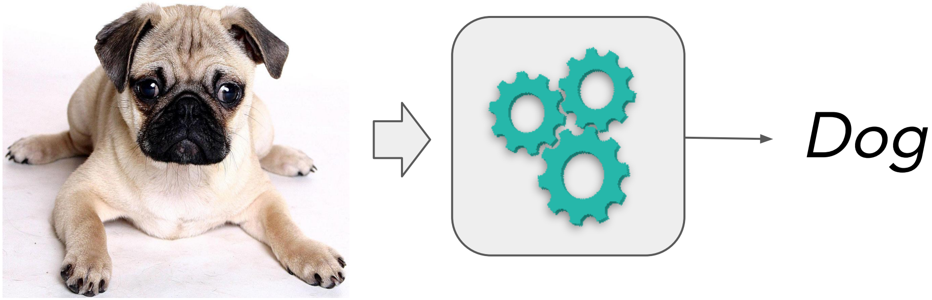
POLITECNICO DI TORINO

MLDL2020

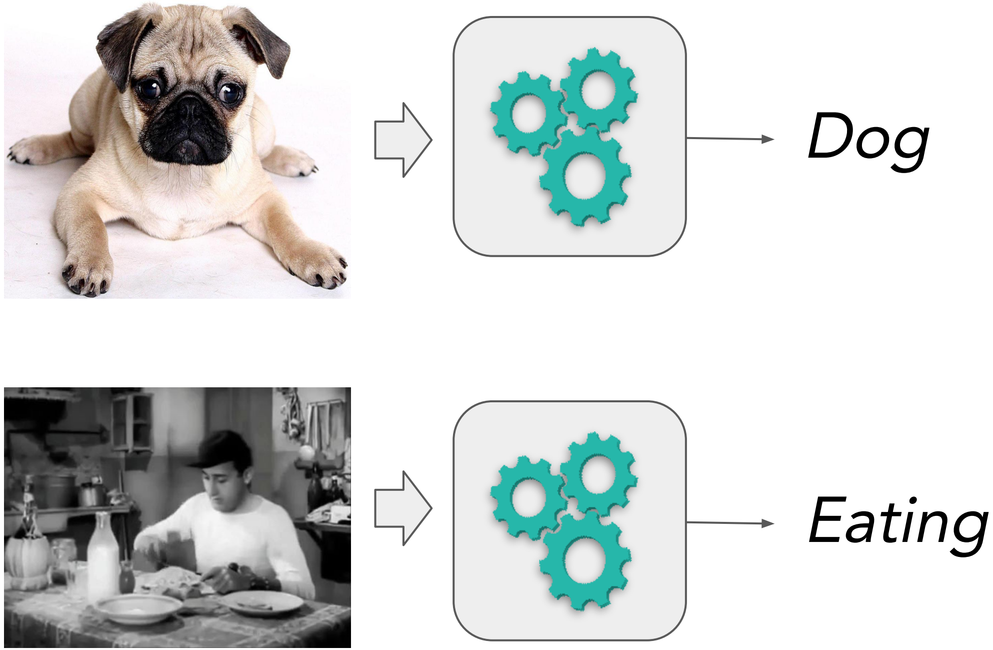
# Action Recognition

Mirco Planamente
mirco.planamente@polito.it

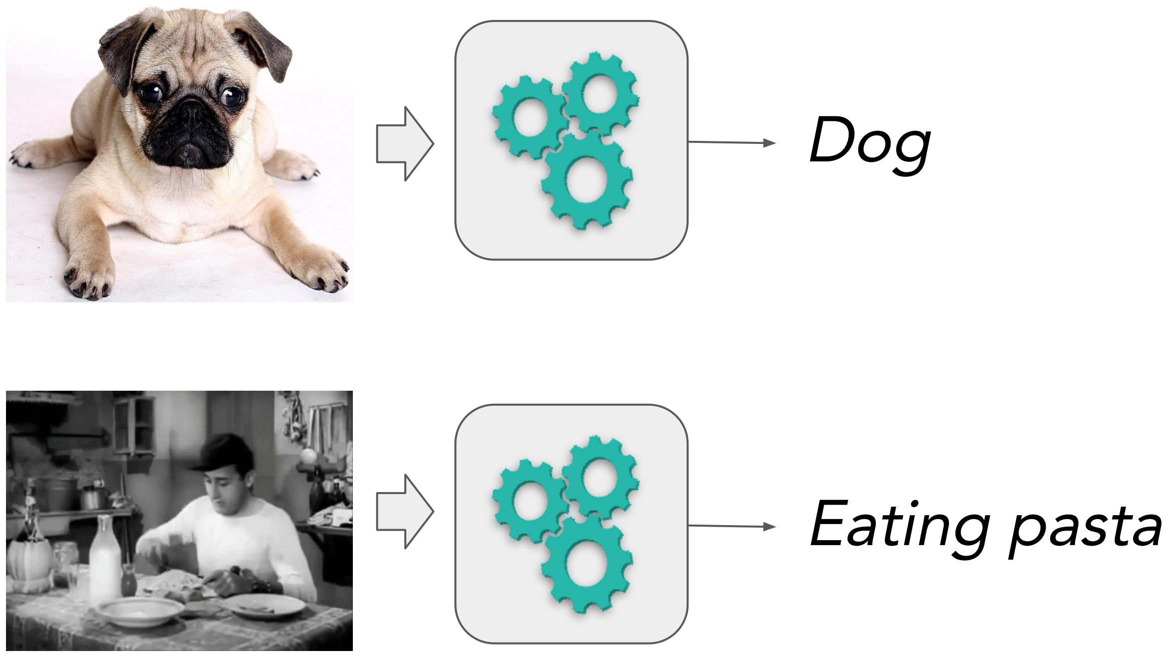
# Images Vs Videos



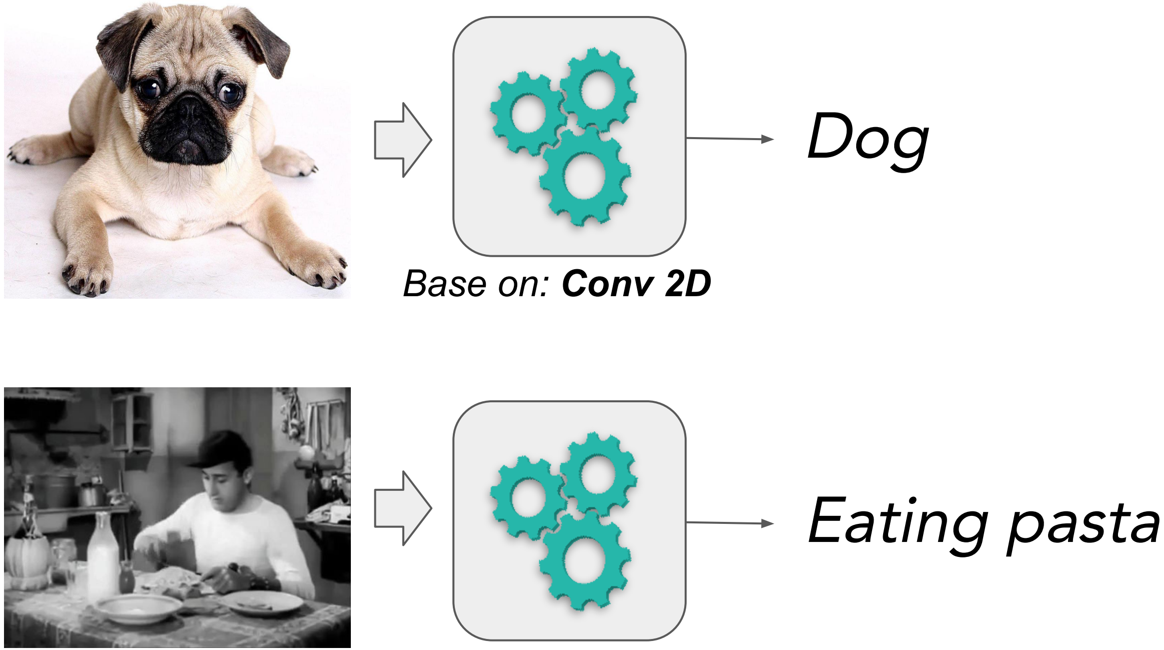
# Images Vs Videos



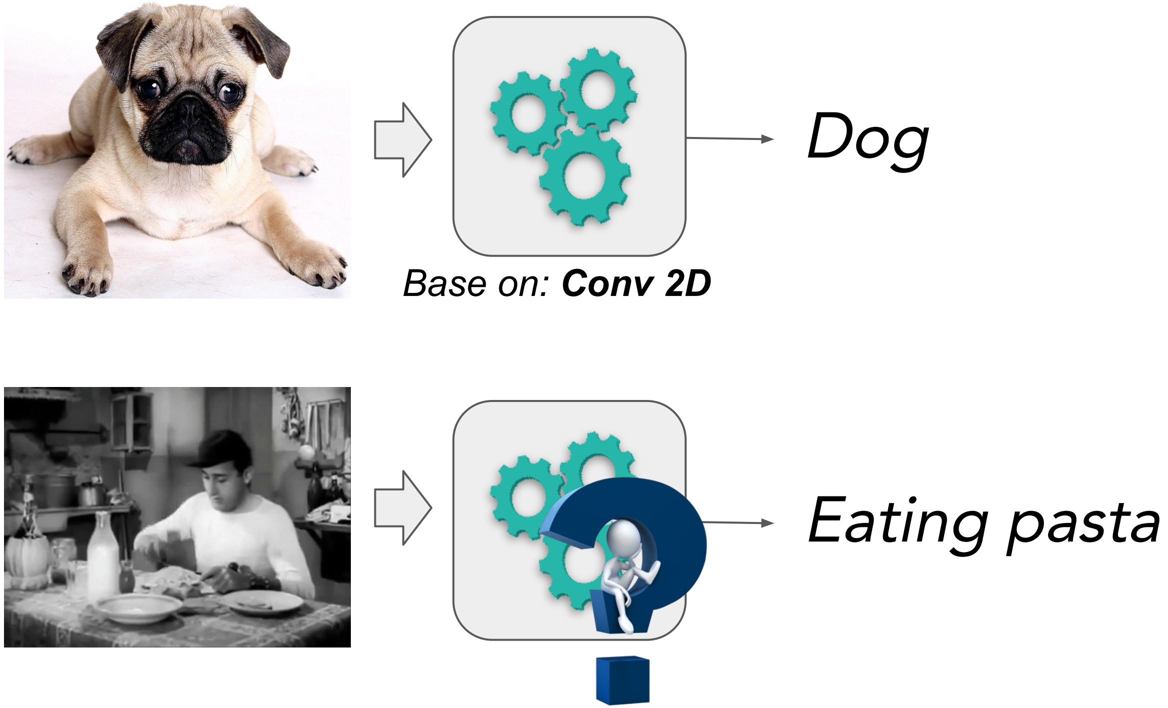
# Images Vs Videos



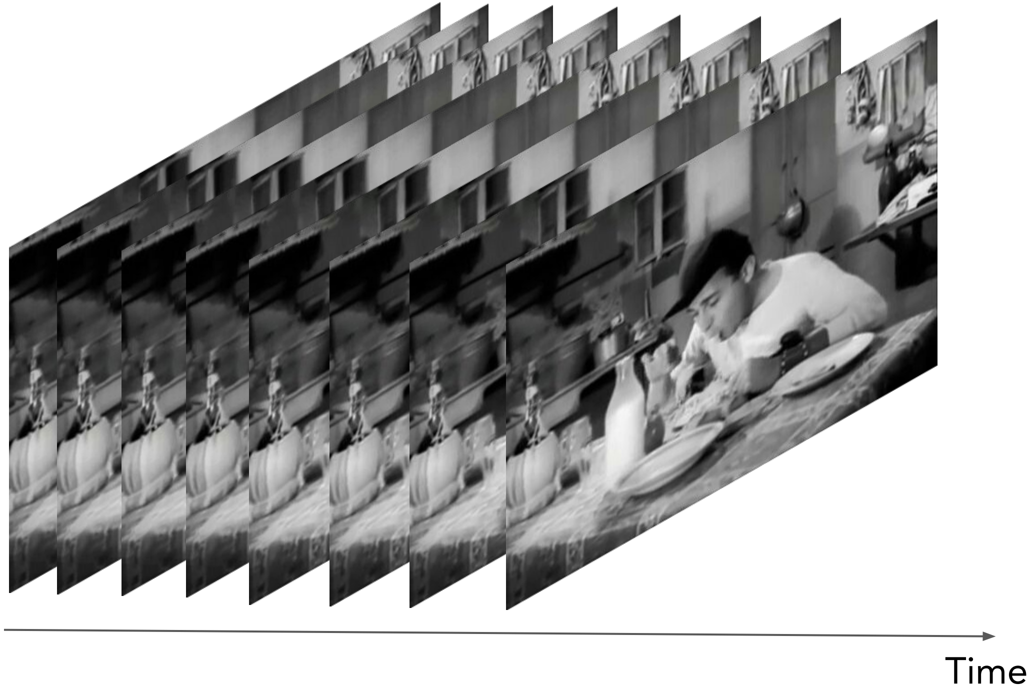
# Images Vs Videos



# Images Vs Videos



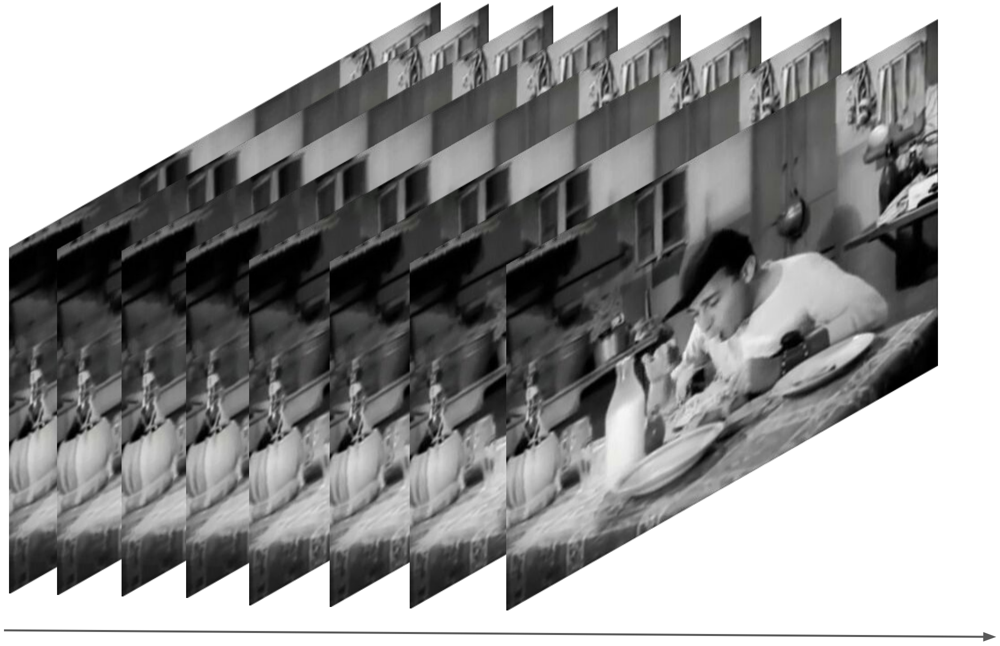
