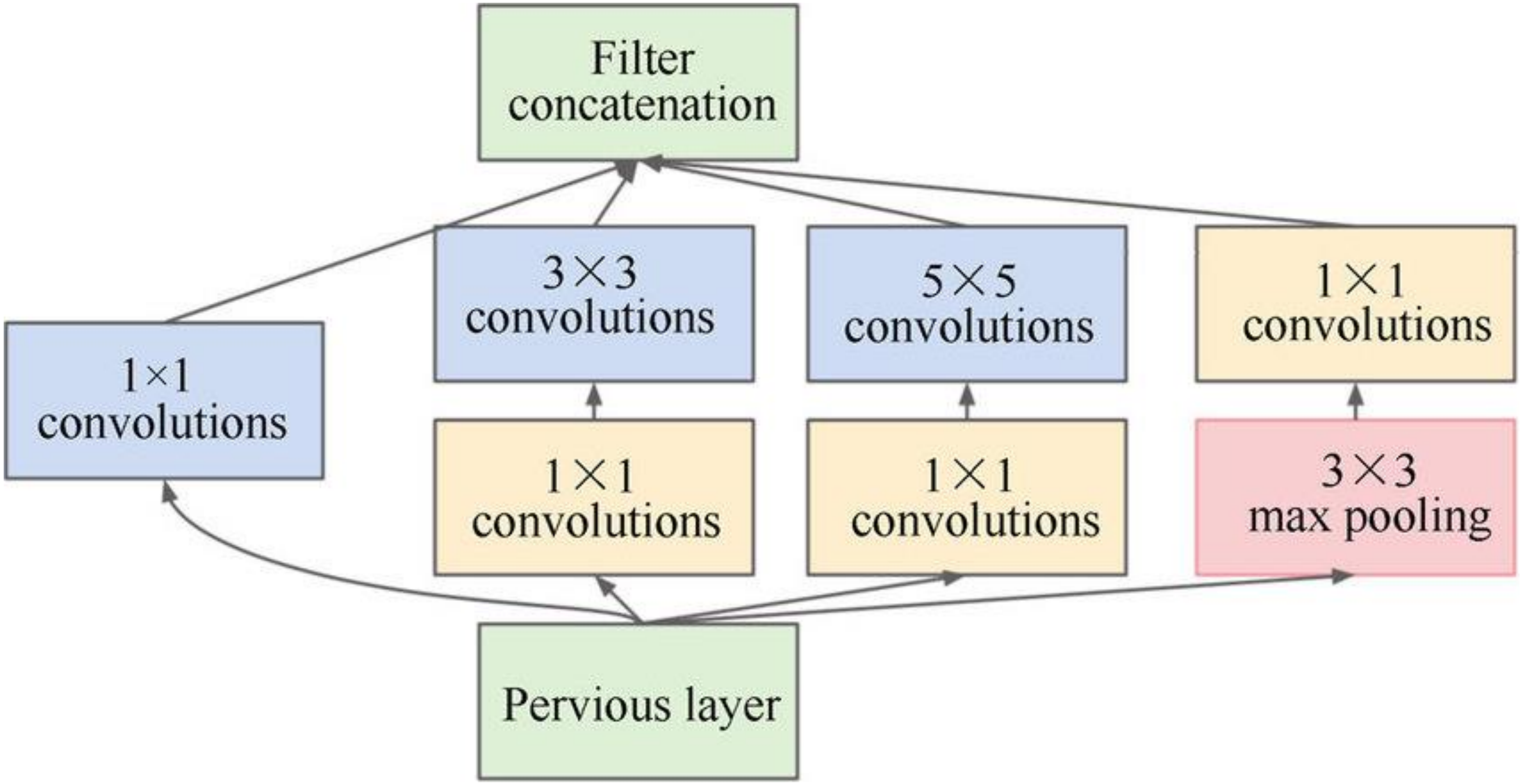
2D CNN: feature extraction + classification



video representation = set of features

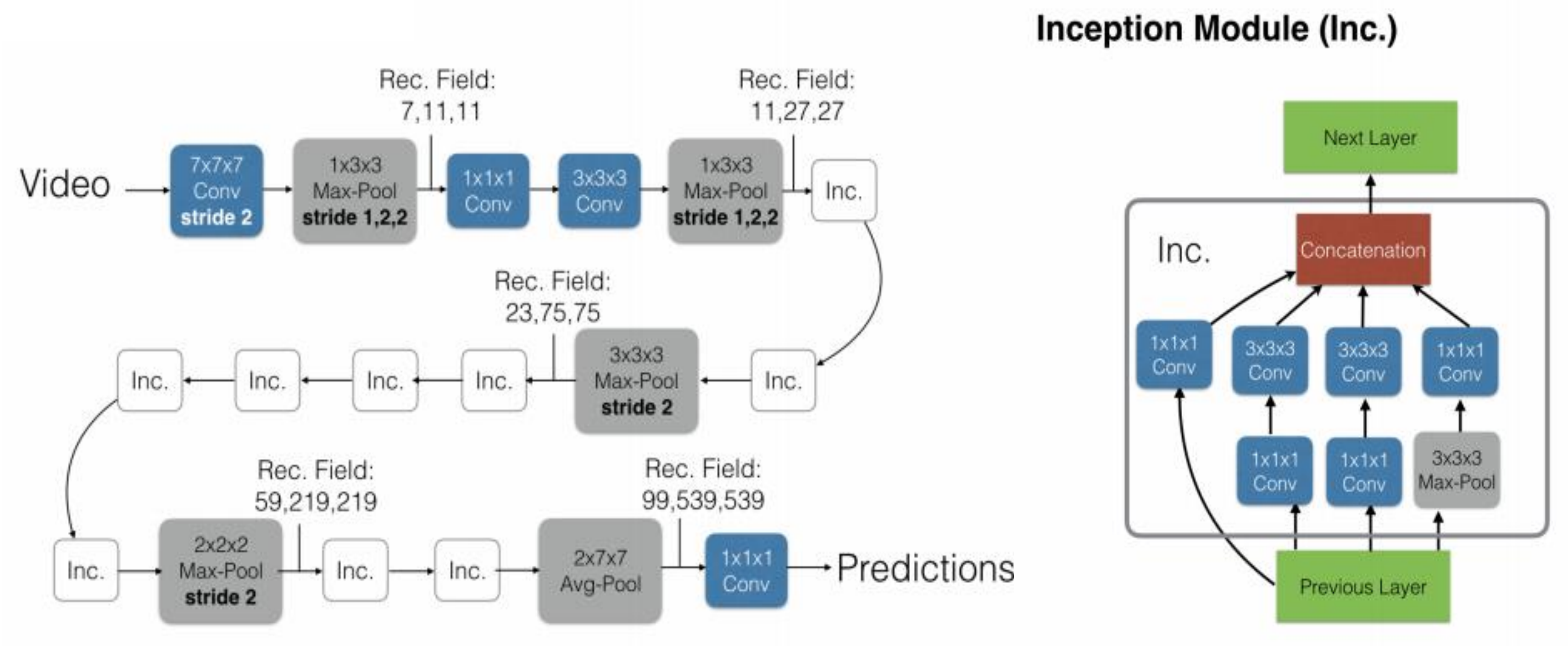
CNN: Video Classification

Inception Module



CNN: Video Classification

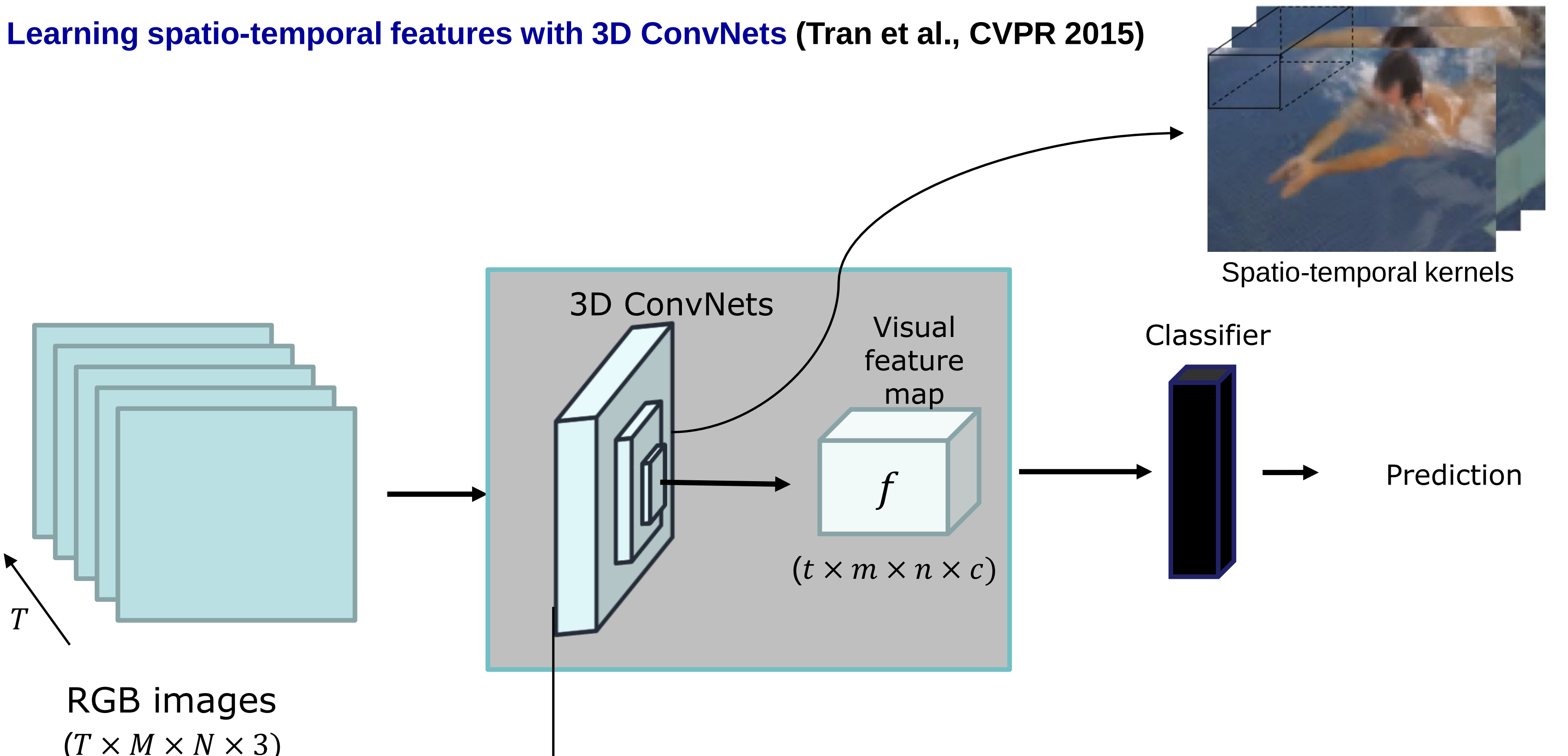
I3D Network [CVPR'17]



Same structure as 2D CNN (GoogleNet!)

Activity Recognition for Daily-living activities
3D CNN: Approaches based on RGB

Learning spatio-temporal features with 3D ConvNets **(Tran et al., CVPR 2015)**



3D ConvNets like I3D, ResNext3D-101, slow-fast network, X3D

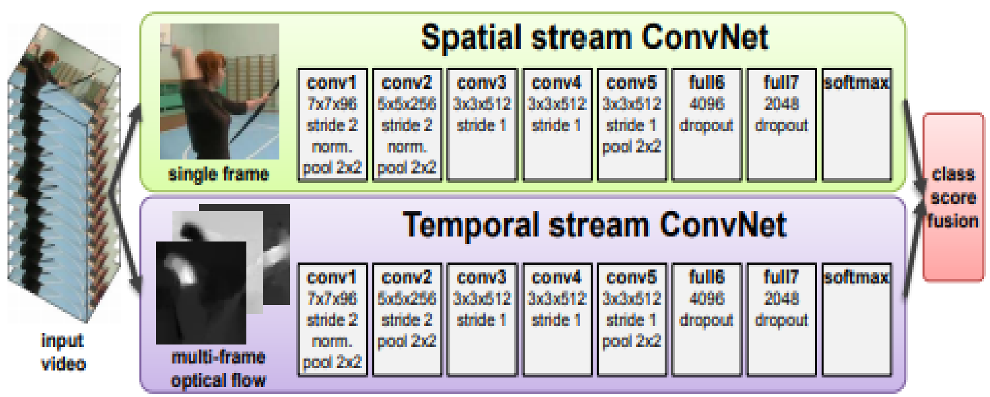
Section 3

Modalities

Two-stream Network [NIPS'14]

- **Using multiple modalities as input!**
- **RGB**: One image randomly sampled from the video. (<u>Spatial</u>: encodes object/appearance information)
- **Flow**: *2L* optical flow images from a video. (<u>Temporal</u>: encodes short-term motion)

$224 \times 224 \times 3$



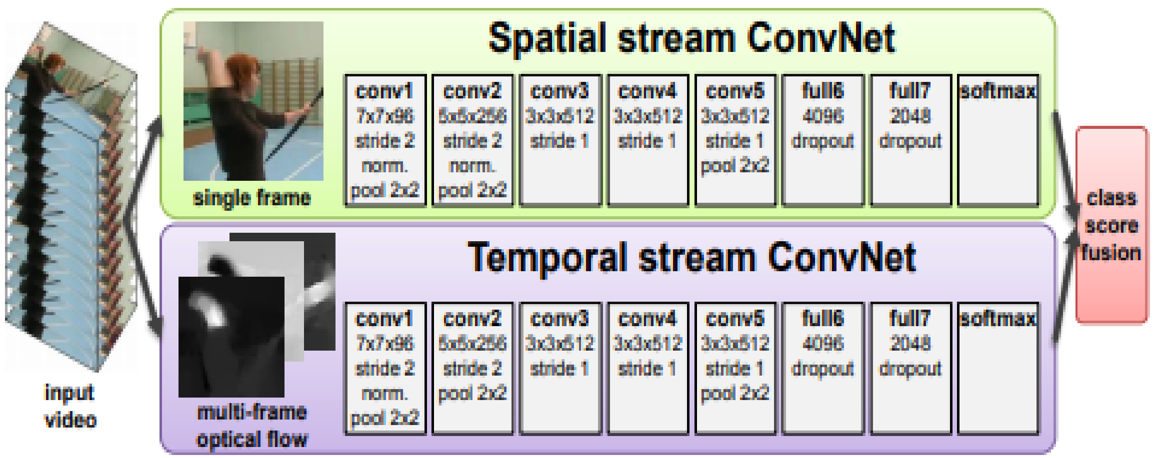
Just Fuse the Score

$224 \times 224 \times 2L$

Two-stream Network [NIPS'14]

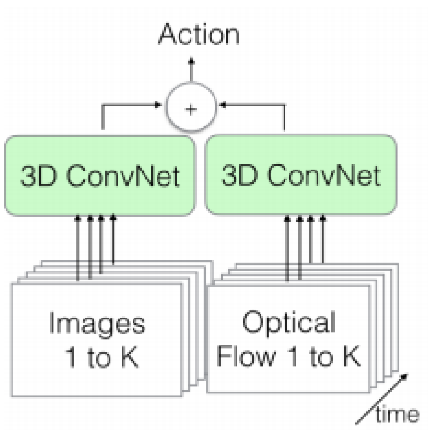
Drawbacks:

- Temporal information is not encoded along space.
- Long-term motion is ignored!