# First Solution




S. Ji, W. Xu, M. Yang, and K. Yu. *“3d convolutional neural networks for human action recognition.”*(ICML10)
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh *“Learning Spatio-Temporal Features With 3D Residual Networks for Action Recognition”*(ICCV17)

# First Solution

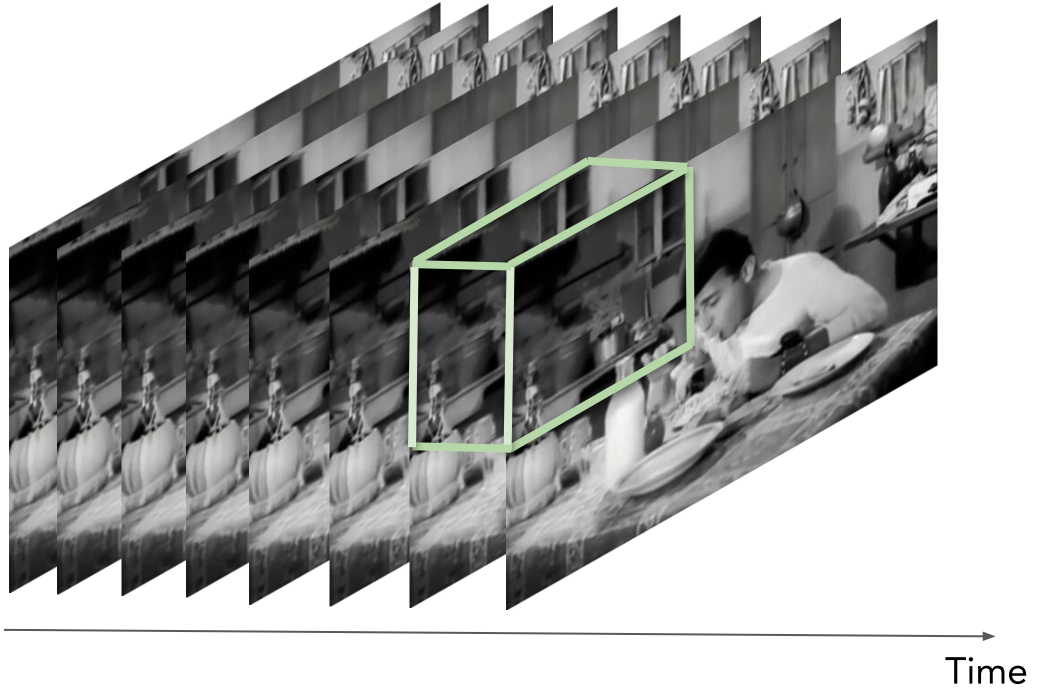
3D Convolution:



Time

S. Ji, W. Xu, M. Yang, and K. Yu. *"3d convolutional neural networks for human action recognition."*(ICML10)
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh *"Learning Spatio-Temporal Features With 3D Residual Networks for Action Recognition"*(ICCV17)

# First Solution

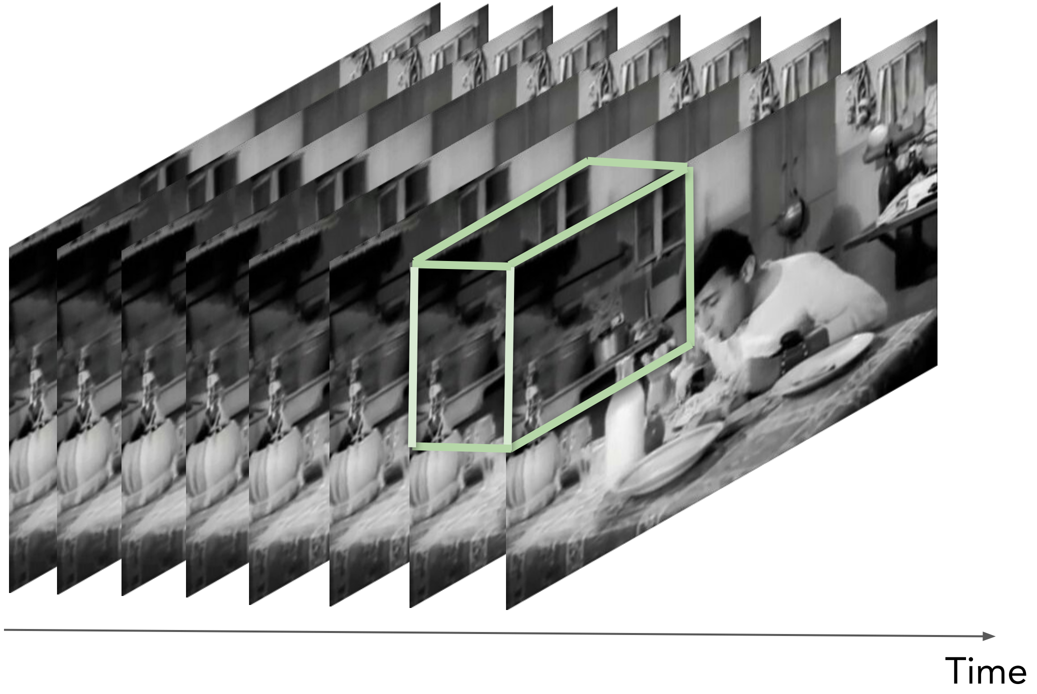



3D Convolution:

- The filter is 3-dimensional

S. Ji, W. Xu, M. Yang, and K. Yu. *“3d convolutional neural networks for human action recognition.”*(ICML10)
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh *“Learning Spatio-Temporal Features With 3D Residual Networks for Action Recognition”*(ICCV17)

# First Solution

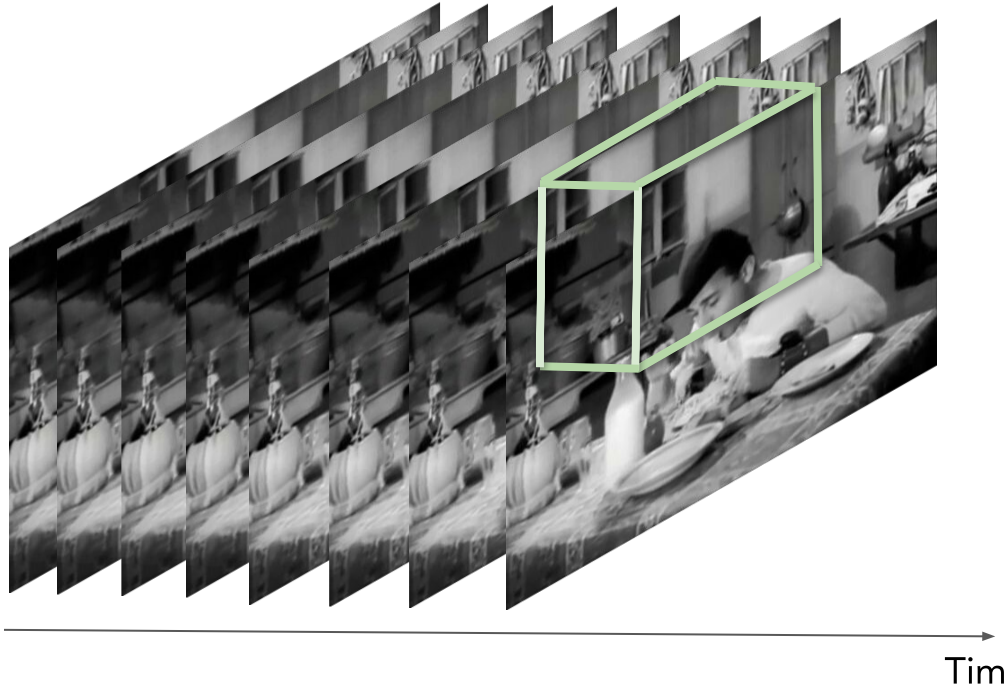


3D Convolution:

- The filter is 3-dimensional
- It slides in 3 directions


S. Ji, W. Xu, M. Yang, and K. Yu. *"3d convolutional neural networks for human action recognition."*(ICML10)
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh *"Learning Spatio-Temporal Features With 3D Residual Networks for Action Recognition"*(ICCV17)

# First Solution

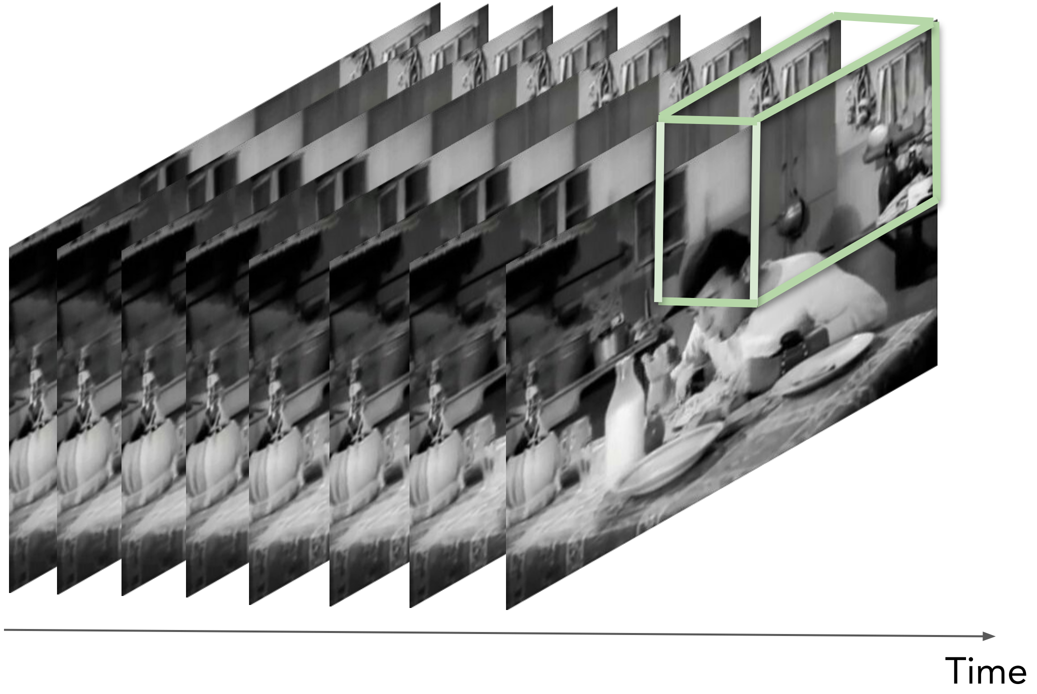



3D Convolution:

- The filter is 3-dimensional
- It slides in 3 directions

S. Ji, W. Xu, M. Yang, and K. Yu. *“3d convolutional neural networks for human action recognition.”*(ICML10)
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh *“Learning Spatio-Temporal Features With 3D Residual Networks for Action Recognition”*(ICCV17)

# First Solution




3D Convolution:

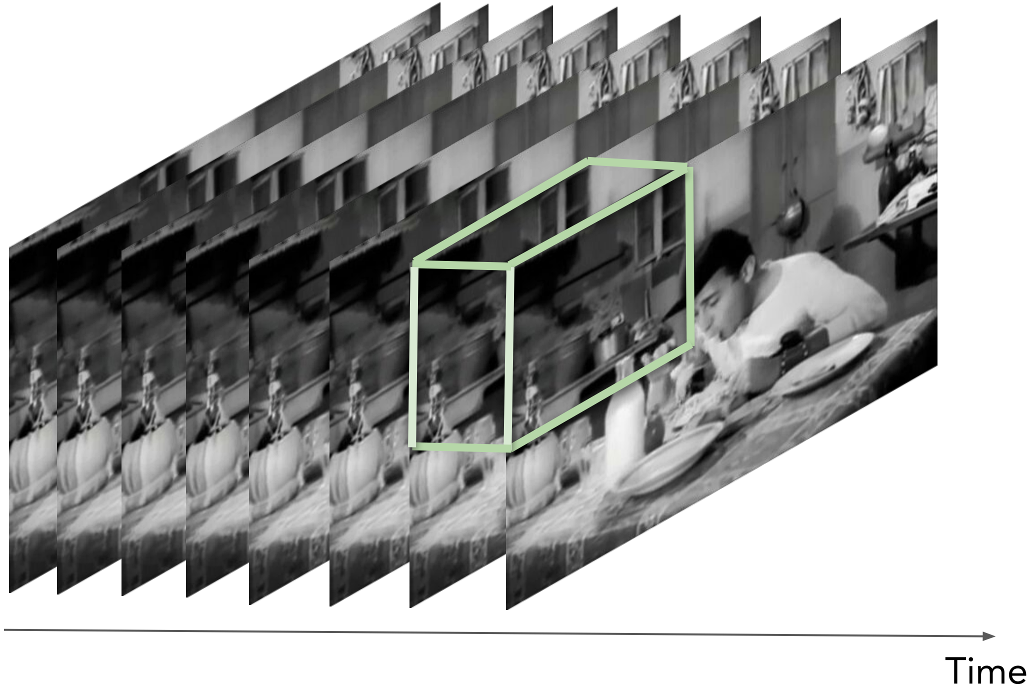
- The filter is 3-dimensional
- It slides in 3 directions