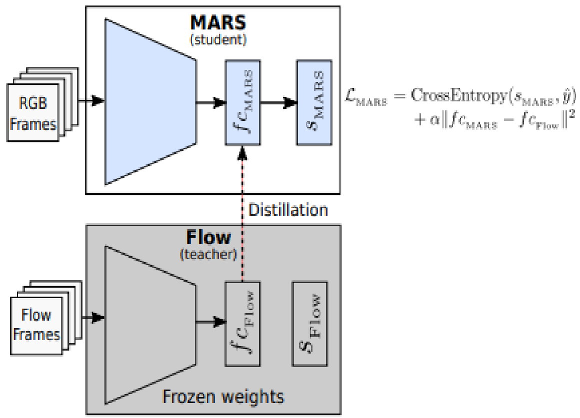
Multi-Modal : Video Classification

State-of-the-Art : RGB + Optical Flow



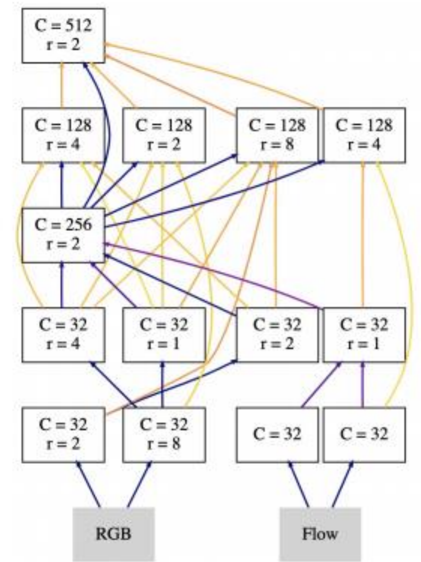
(a) Two-streams in I3D (*Carreira et al., CVPR 2017*).

Late fusion of RGB and Optical Flow streams.



(b) Teacher-student network in MARS (*Crasto et al., CVPR 2019*).

Knowledge distillation from Optical flow stream to RGB stream.



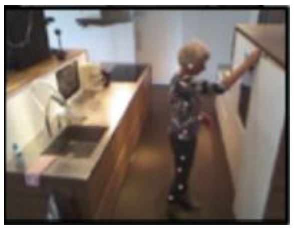
(c) NAS in AssembleNet (*Ryoo et al., ICLR 2020*)

NAS to combine Optical flow stream to RGB stream.

Less attention has been given to combine Poses with RGB

Multi-Modal : Video Classification

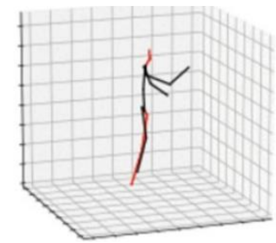
Different input modalities : RGB based and others: Audio, Text, Depth image, Bio Signals (EEG, ECG, EDA, HR) …



RGB

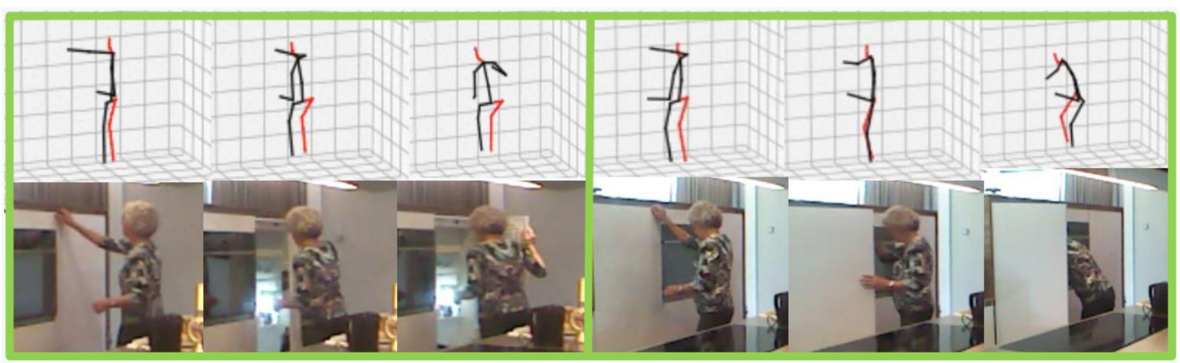
Depth

Optical Flow

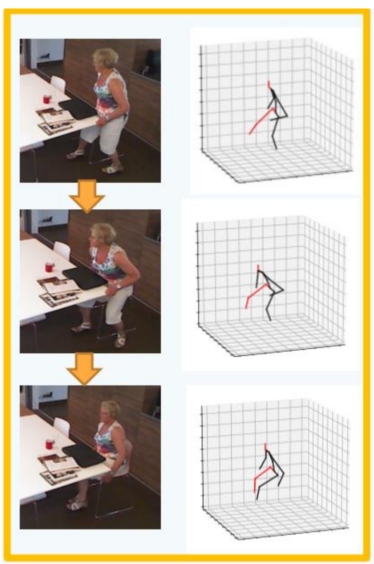


3D skeleton

Complementary Nature



Open fridge Open cupboard



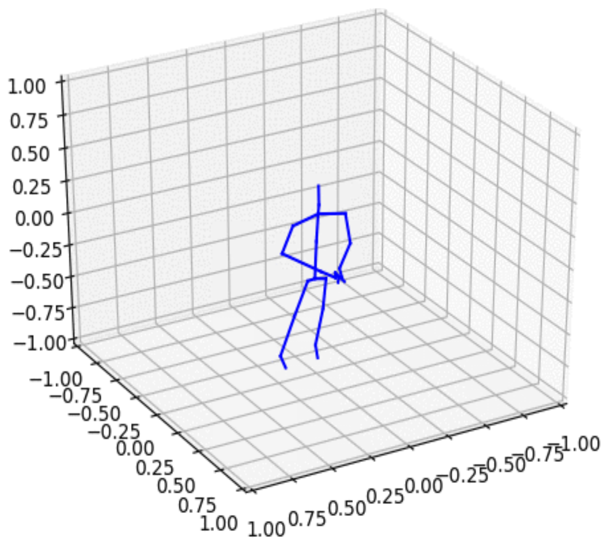
Sit down

Filtering the noisy appearance patterns

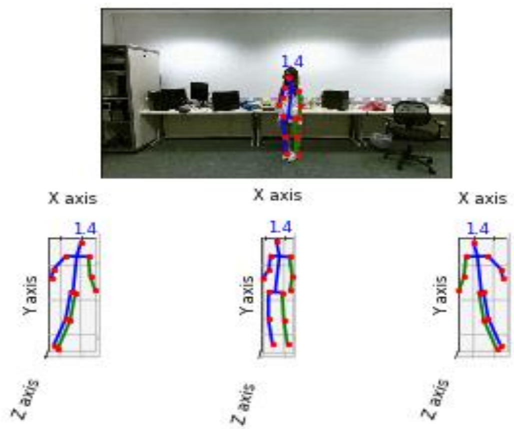
Help capturing the body motion

Estimation of 3D Poses - Skeletons

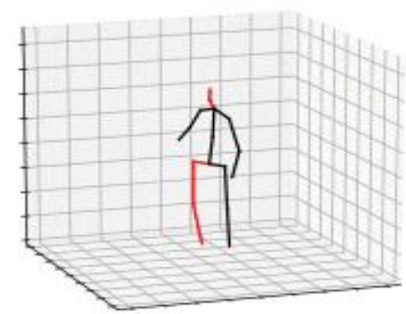
Poses computed from RGB images



3D poses captured using Kinect (*Shotton et al., CVPR 2009*)



3D poses computed using LCRNet++ (*Rogez et al., TPAMI 2019*)

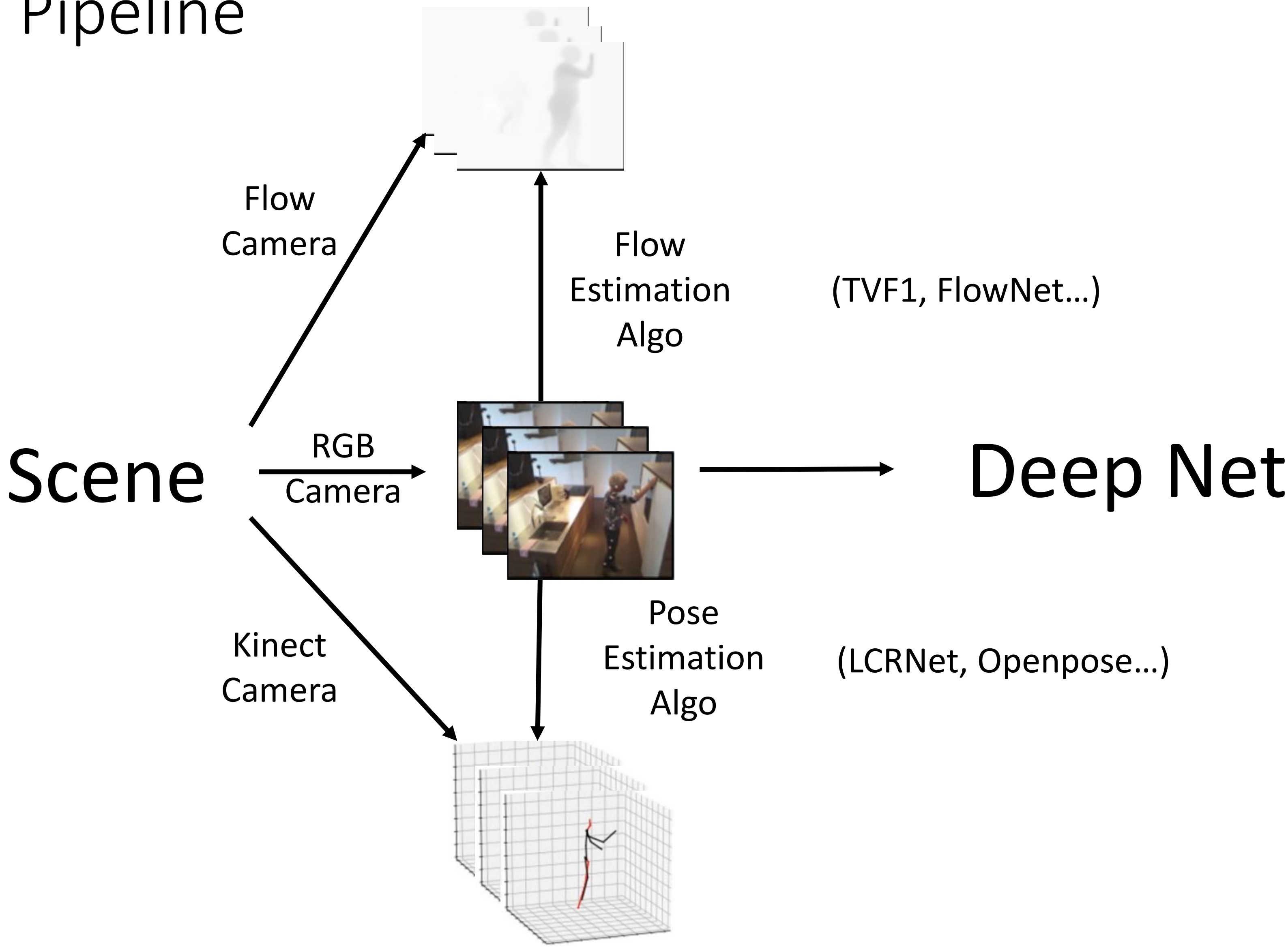


3D poses computed using VideoPose3D (*Pavllo et al., CVPR 2019*)

Robust to **illumination changes**, **view-invariant**, and provide **geometric information** associated to actions.

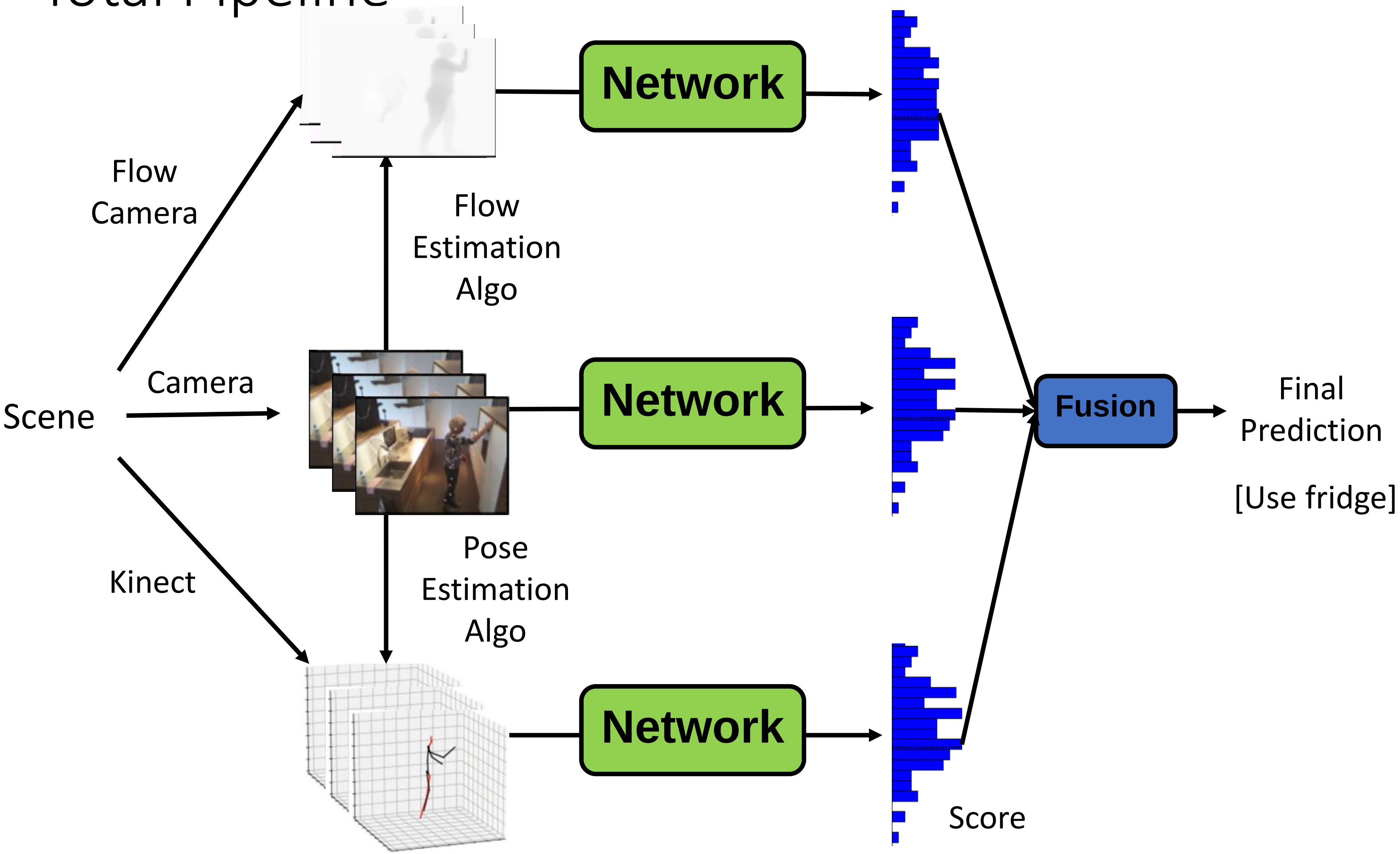
Multi-Modal : Video Classification

Pipeline



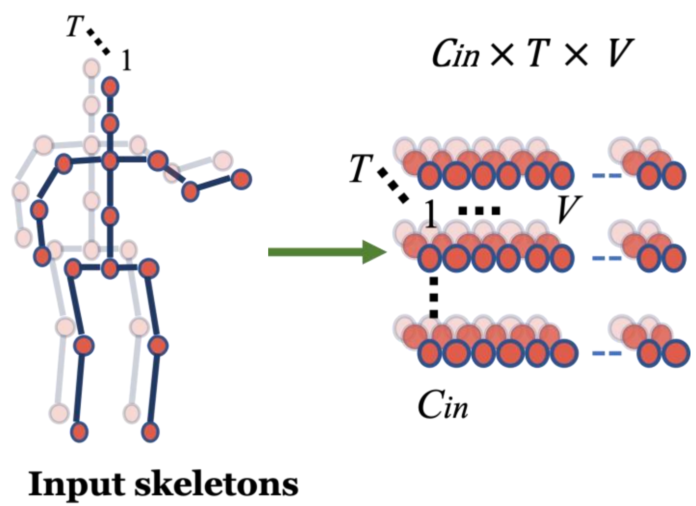
Multi-Modal : Video Classification

Total Pipeline

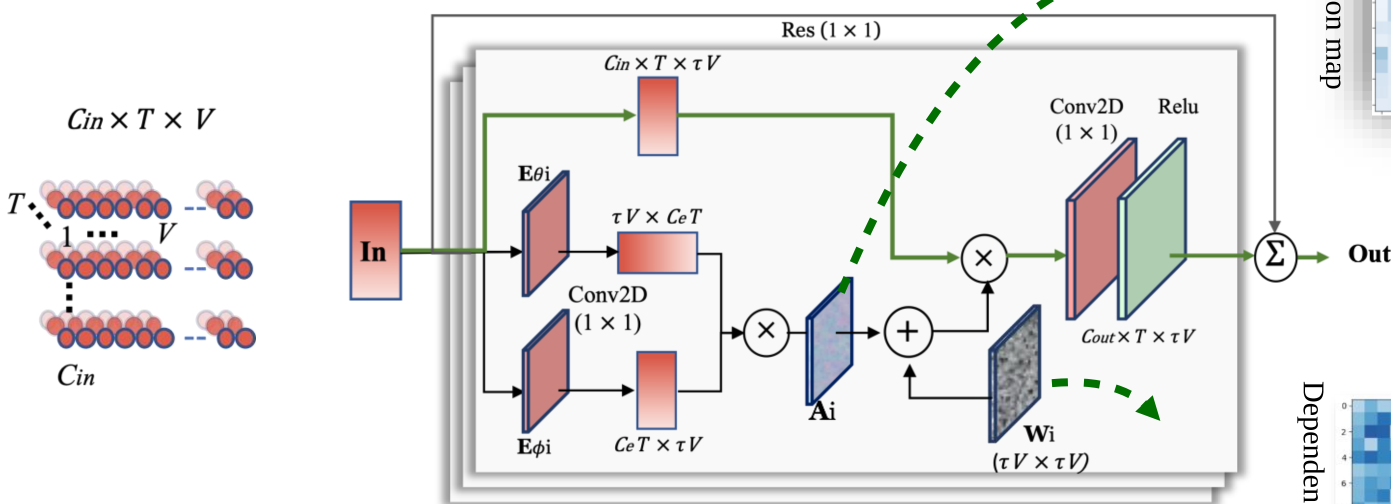


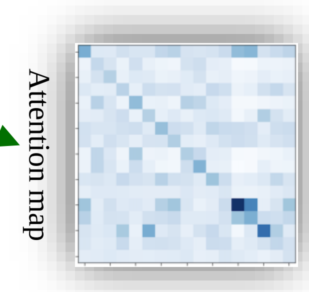
Skeleton based Action Recognition

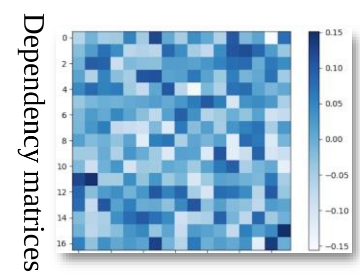
Skeleton modeling



Approach: Spatial processing (S-LSU)



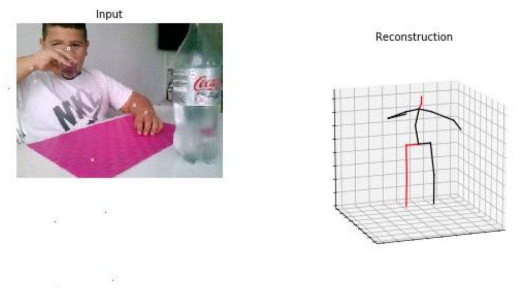




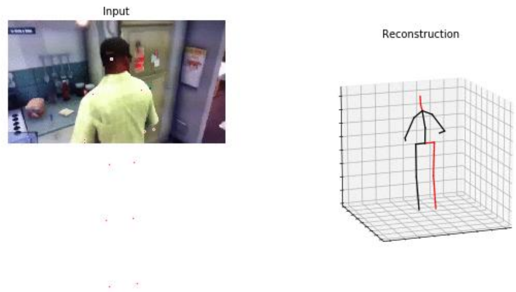
$$\mathbf{f}_{\mathbf{out}} = \text{S-LSU}(\mathbf{f}_{\mathbf{in}}) = \sum_{i=1}^{N} \mathbf{E}_i \cdot \left(\mathbf{f}_{\mathbf{in}} \times (\mathbf{W}_i + \mathbf{A}_i)\right)$$