S. Ji, W. Xu, M. Yang, and K. Yu. *"3d convolutional neural networks for human action recognition."*(ICML10)
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh *"Learning Spatio-Temporal Features With 3D Residual Networks for Action Recognition"*(ICCV17)

# First Solution

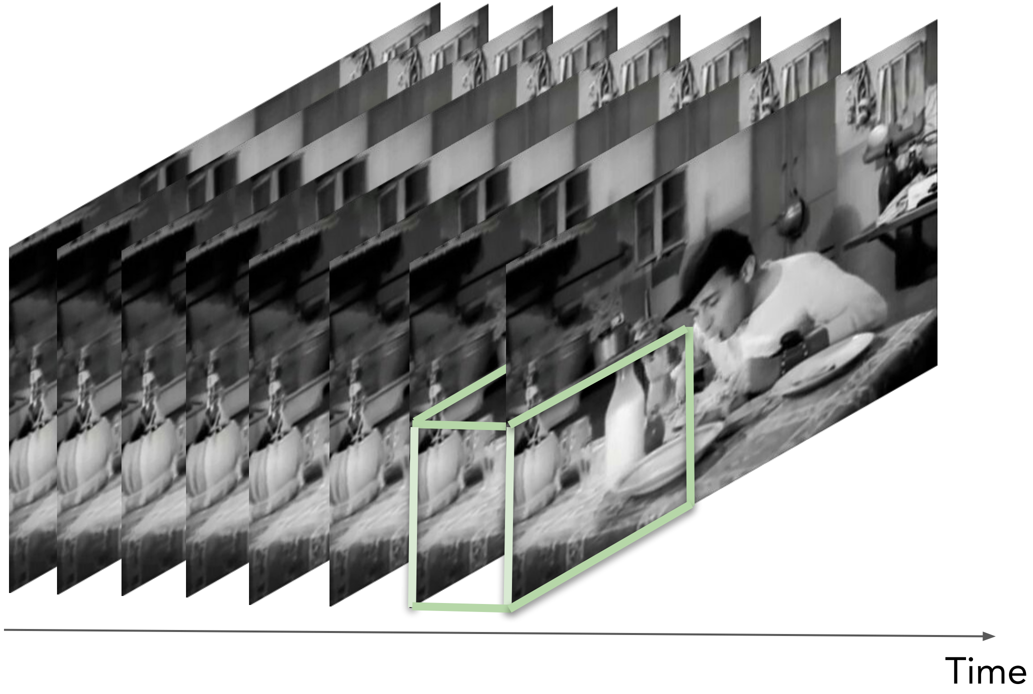
3D Convolution:

- The filter is 3-dimensional
- It slides in 3 directions




S. Ji, W. Xu, M. Yang, and K. Yu. *"3d convolutional neural networks for human action recognition."*(ICML10)
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh *"Learning Spatio-Temporal Features With 3D Residual Networks for Action Recognition"*(ICCV17)

# First Solution

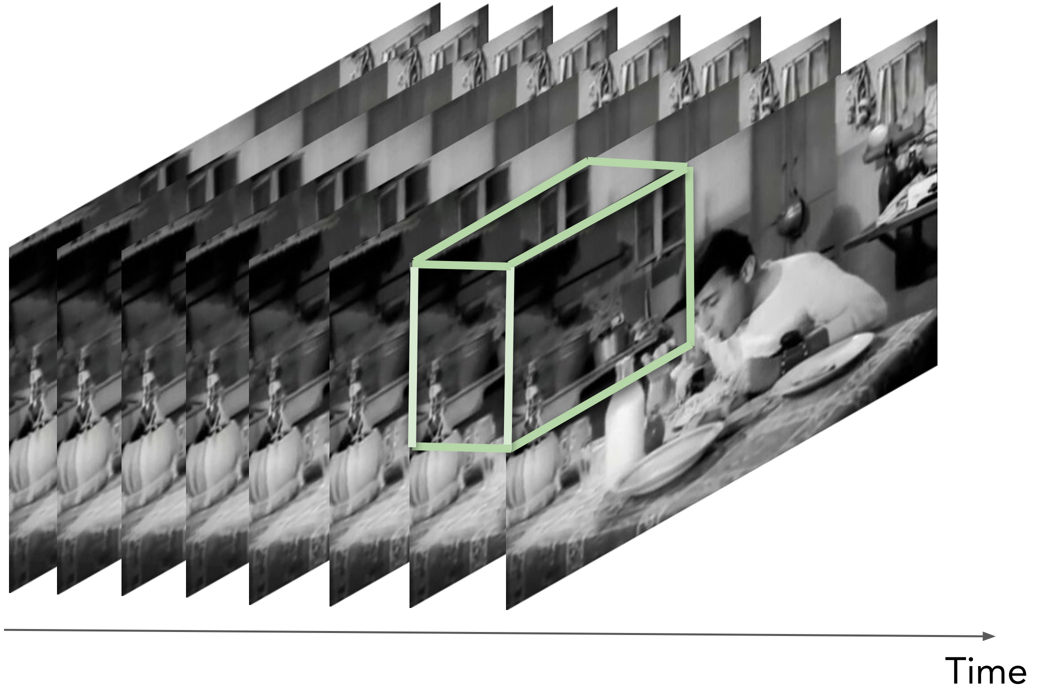



3D Convolution:

- The filter is 3-dimensional
- It slides in 3 directions


S. Ji, W. Xu, M. Yang, and K. Yu. *“3d convolutional neural networks for human action recognition.”*(ICML10)
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh *“Learning Spatio-Temporal Features With 3D Residual Networks for Action Recognition”*(ICCV17)

# First Solution

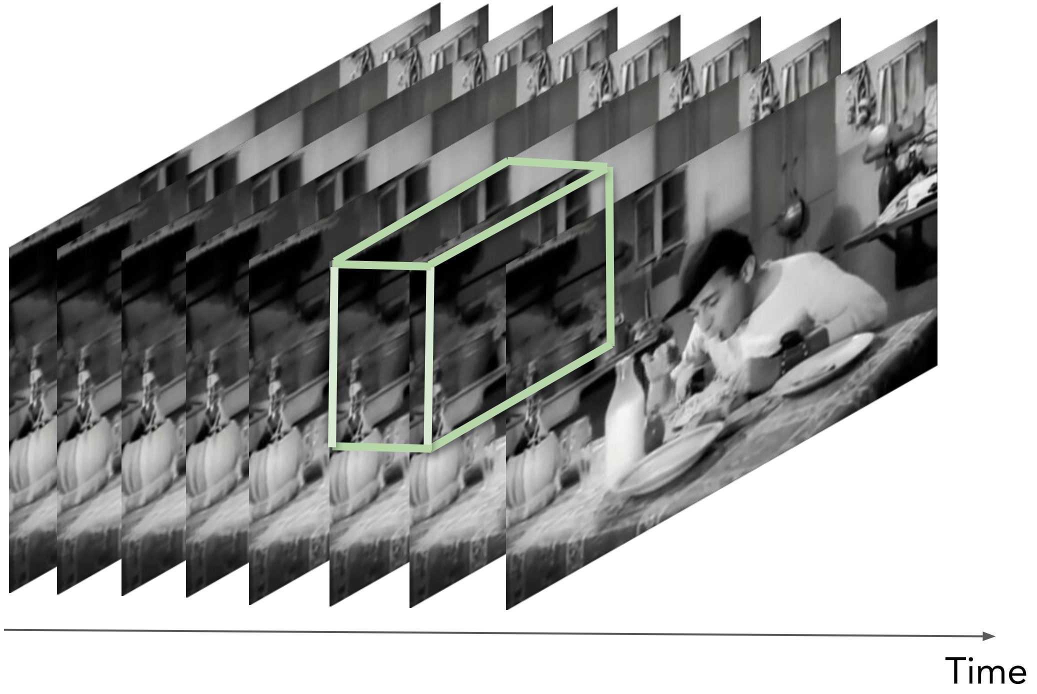



3D Convolution:

- The filter is 3-dimensional
- It slides in 3 directions


S. Ji, W. Xu, M. Yang, and K. Yu. *“3d convolutional neural networks for human action recognition.”*(ICML10)
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh *“Learning Spatio-Temporal Features With 3D Residual Networks for Action Recognition”*(ICCV17)