Pre-training Dataset: Posetics

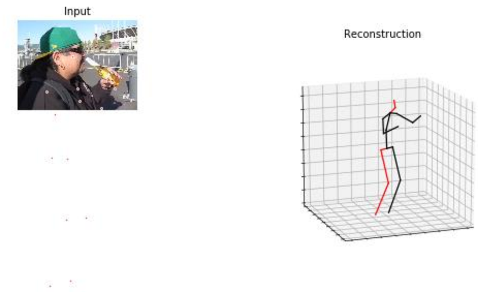
Real-world actions with skeletons



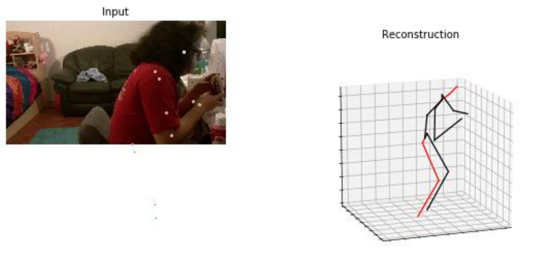
Drinking



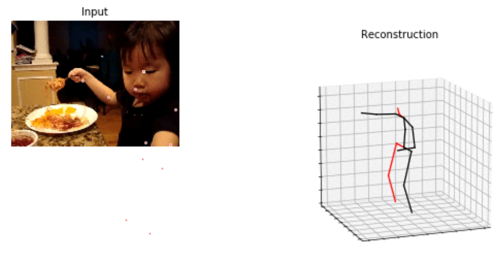
Drinking



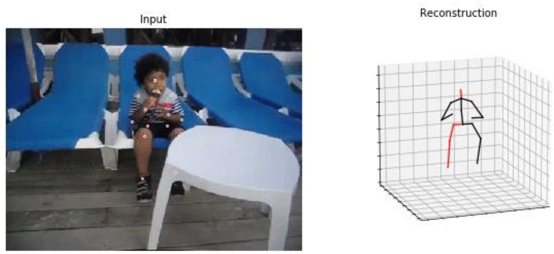
Drinking



Eating

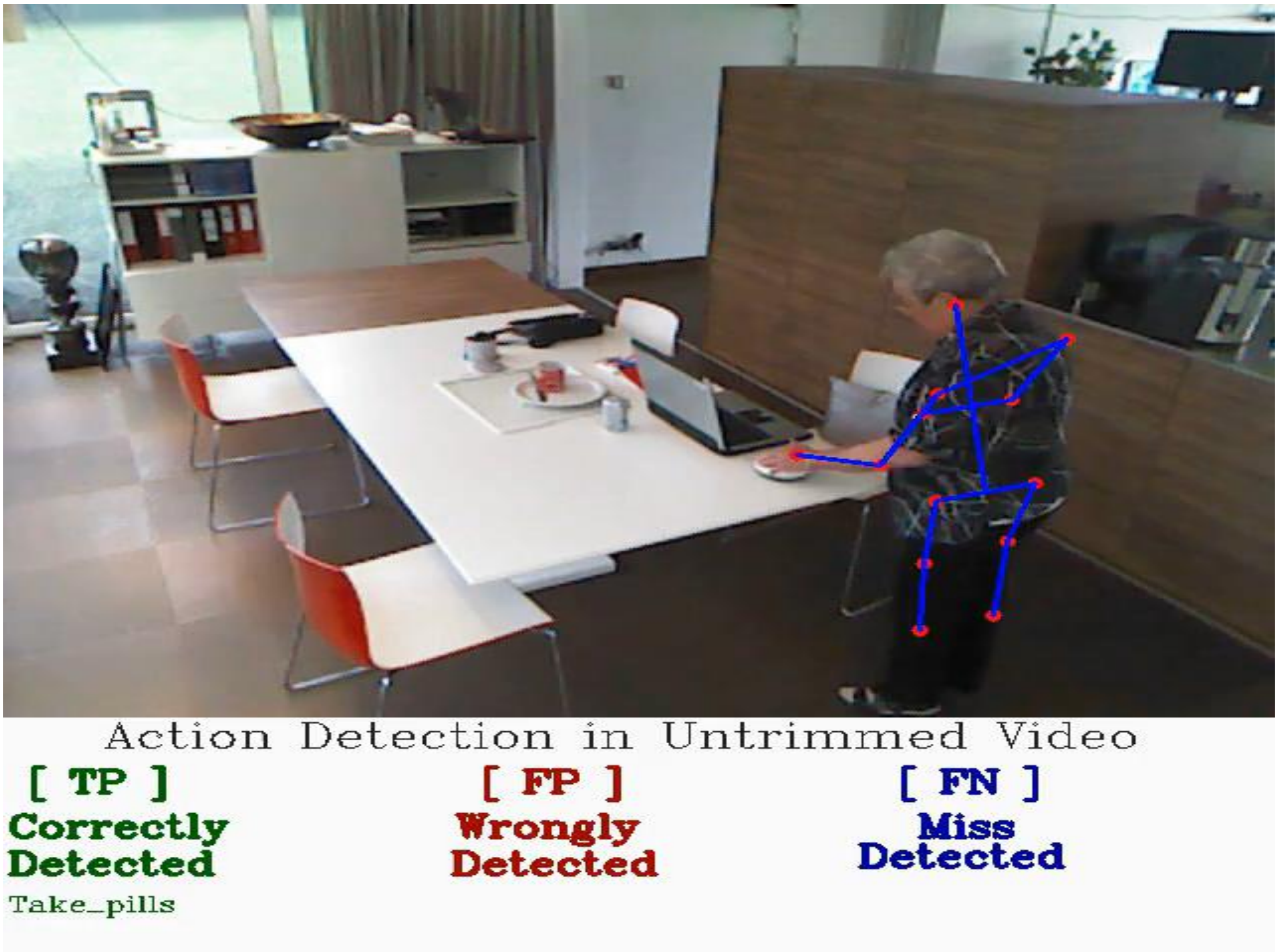


Eating



Eating

Toyota Smart-Home
Large scale daily living dataset

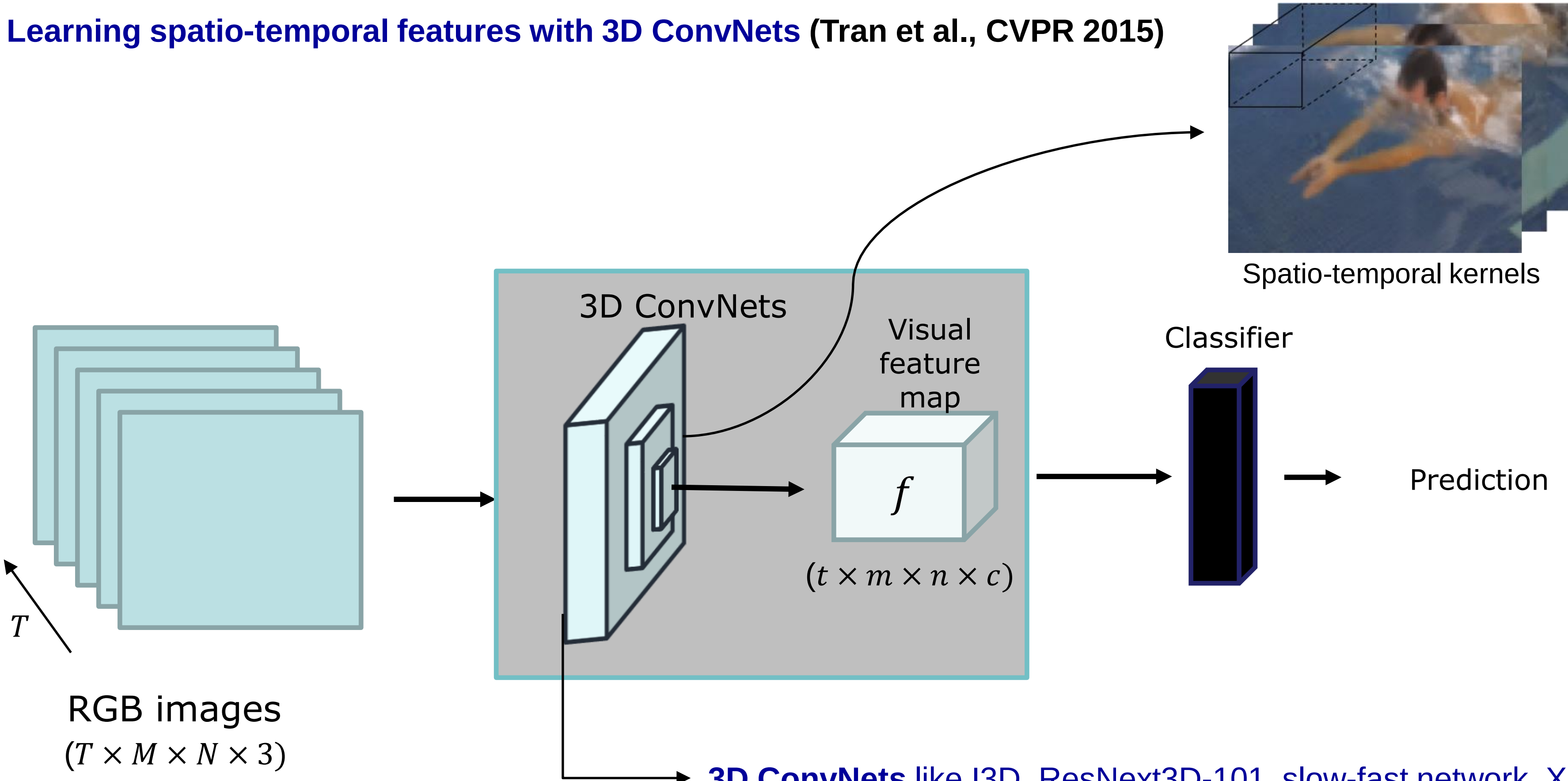


Section 4

RGB based Deep Networks for Action Recognition with Attention

Activity Recognition for Daily-living activities
3D CNN: Approaches based on RGB

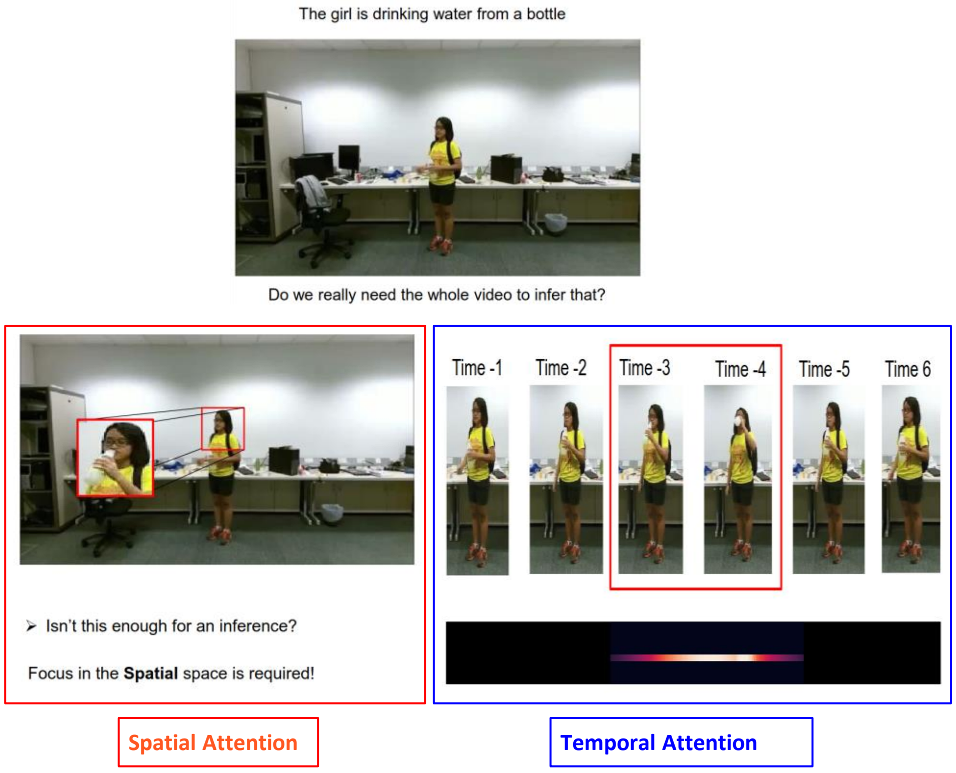
Learning spatio-temporal features with 3D ConvNets **(Tran et al., CVPR 2015)**



3D ConvNets like I3D, ResNext3D-101, slow-fast network, X3D

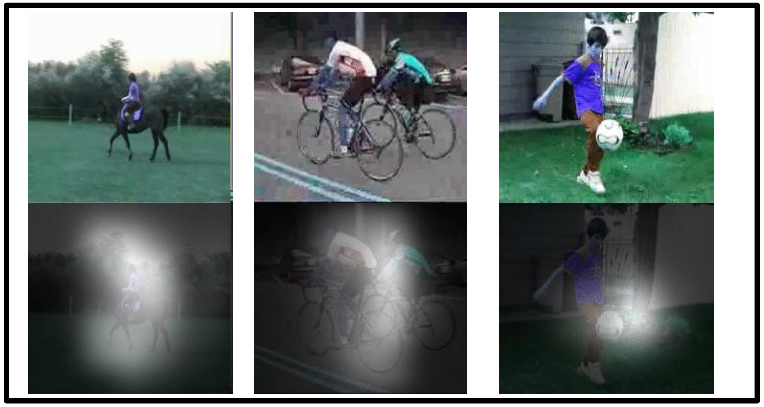
Several Attention Mechanism:

- **Primary purpose of Attention:** To imitate human visual cognitive systems and focus on essential features. (or) <u>Learn how to pick relevant information from input data</u>
- **Key Idea:** To focus on the significant parts in an image and suppress unnecessary information.
- **CNN with Attention:** are used to make CNN learn and focus more on the important information, rather than learning non-useful background information.

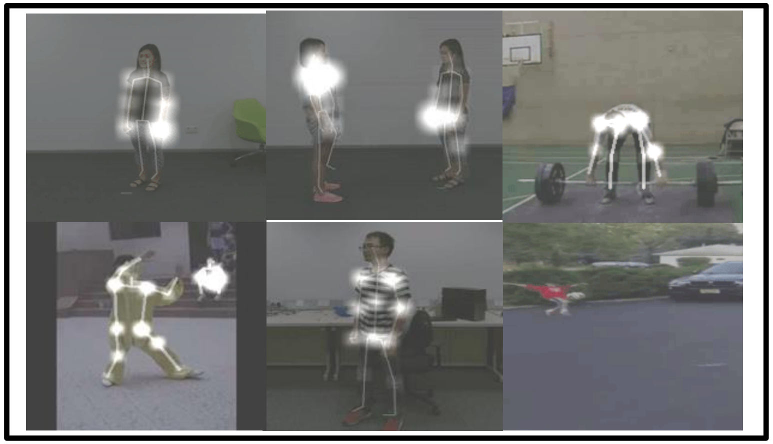


Several Attention Mechanisms:

- Squeeze-and-Excitation Attention (Channel Attention)
- Convolutional Block Attention Module (Channel + Spatial Attention)
- Self-Attention
- Spatial-Temporal Attention



Sharma et al., (ICLRW 2015)

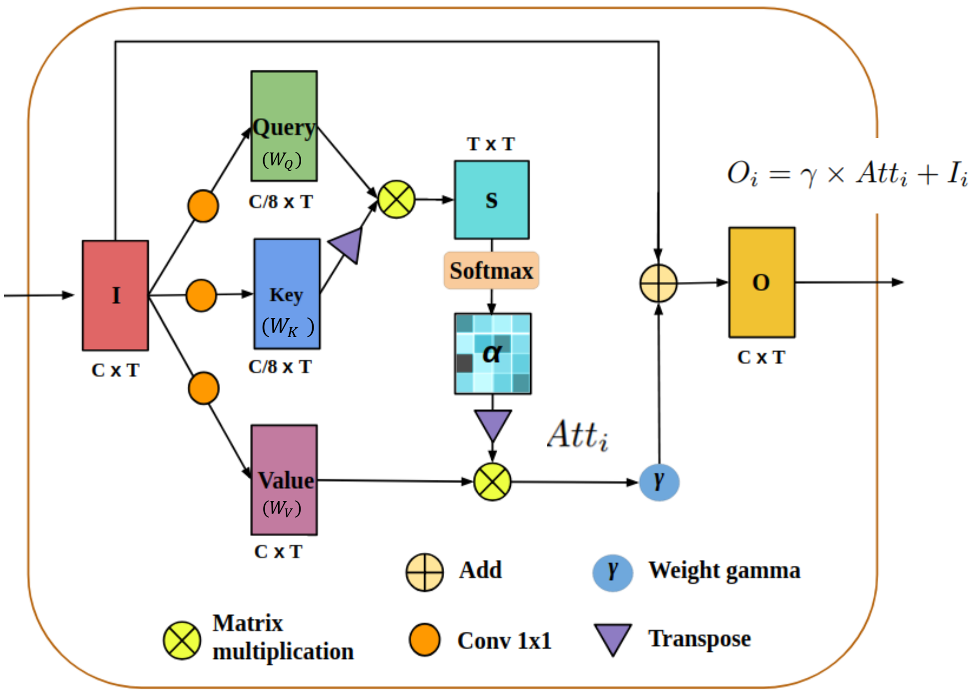


Xiaong et al., (AAAI 2018)

Activity Recognition for Daily-living activities

CNN : Approaches based on Self-Attention

Self-Attention Block

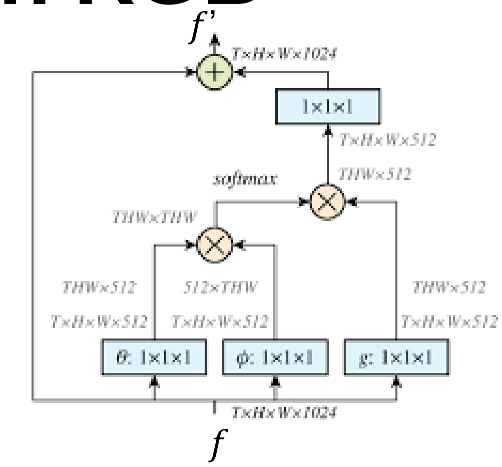


Self-attention
extracts weight-map from itself (without external information).

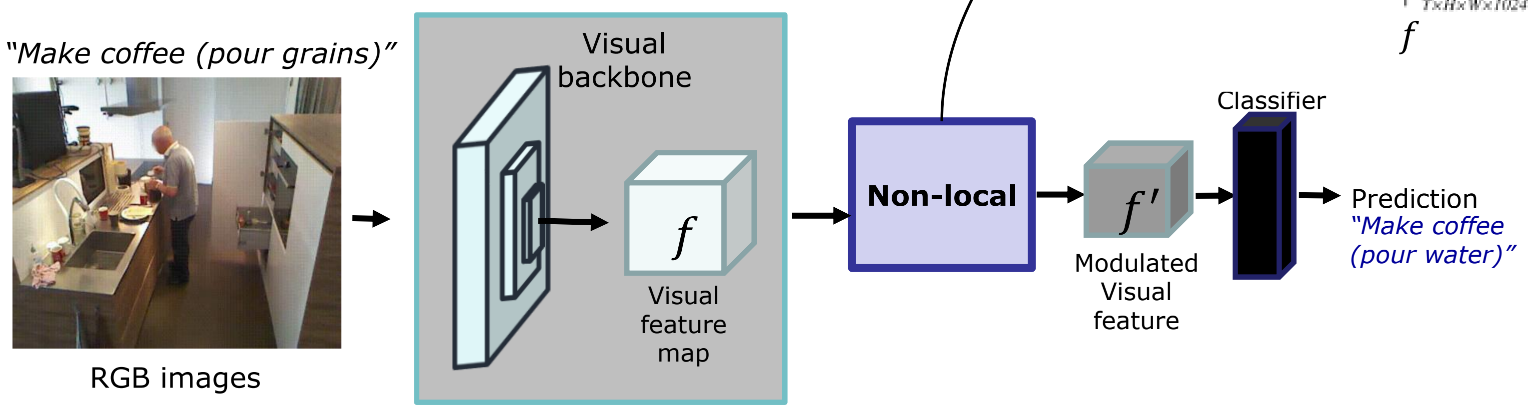
A normalizing process
Find one-to-one relations (correlations) between features at each time point

X. Wang, R. B. Girshick, A. Gupta, and K. He. Non-local neural networks. CVPR 2018

Activity Recognition for Daily-living activities
3D CNN : Approaches based on RGB



RGB based self-attention mechanism (Wang et al., CVPR 2018)



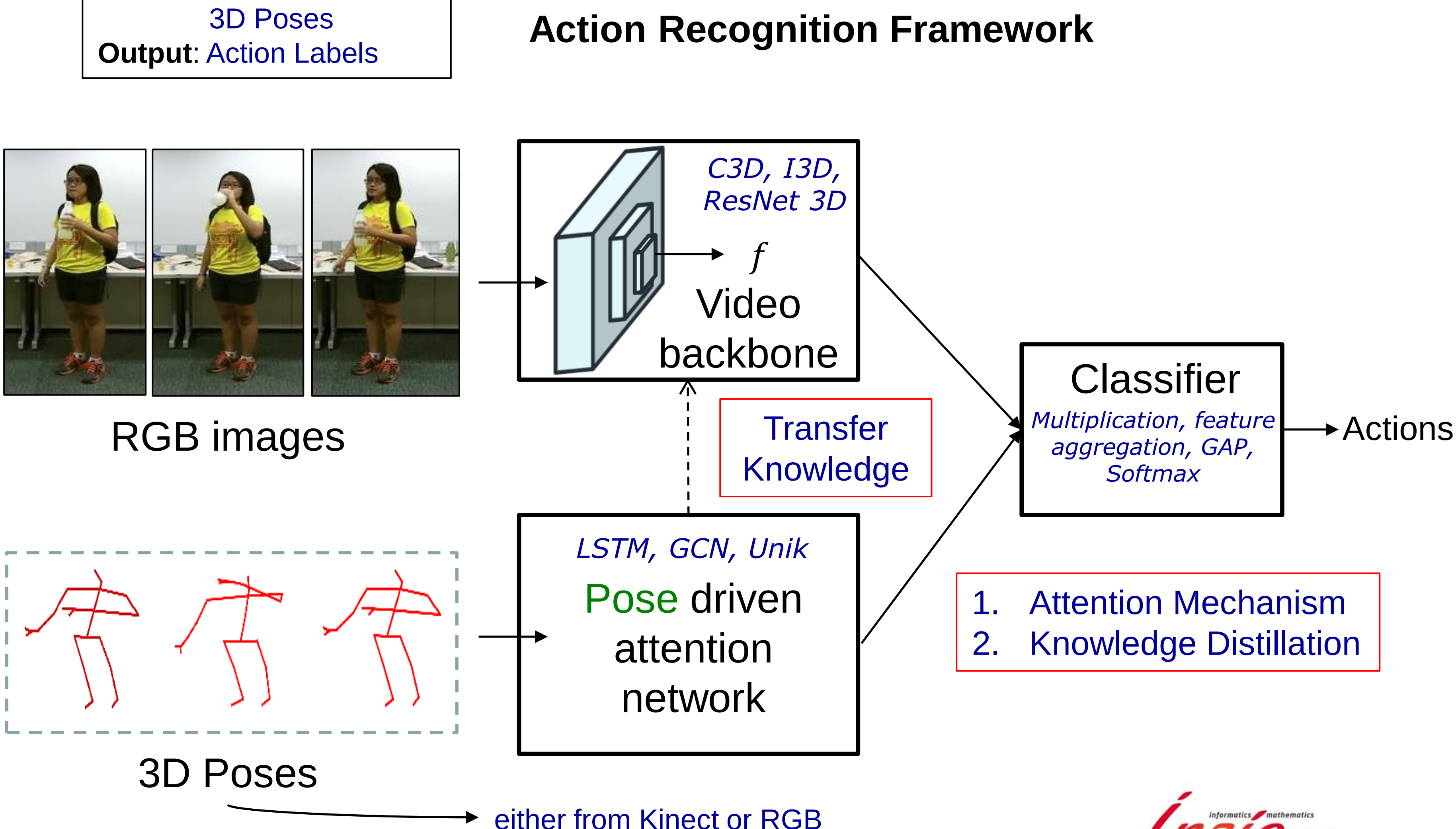
- Computes attention of each pixel as a weighted sum of the features of all the pixels in the space-time volume.
- Relies too much on the appearance of the actions, i.e., pixel position within space-time volume.