# First Solution

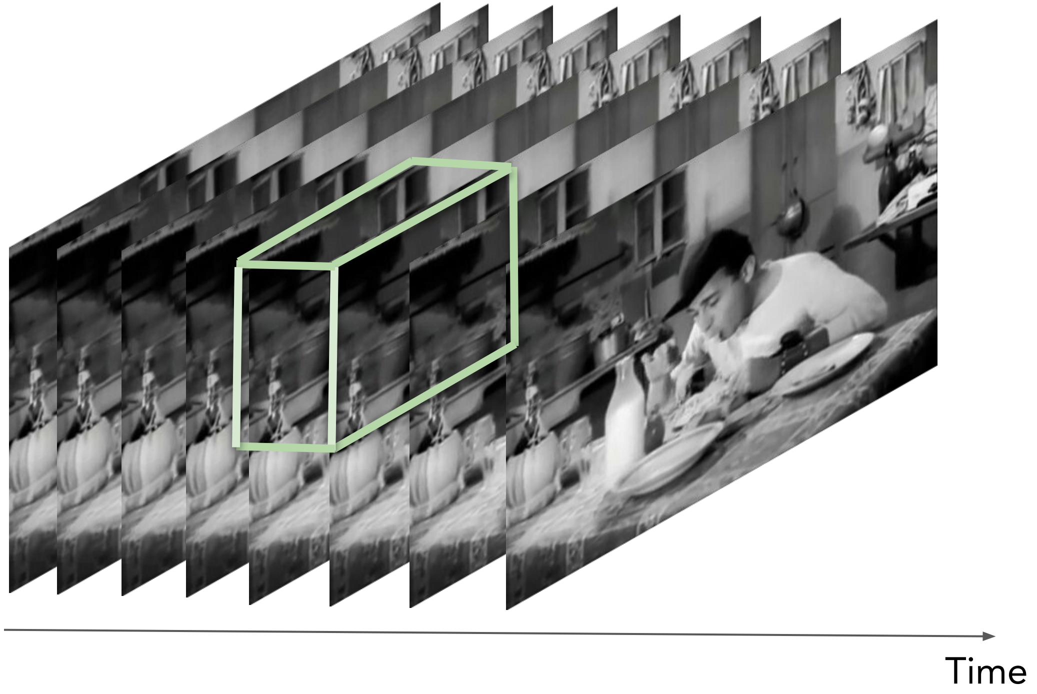



3D Convolution:

- The filter is 3-dimensional
- It slides in 3 directions

S. Ji, W. Xu, M. Yang, and K. Yu. *“3d convolutional neural networks for human action recognition.”*(ICML10)
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh *“Learning Spatio-Temporal Features With 3D Residual Networks for Action Recognition”*(ICCV17)

# First Solution

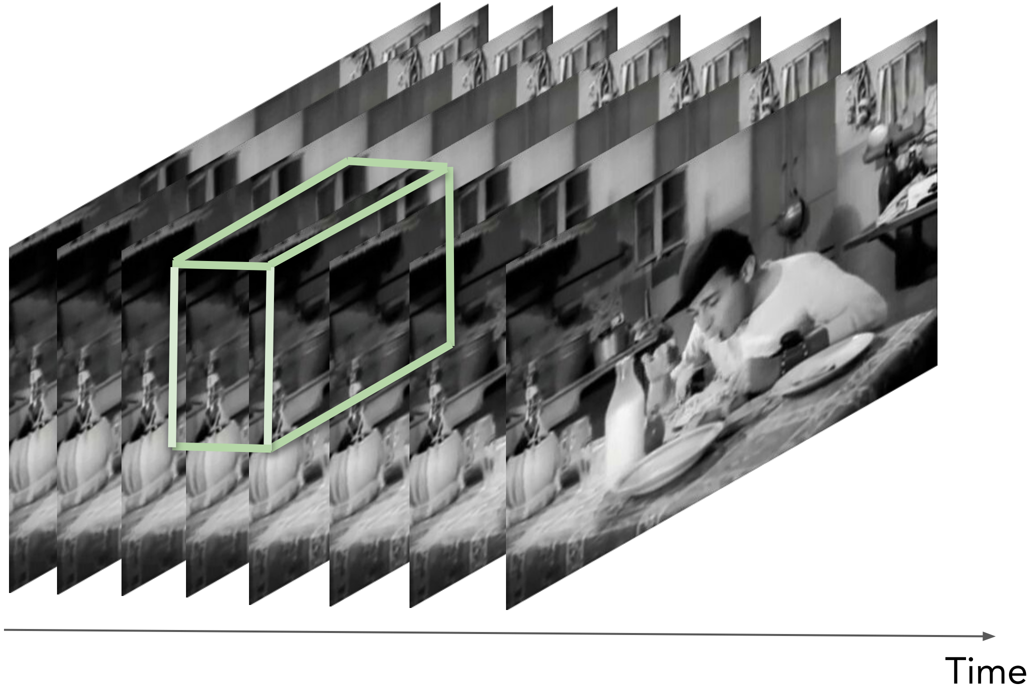



3D Convolution:

- The filter is 3-dimensional
- It slides in 3 directions

S. Ji, W. Xu, M. Yang, and K. Yu. *“3d convolutional neural networks for human action recognition.”*(ICML10)
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh *“Learning Spatio-Temporal Features With 3D Residual Networks for Action Recognition”*(ICCV17)

# First Solution

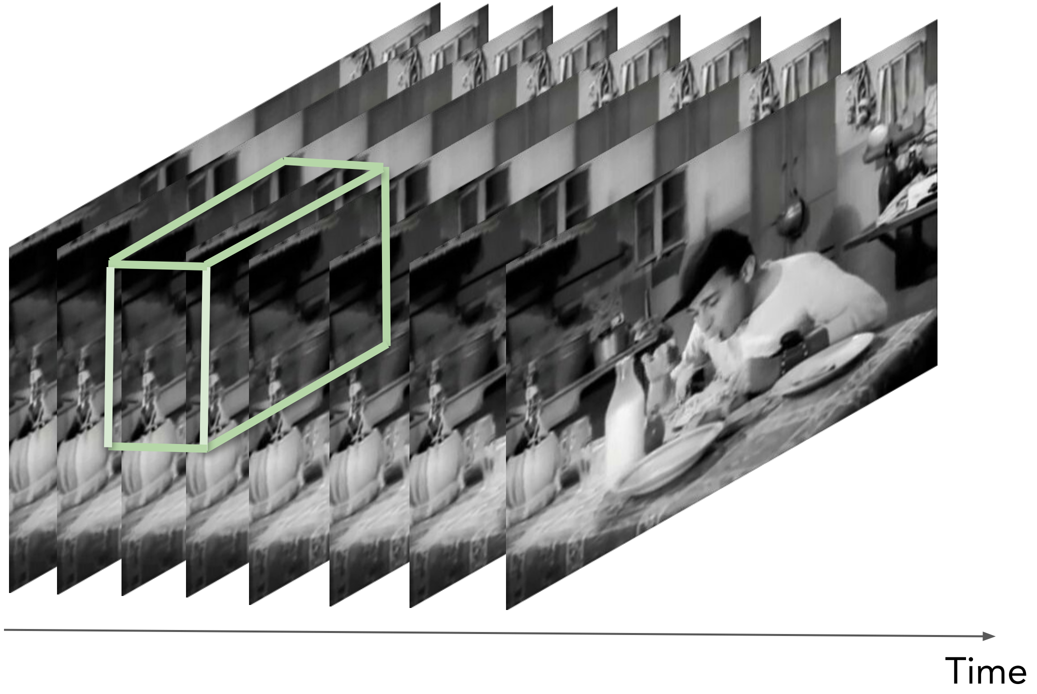



3D Convolution:

- The filter is 3-dimensional
- It slides in 3 directions

S. Ji, W. Xu, M. Yang, and K. Yu. *"3d convolutional neural networks for human action recognition."*(ICML10)
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh *"Learning Spatio-Temporal Features With 3D Residual Networks for Action Recognition"*(ICCV17)

# First Solution

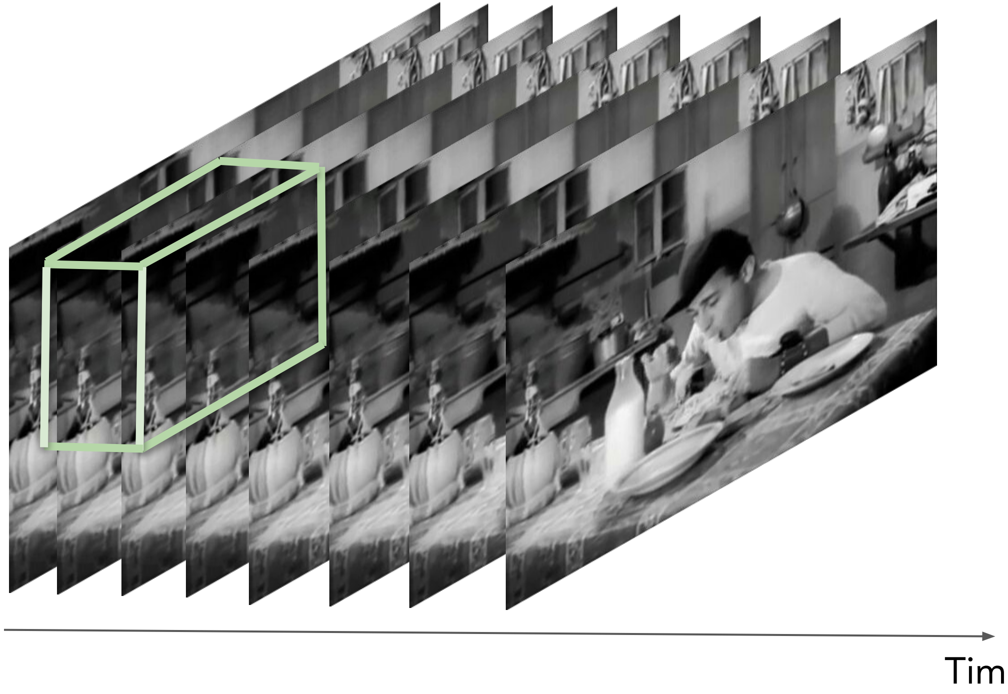



3D Convolution:

- The filter is 3-dimensional
- It slides in 3 directions

S. Ji, W. Xu, M. Yang, and K. Yu. *"3d convolutional neural networks for human action recognition."*(ICML10)
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh *"Learning Spatio-Temporal Features With 3D Residual Networks for Action Recognition"*(ICCV17)