Drawbacks:

- Too rigid Spatio-temporal kernels to capture salient features for subtle spatio-temporal patterns
- No specific operations to help disambiguate similarity in actions.
- 3D (XYT) CNNs are not view-adaptive.

3D Pose guided Attention Mechanisms for Action Recognition

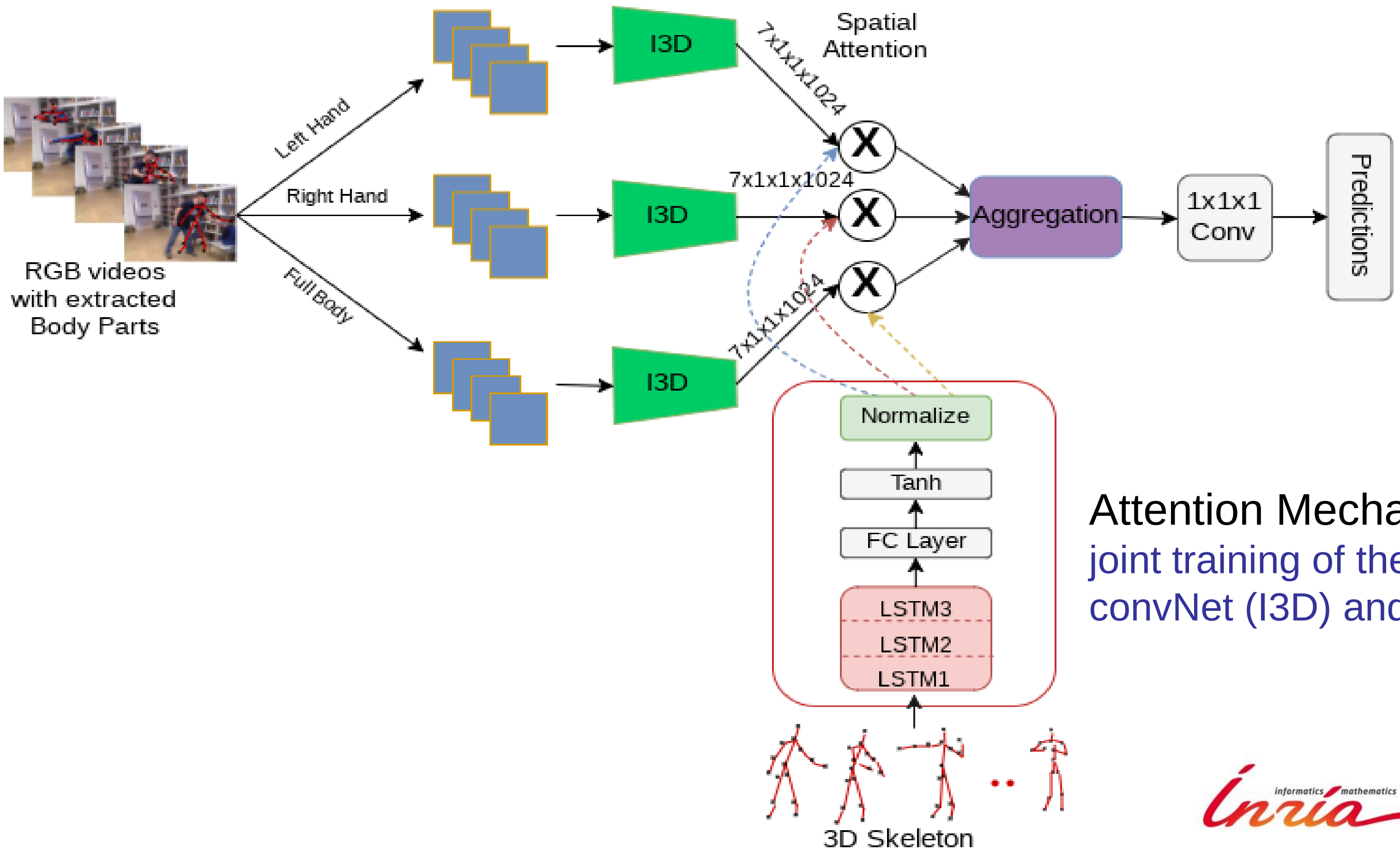
Input: RGB images
3D Poses
Output: Action Labels



Activity Recognition : Feature Selection
Parts based I3D (P-I3D) [Srijan WACV19]

To address low inter-class variation: we propose new combination of Deep Features:

- **Attention Mechanism** to learn the attention weights along 3D joints for the action features depending on the current action



Attention Mechanism:
joint training of the 3D convNet (I3D) and LSTM

Activity Recognition : Feature Selection

Parts based I3D (P-I3D) [Srijan WACV19]

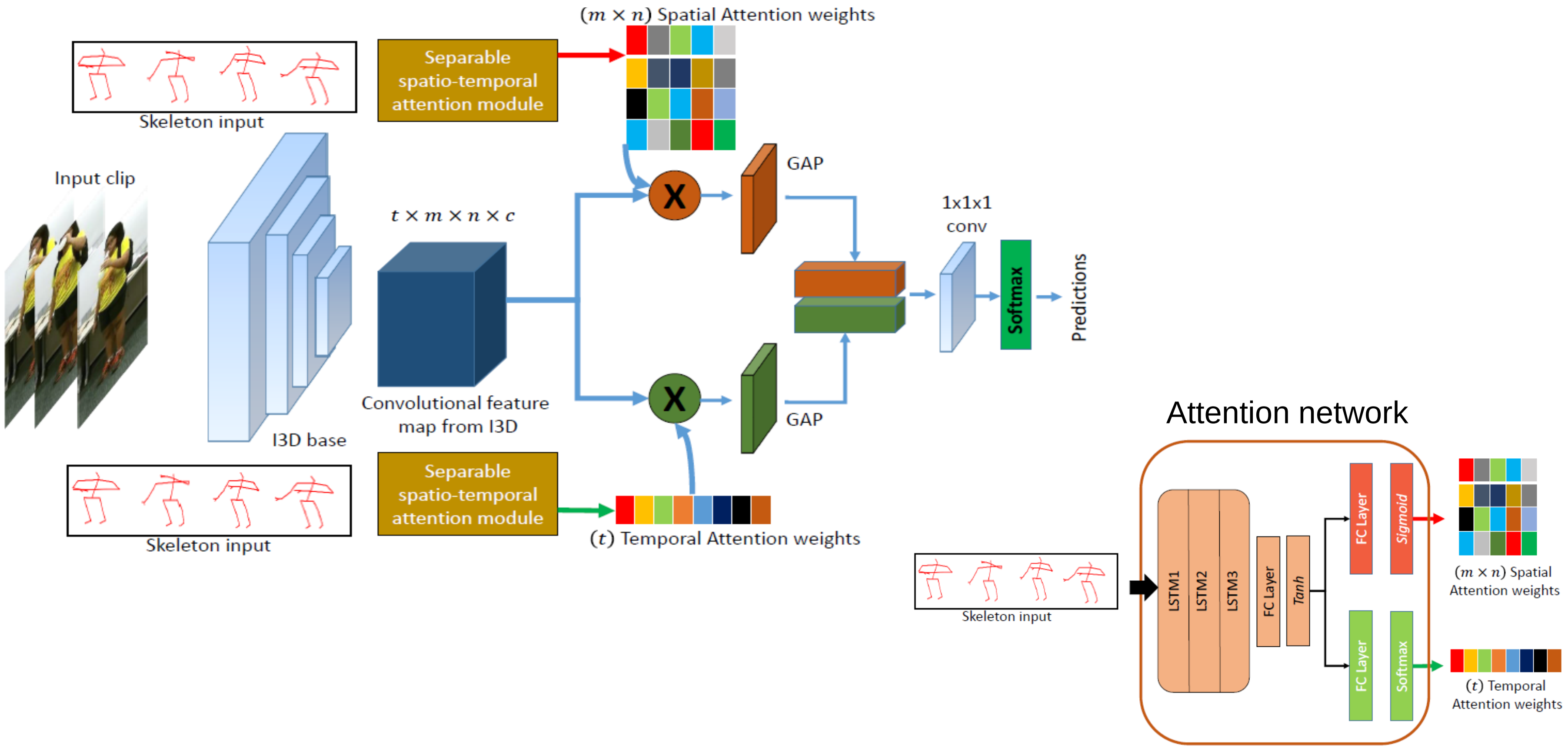
Drawbacks

- Body part Representation is not a flexible representation for human actions
- Costly in terms of # training parameters
- Temporal attention is missing.

How should we incorporate temporal attention in the current framework?

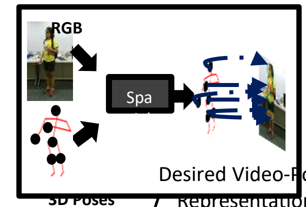
3D Pose guided Attention Mechanisms for Action Recognition [Srijan ICCV19]

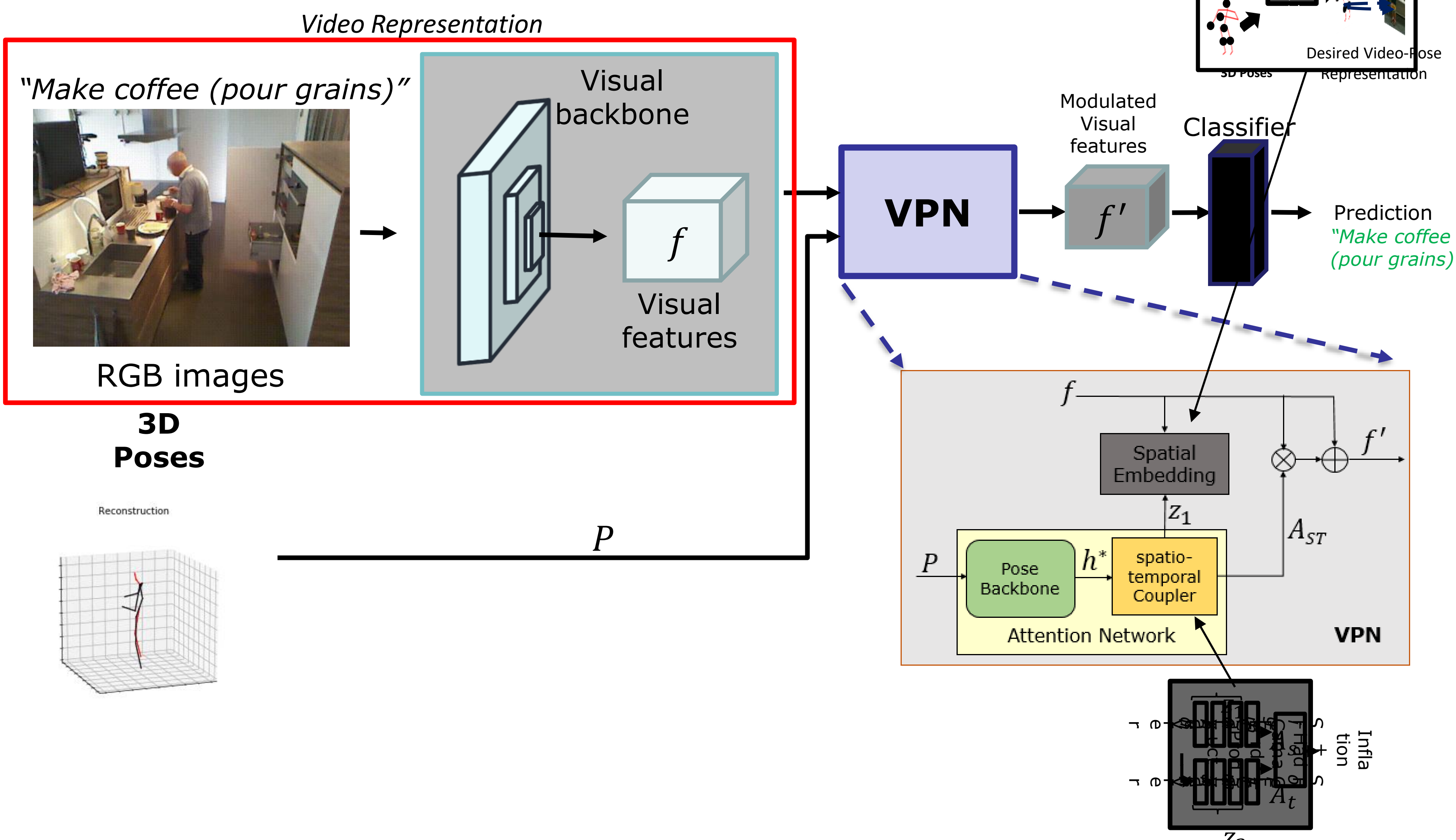
Separable Spatio-Temporal Attention (STA)