# First Solution

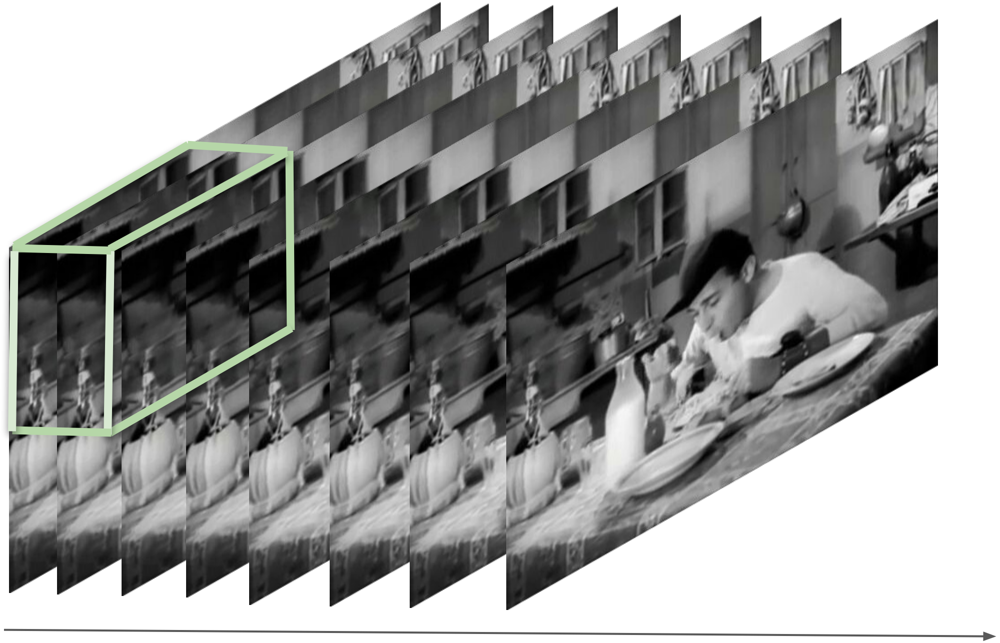



3D Convolution:

- The filter is 3-dimensional
- It slides in 3 directions

S. Ji, W. Xu, M. Yang, and K. Yu. *“3d convolutional neural networks for human action recognition.”*(ICML10)
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh *“Learning Spatio-Temporal Features With 3D Residual Networks for Action Recognition”*(ICCV17)

# First Solution



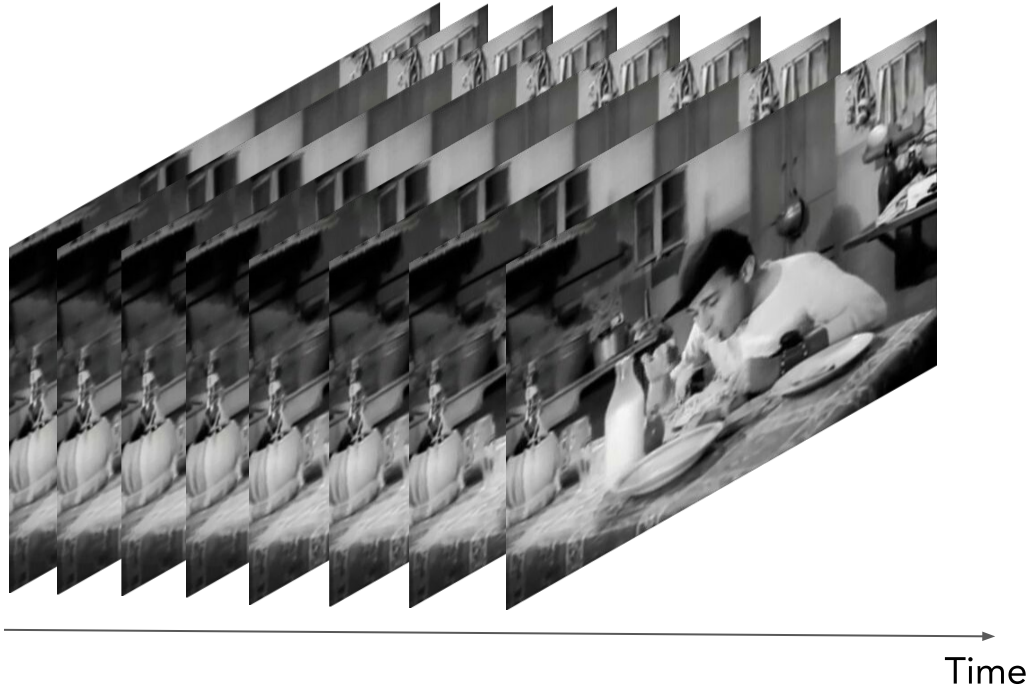
Time

3D Convolution:

- The filter is 3-dimensional
- It slides in 3 directions

S. Ji, W. Xu, M. Yang, and K. Yu. *“3d convolutional neural networks for human action recognition.”*(ICML10)
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh *“Learning Spatio-Temporal Features With 3D Residual Networks for Action Recognition”*(ICCV17)

# First Solution




3D Convolution:

- The filter is 3-dimensional
- It slides in 3 directions
- Pretrained?

S. Ji, W. Xu, M. Yang, and K. Yu. *"3d convolutional neural networks for human action recognition."*(ICML10)
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh *"Learning Spatio-Temporal Features With 3D Residual Networks for Action Recognition"*(ICCV17)