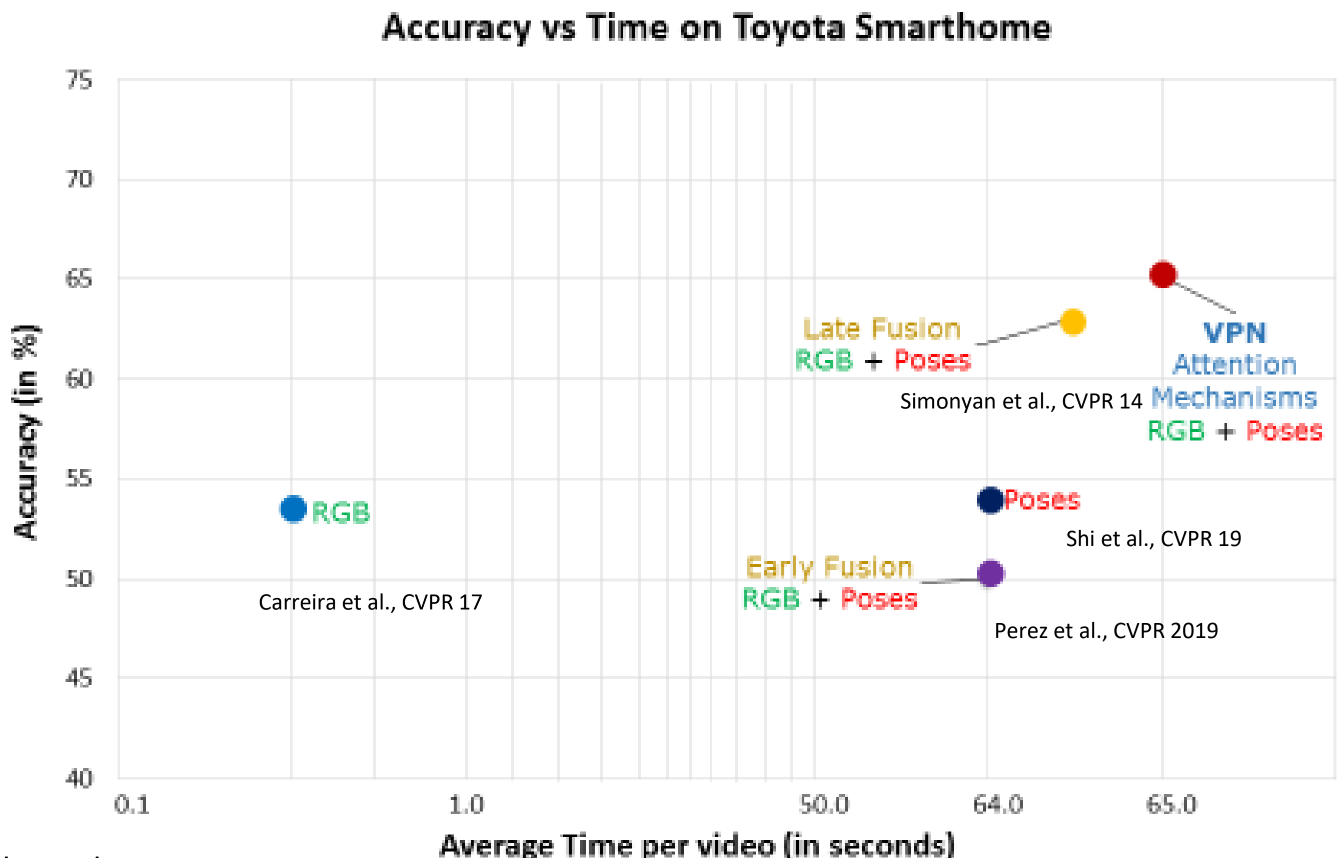
3D Pose guided Attention Mechanisms for Action Recognition (VPN) [Srijan ECCV20]

Video-Pose Network (VPN)





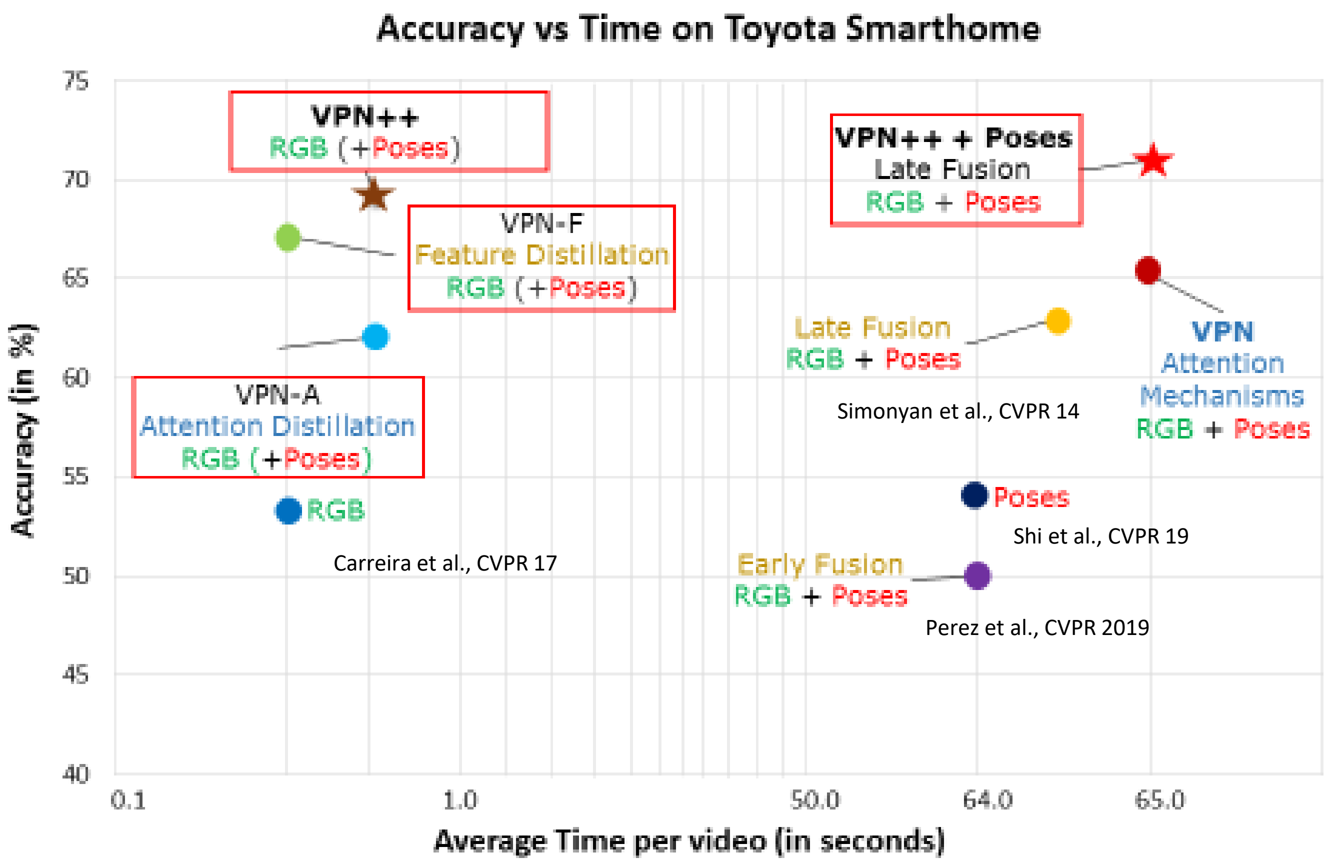
Rethinking Video-Pose embeddings VPN++



- Video-Pose models (like VPN) rely on the availability of 3D Poses.
- **Model inference time** is significantly higher than RGB based methods.

What is the best way of transferring cross-modal knowledge?

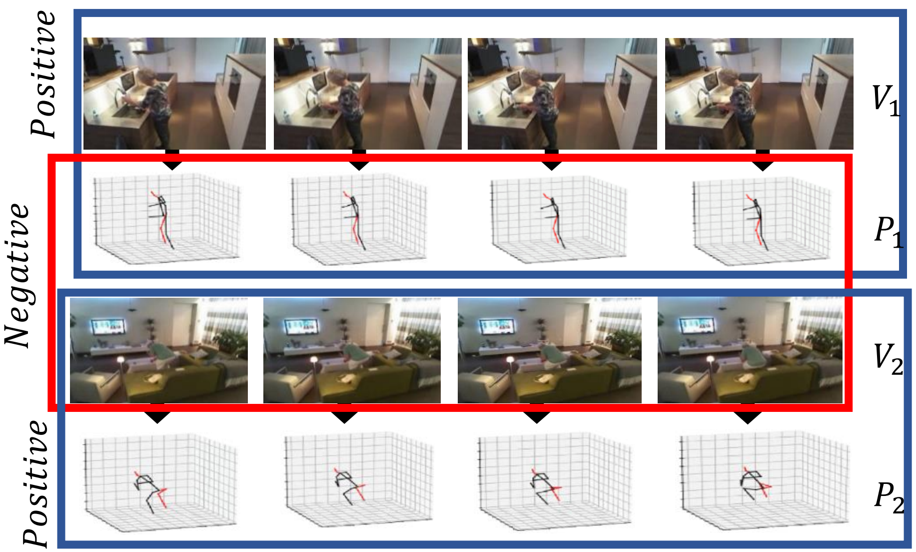
Rethinking Video-Pose embeddings

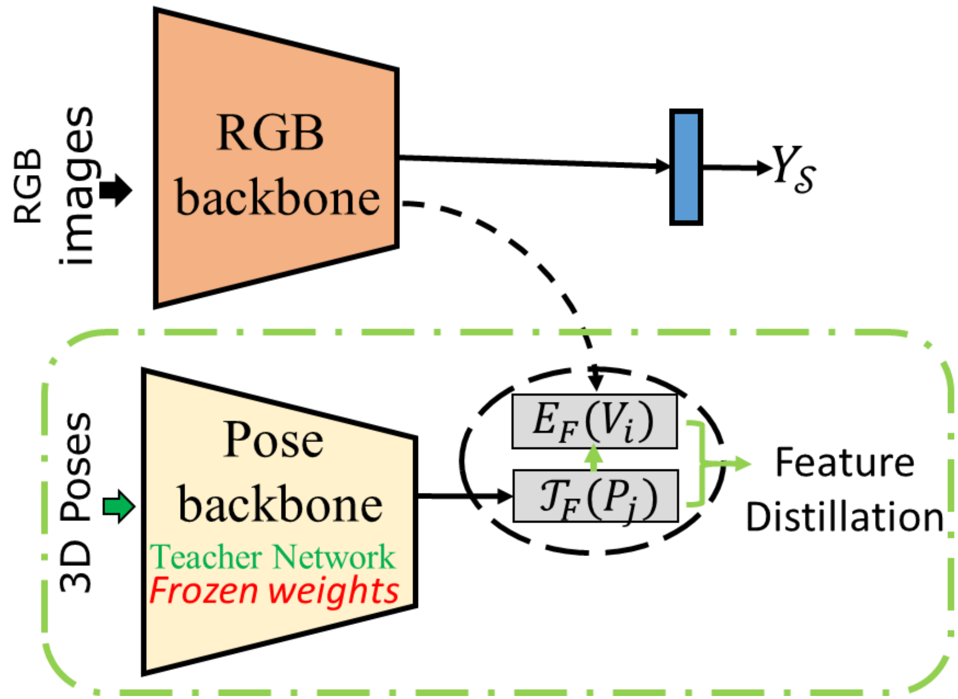


- We introduces **VPN++** that explores the concept of knowledge distillation to infuse pose stream into RGB stream.

Rethinking Video-Pose embeddings

1. **Feature-level distillation (VPN-F)**





- Positive-negative pair construction
- **Supervised Contrastive Distillation**

$$\mathcal{L}_{SCD} = \frac{1}{|\mathcal{B} - \mathcal{N}|} \sum_{i} \log[\mathcal{T}_F(P_j), E_F(V_i)] + \sum_{j \neq i} \log(1 - [\mathcal{T}_F(P_j), E_F(V_i)])$$

$$\text{where} \quad [\mathcal{T}_F(P_j), E_F(V_i)] = \frac{e^{\mathcal{T}_F(P_j)^\intercal E_F(V_i)}}{e^{\mathcal{T}_F(P_j)^\intercal E_F(V_i)} + \mathcal{M}}$$

Rethinking Video-Pose embeddings

2. **Attention-level distillation (VPN-A)**

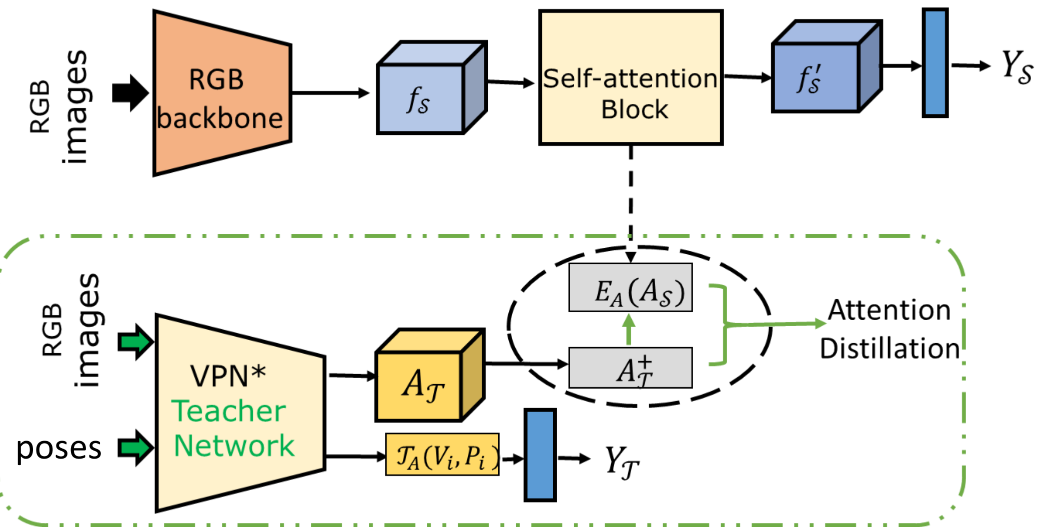
Teacher-Student choice

- Teacher – VPN (Pose driven attention network)
- Student – RGB + self-attention block

Where and how should you distillate?

- Attention or modulated features!
  (*To learn the modulated features, you need to learn the attention weights*)

- Contrastive training or Collaborative training!
  (*learning attention is an iterative process*)



$$\mathcal{L}_D = ||A^+_{\mathcal{T}} - E_A(A_{\mathcal{S}})||^2$$

Rethinking Video-Pose embeddings

Comparison to the State-of-the-art

proposed models

Action classification accuracy (in %) on **Toyota Smarthome** dataset

Methods	Pose	RGB	Att.	CS	CV_2
LSTM [Mahasseni et al., CVPR 2016]	✓	x	×	42.5	17.2
AGCN-J* [Shi et al., CVPR 2019]	✓	x	×	49.5	50
DT [Wang et al., CVPR 2011]	x	✓	×	41.9	23.7
I3D [Carreira et al., CVPR 2017]	x	✓	×	53.4	45.1
I3D + NL [Wang et al., CVPR 2018]	x	✓	✓	53.6	43.9
AssembleNet++ [Ryoo et al., ECCV 2020]	x	✓	✓	63.6	-
NPL [Piergiovanni et al., CVPR 2021]	x	✓	×	-	54.6
P-I3D	✓	✓	✓	54.0	48.7
Separable STA	✓	✓	✓	54.2	50.3
VPN	✓	✓	✓	60.8	53.5
VPN++	O	✓	✓	69.0	54.9
VPN++ + 3D Poses	✓	✓	✓	**<u>71.0</u>**	**<u>58.1</u>**

Action classification accuracy (in %) on **NUCLA** dataset

Methods	Data	Att.	$V^{3}_{1,2}$
HPM+TM [Rahmani et al., CVPR, 2016]	Depth	x	91.9
View-invariant [Liu et al., PR, 2017]	Pose	x	86.1
Ensemble TS-LSTM [Lee et al., ICCV, 2017]	Pose	x	89.2
I3D* [Carreira et al., CVPR 2017]	RGB	x	86.0
Glimpse Cloud [Baradel et al., CVPR, 2018]	RGB + $\overline{Pose}$	✓	90.1
P-I3D	RGB + Pose	✓	93.1
Separable STA	RGB + Pose	✓	92.4
VPN	RGB + Pose	✓	<u>93.5</u>
VPN++	RGB + $\overline{Pose}$	✓	91.9
VPN++ + 3D Poses	RGB + Pose	✓	**<u>93.5</u>**

Going Beyond Video-Pose

Can we generalize VPN++ over other modalities?

VFN++ - Combining RGB and Optical Flow

Stream	SH (CS)	NTU-60 (CS)	NTU-60 (CV)
RGB	53.4	85.5	87.3
OF	51.8	85.7	92.8
RGB + OF	57.3	87.1	93.6
MARS + RGB [50]	58.1	88.2	92.9
VFN++	59.0	90.1	93.4
VFN++ + OF	**66.4**	**94.6**	**97.2**

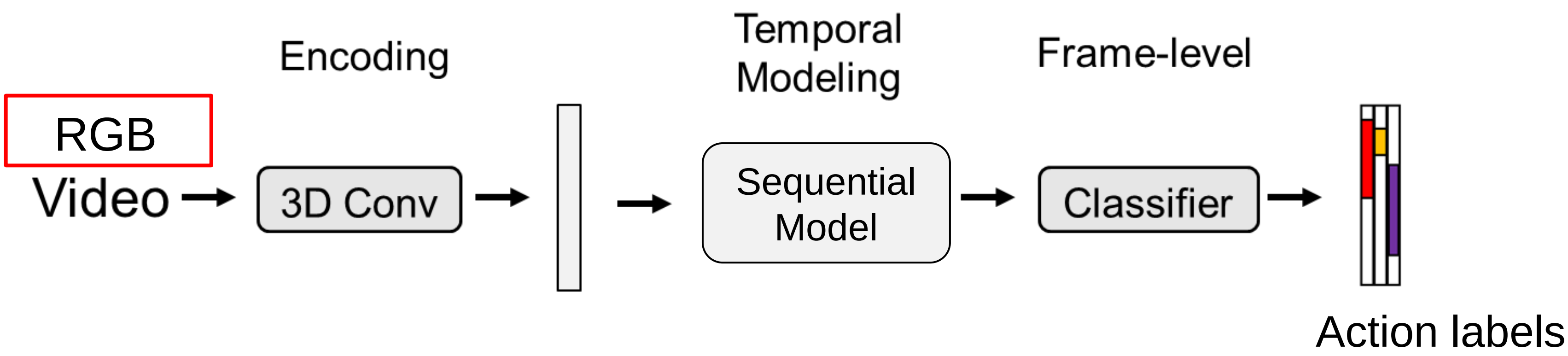
Effectiveness of Video-Flow Network++ representation using our SCD loss

VPFN++ - Combining RGB, Pose and Optical Flow

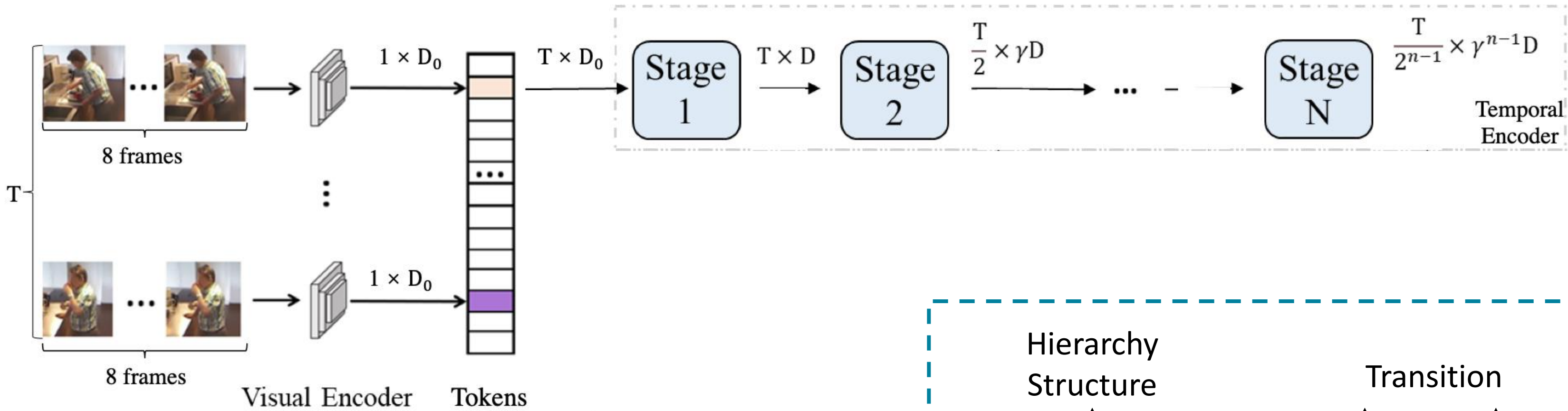
Fusion	SH (CS)	NTU-60 (CS)	NTU-60 (CV)
RGB + OF + Pose	64.4	90.2	95.9
VPN++	69.0	91.9	94.9
VPFN	69.7	92.1	95.5
VPFN + Pose	71.7	95.1	98.2
VPFN + Pose + OF	**72.9**	**96.7**	**99.1**

TABLE 12: Combination of RGB, Pose and Flow modalities into a single model. Here VPFN is VPN++ + FARS.

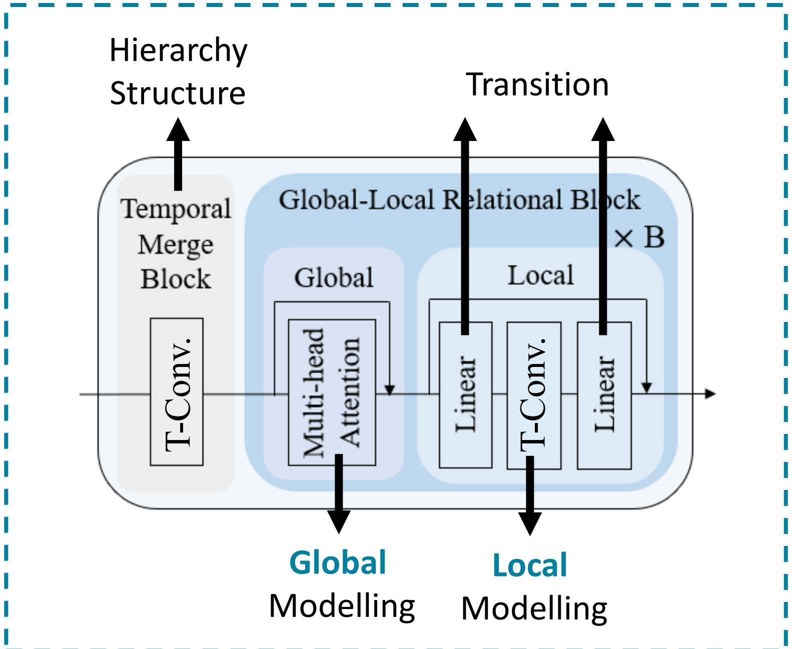
Action Detection Framework



Temporal Modelling/Encoder: MS-TCT



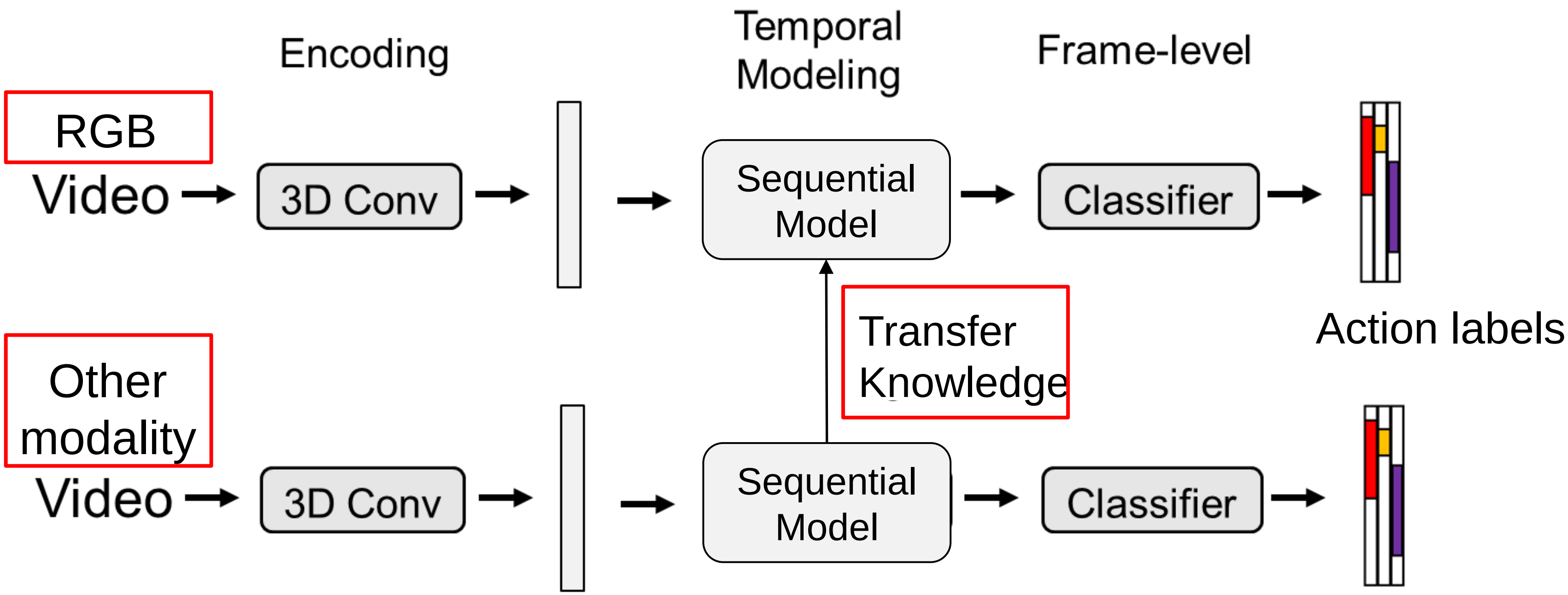
- Temporal merge block introduces temporal hierarchy (T-Conv., Stride 2)
- Multi-head Attention and Convolution are used for temporal modelling
- Linear layer is for transition between two layers
- Modelling Global and Local temporal relations at multiple scales.



Temporal Encoder Block

R. Dai, S. Das, K. Kahatapitiya, M. Ryoo and F. Bremond. MS-TCT: Multi-Scale Temporal ConvTransformer for Action Detection. CVPR 2022.

Action Detection Framework



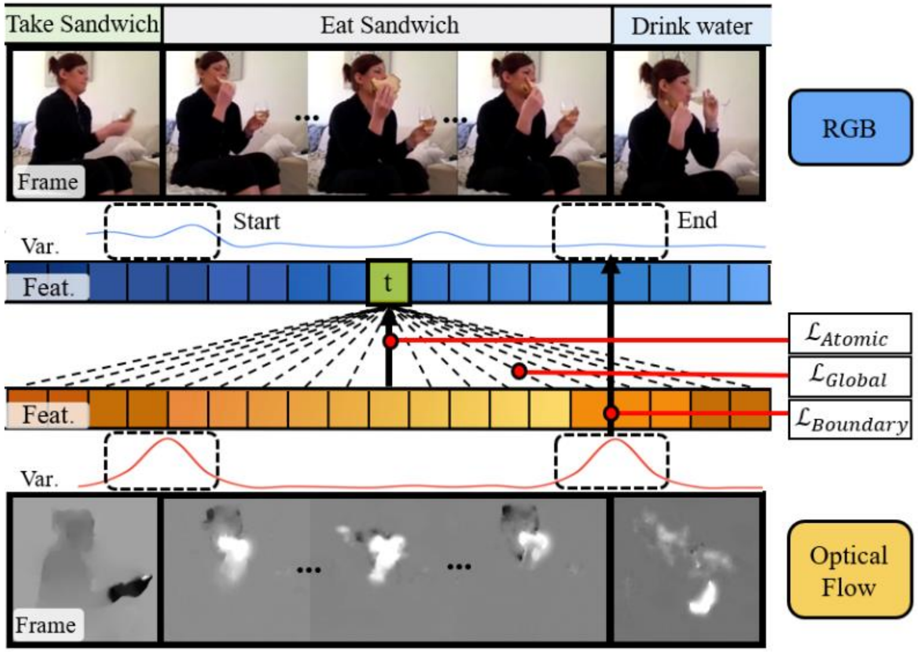
How should we leverage the modalities in action detection?

1. Early or Late fusion
2. Attention-based Guidance
3. Knowledge Distillation (other modalities only at training time)

Action Detection Framework

Extension of VPN++ for long untrimmed videos

Augmented RGB – RGB stream hallucinating Poses / OF.



	Charades	PKU-MMD	TSU-CS	TSU-CV
Teacher-OF	18.6	68.4	29.4	17.5
Teacher-Pose	9.8	65.0	26.2	22.4
Vanilla-RGB	22.3	79.6	29.2	18.9
Two-stream RGB + Pose	23.0	82.9	32.6	23.7
Two-stream RGB + OF	24.8	83.4	33.5	19.5
Pose Augmented RGB	23.2	84.7	32.4	23.6
OF Augmented RGB	24.6	85.5	32.8	19.3
Pose + OF Augmented RGB	24.9	86.3	33.7	23.8

Rui Dai, Srijan Das, Francois Bremond, "Learning an Augmented RGB Representation with Cross-Modal Knowledge Distillation for Action Detection", ICCV 2021.

3D Pose guided Attention Mechanisms for Action Recognition

Conclusion

VPN : effective strategy to exploit 3D poses to guide RGB cues for recognizing Activities of Daily Living (ADL) in real-world scenarios.