# Second Solution

# Second Solution

=

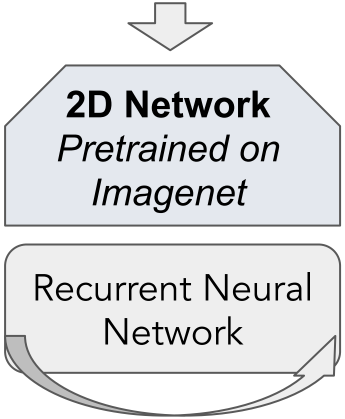
# Second Solution

=

**2D Network**
*Pretrained on Imagenet*

# Second Solution

=

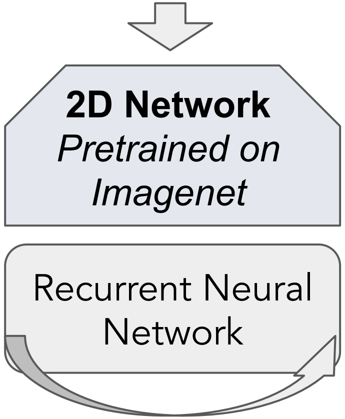


# Second Solution

=

+

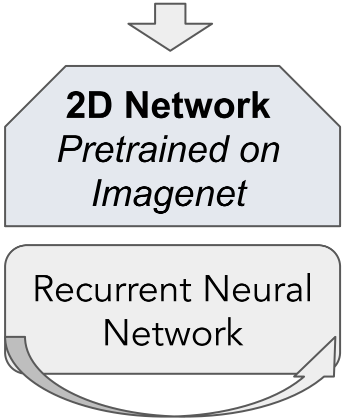
*Motion*



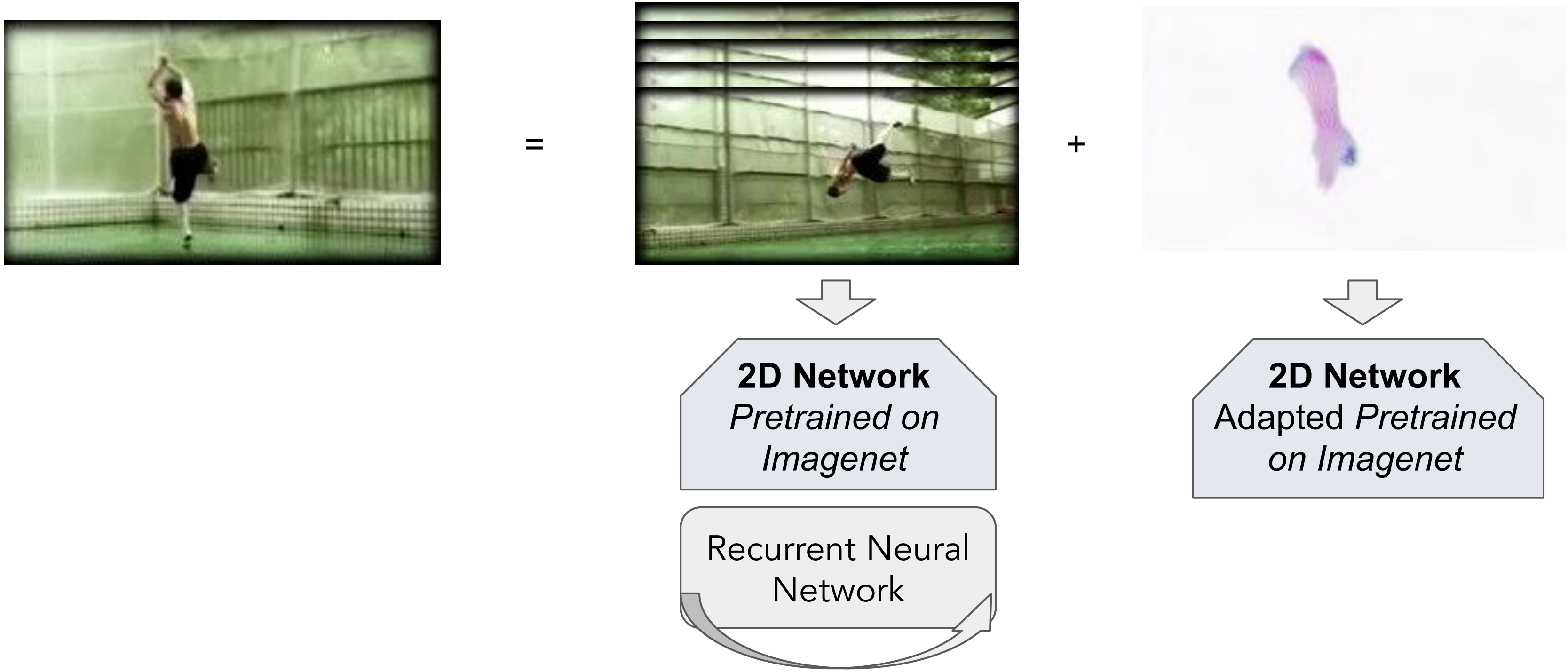
# Second Solution

=

+



# Second Solution



# Second Solution

=

+

**2D Network** *Pretrained on Imagenet*

Recurrent Neural Network

**2D Network** Adapted *Pretrained on Imagenet*

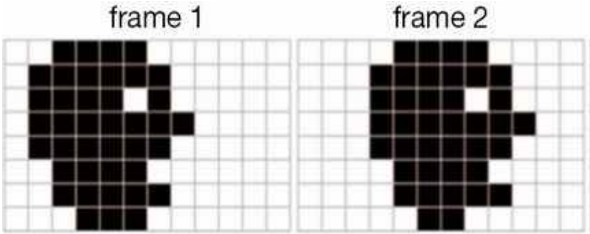
Introduction to Recurrent Neural Network: Understanding LSTM Networks

Jeffrey Donahue et. al. “Long-term Recurrent Convolutional Networks for Visual Recognition and Description.” (CVPR2015)

# Optical Flow

It extracts image motion, at each pixel, from spatio-temporal image brightness variations.

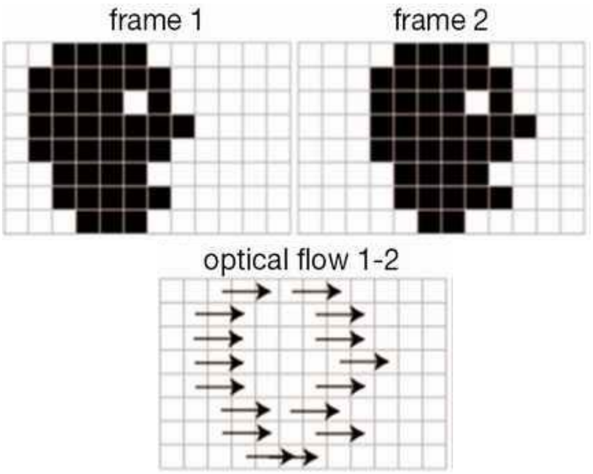
# Optical Flow




It extracts image motion, at each pixel, from spatio-temporal image brightness variations.

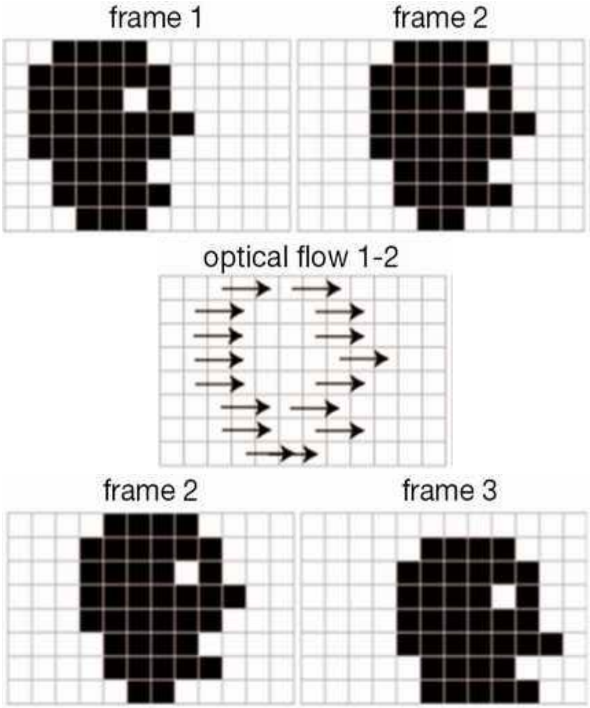
# Optical Flow

It extracts image motion, at each pixel, from spatio-temporal image brightness variations.



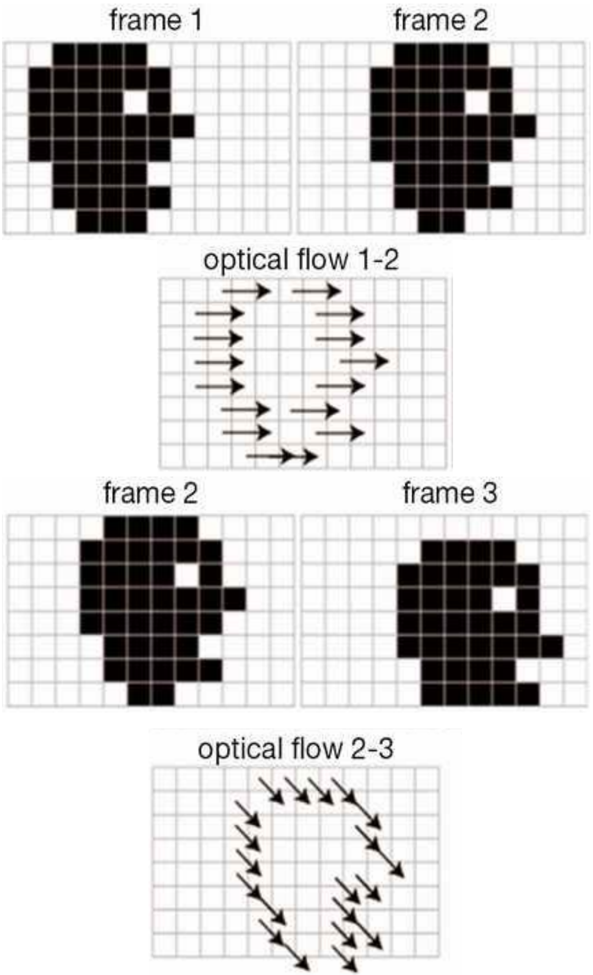
# Optical Flow

It extracts image motion, at each pixel, from spatio-temporal image brightness variations.



# Optical Flow

It extracts image motion, at each pixel, from spatio-temporal image brightness variations.

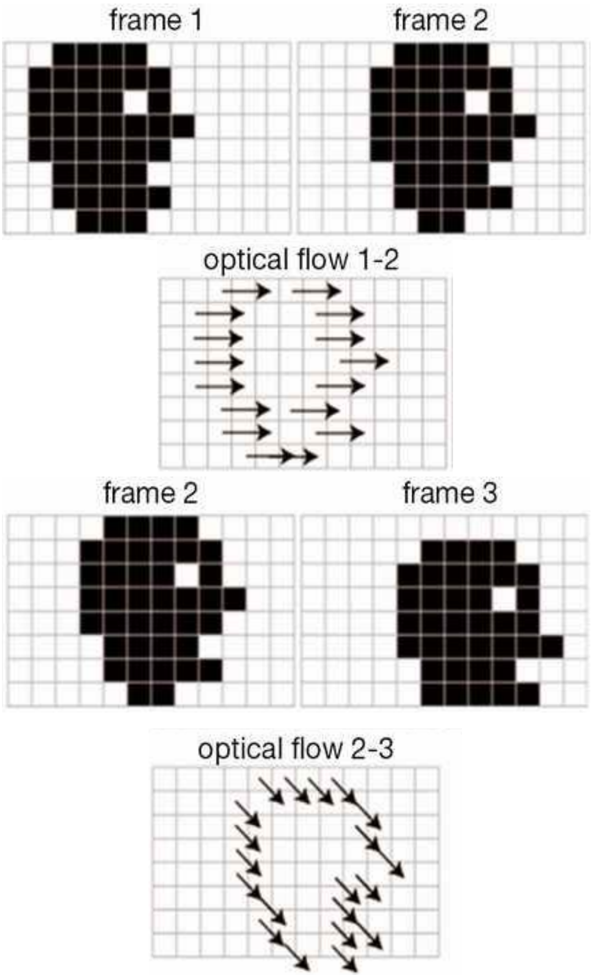


# Optical Flow

It extracts image motion, at each pixel, from spatio-temporal image brightness variations.

*Brightness constancy constraint*

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t)$$



# Optical Flow

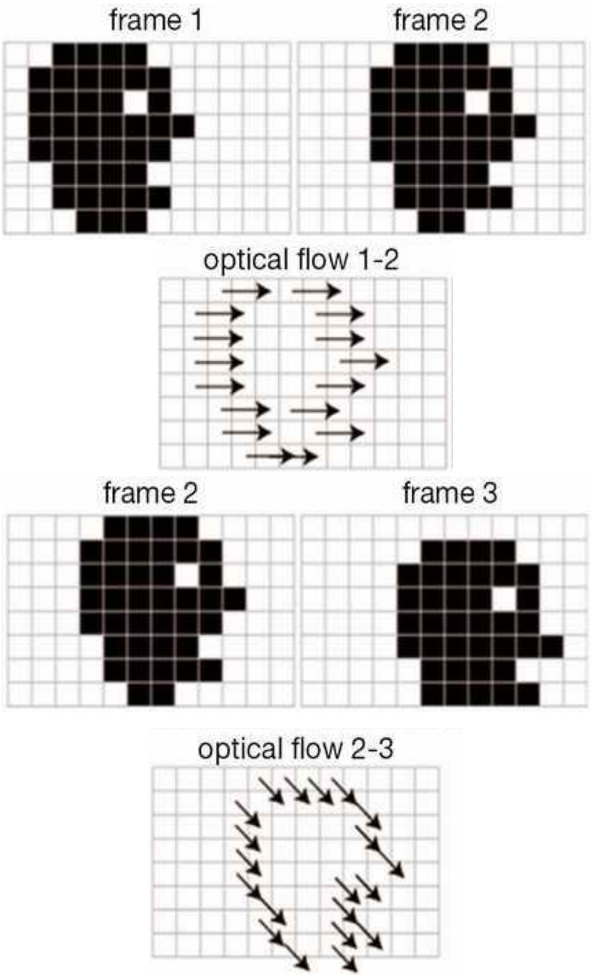
It extracts image motion, at each pixel, from spatio-temporal image brightness variations.