- 3 variants of spatio-temporal attention mechanisms for the recognition of ADL.
- The codes and models are available at https://github.com/srijandas07. The inference time of these models are close to real-time given the poses.

VPN++ : effective strategy to exploit 3D poses when poses are not available at inference time.

- These models have been evaluated on four public datasets achieving state-of-the-art results.

For datasets with Laboratory settings, the accuracy is > 92%
For real-world datasets, the accuracy is up to 69%

what about the rest 31%?

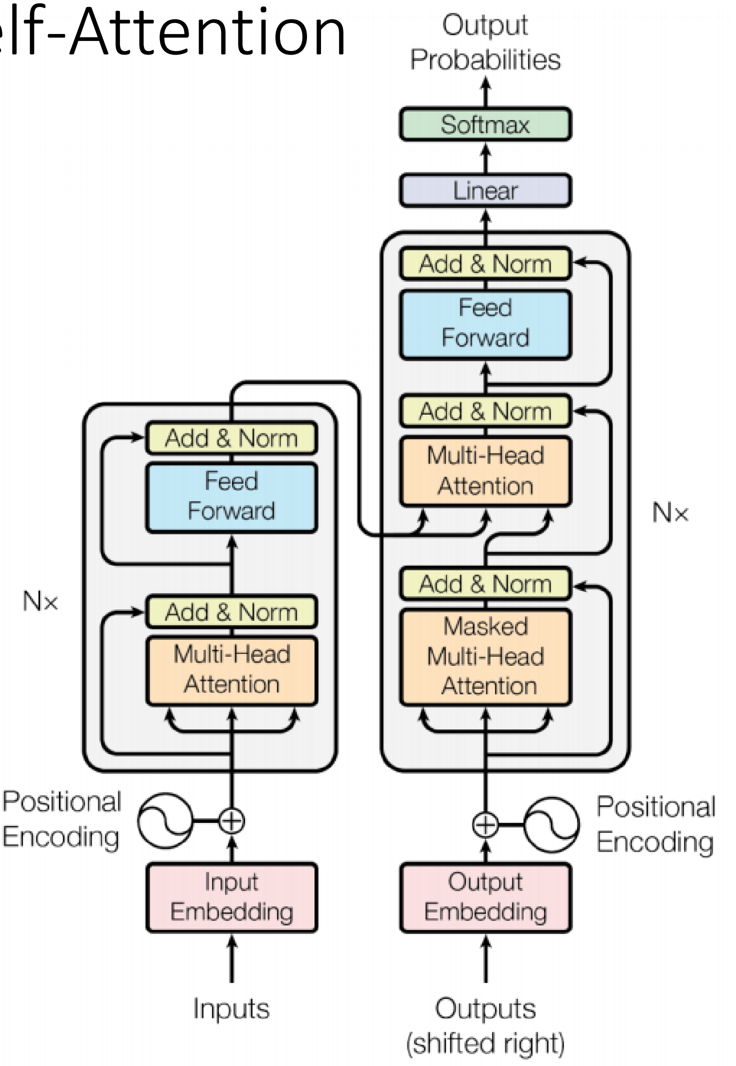
Section 4

Transformer

Vision Transformer

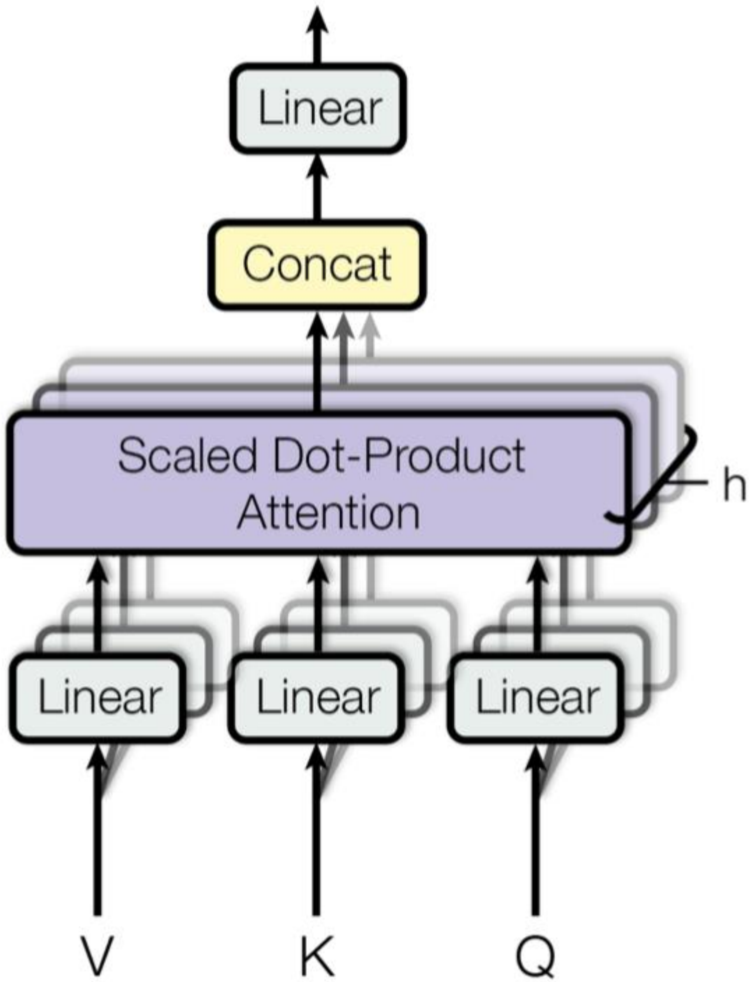
Transformers are based on Self-Attention

- Positional Encoding
- Multi-head attention
- Feed Forward
- Outputs



Vision Transformer

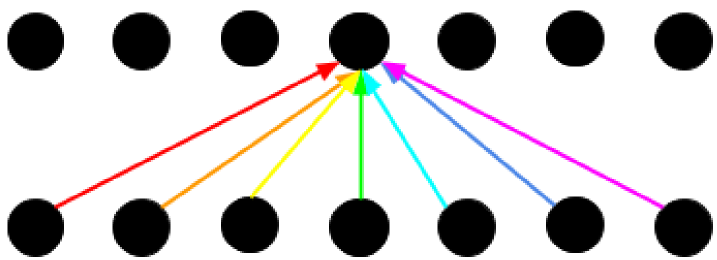
Multi-head Attention



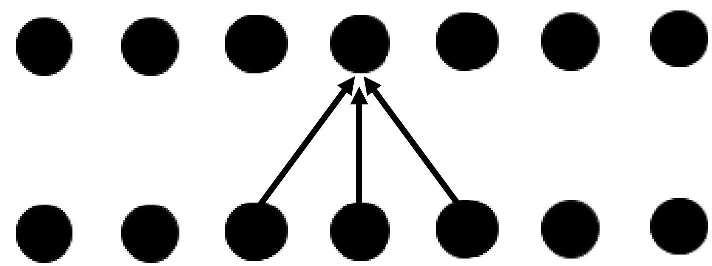
Vision Transformer

Transformer and Convolution

Transformer layer



Convolution layer



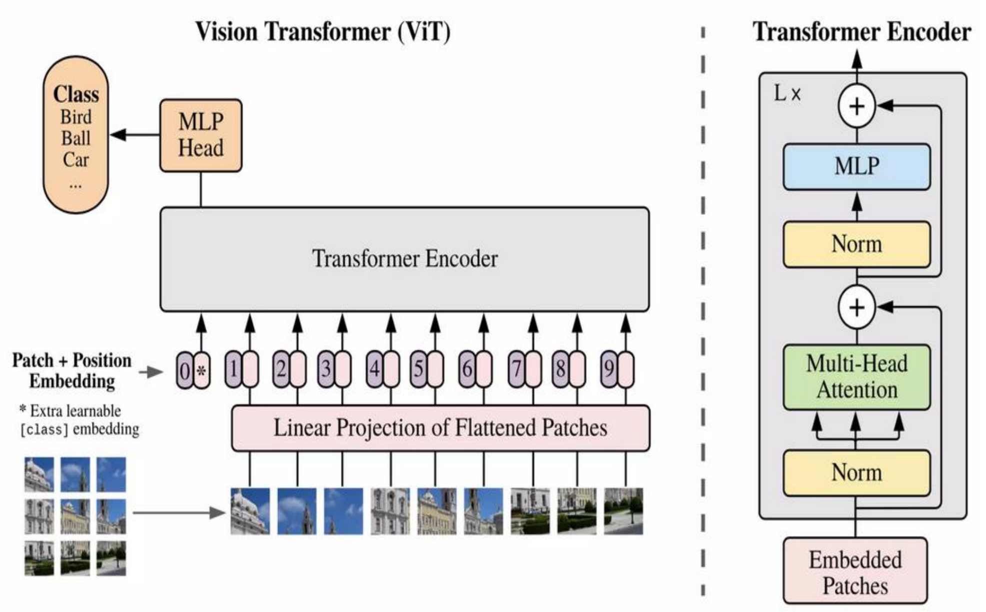
Pros:
Global relation
Attention enhanced

Cons:
More Flops
Loss location info

Vision Transformer (ViT):

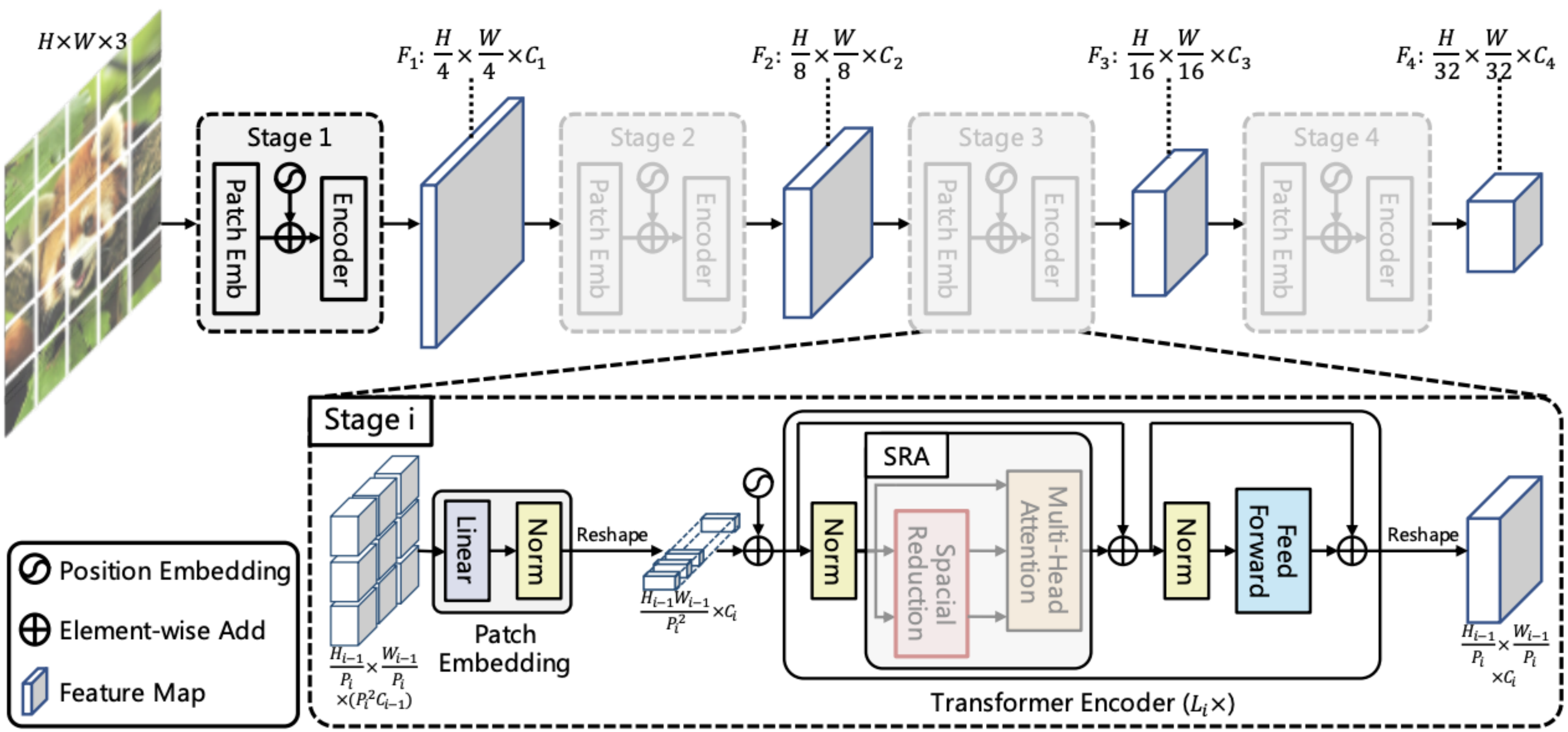
- In ViTs, images are represented as sequences, and class labels for the image are predicted, which allows models to learn image structure independently.

- **How ViT works?**
  - Split an image into patches (Tokenize)
  - Flatten the patches
  - Produce lower-dimensional linear embeddings from the flattened patches
  - Add positional embeddings
  - Feed the sequence as an input to a standard transformer encoder (for interaction among tokens)
  - Pretrain the model with image labels (fully supervised on a huge dataset)
  - Finetune on the downstream dataset for image classification



Vision Transformer

Pyramid Vision Transformer



Vision Transformer

ViT vs. CNN:

- ViT has more similarity between the representations obtained in **shallow and deep layers** compared to CNNs.
- Unlike CNNs, ViT obtains the **global representation** from the shallow layers, but the local representation obtained from the shallow layers is also important.
- **Skip connections** in ViT are even more influential than in CNNs (ResNet) and substantially impact the performance and similarity of representations.
- ViT retains more spatial information than CNN.
- ViT can learn high-quality intermediate representations with large amounts of data.
- ViT is more **Scalable and Efficient** compared to CNN.

Attention Mechanisms for Action Recognition

Future Work

- ➢ Towards **understanding human object relationships (better visual features)**
- Better visual base network (e.g. ViViT? VideoMAE? DinoV2? VifiCLIP?)
- handling the spatial resolution (super-resolution) which is crucial for real-world ADL recognition (object)
- End-to-End training for the visual encoder (e.g. memory bank, adapters)

- > spatio-temporal context by encoding human object relationships

- ➢ Towards **multi-modal video representation**
- Going beyond RGB + one modality: + depth, optical flow, 3D poses, audio, text, physiological, etc.

- ➢ Towards **unsupervised video representation**

Pre-trained the model to do complicated tasks by self-supervision/ contrastive learning/ masking

- learning the synchronization between RGB & Poses (i.e. common subspace)
- Video MAE Mask Auto-Encoder

- ➢ **Real-World Activity Detection** is the next step!
  - – Extending Temporal Model for activity detection

    utilizing 3D poses for precise prediction of starting and ending of an activity,

    new models (i.e. mamba).

Questions ?