*Brightness constancy constraint*

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t)$$

*Aperture problem*:

- Lucas–Kanade method



# First Person Action Recognition

# Main Task

- First Person Action Recognition

# Main Task

- First Person Action Recognition
- More challenging with respect to Third Person:

# Main Task

- First Person Action Recognition
- More challenging with respect to Third Person:
    - Strong Occlusion

# Main Task

- First Person Action Recognition
- More challenging with respect to Third Person:
  - Strong Occlusion
  - Camera Motion

# Main Task

- First Person Action Recognition
- More challenging with respect to Third Person:
  - Strong Occlusion
  - Camera Motion
- Why is the ego-centric view important ?

# Main Task

- First Person Action Recognition
- More challenging with respect to Third Person:
    - Strong Occlusion
    - Camera Motion
- Why is the ego-centric view important ?

# Main Task

- First Person Action Recognition
- More challenging with respect to Third Person:
    - Strong Occlusion
    - Camera Motion
- Why is the ego-centric view important ?

# Before Starting

# Ego-Rnn

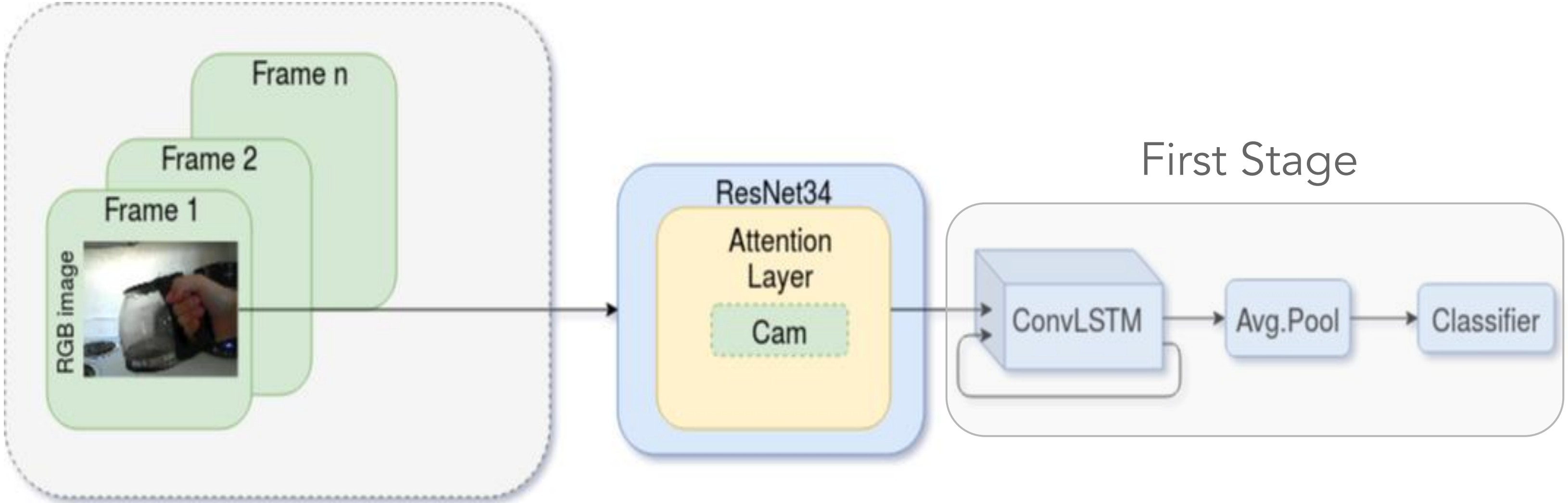
Two Stream Network:

- Network to encode the Appearance:
    - Two stages training
    - convLSTM
    - Class activation maps (CAM) like attention mechanism

- Network to encode the motion (Warp Flow):

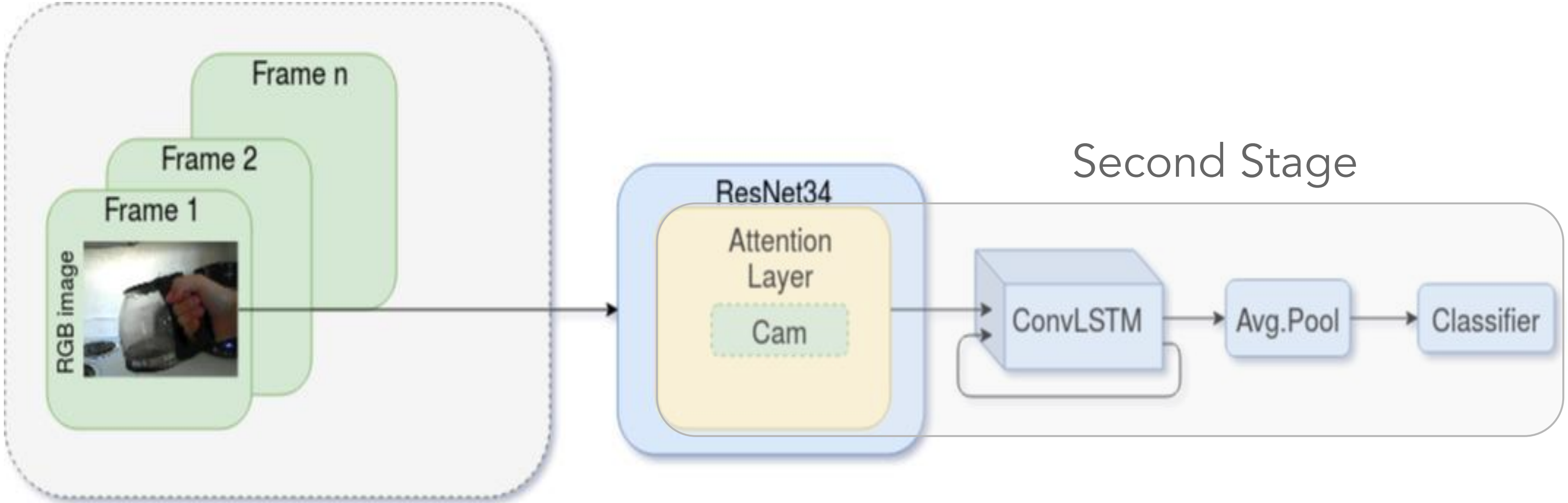
S. Sudhakaran and O. Lanz. “*Attention is all we need: Nailing down object-centric attention for egocentric activity recognition*.” BMVC18

# Ego-Rnn




S. Sudhakaran and O. Lanz. “*Attention is all we need: Nailing down object-centric attention for egocentric activity recognition*.” BMVC18

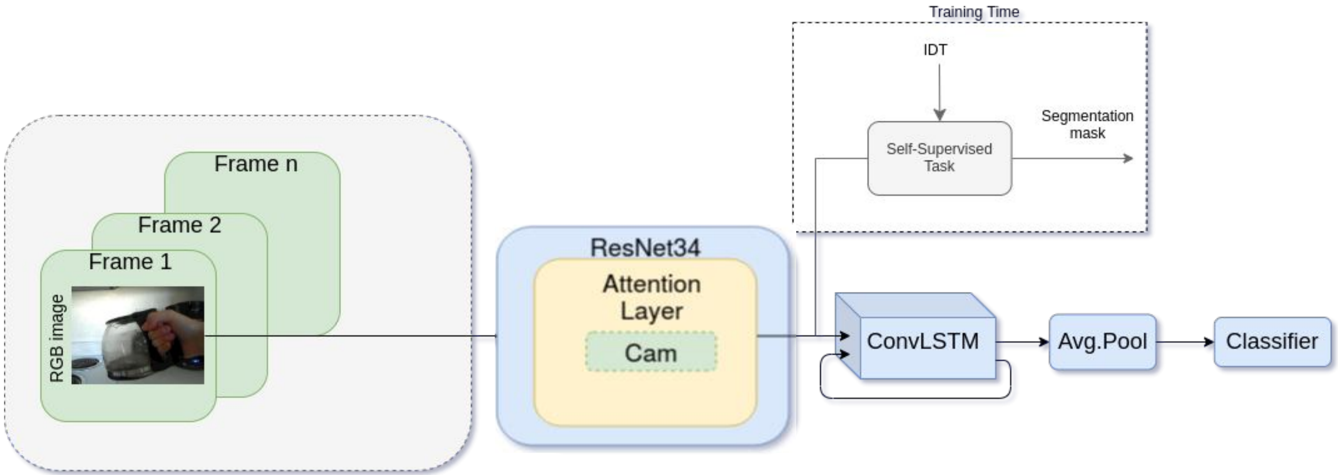
# Ego-Rnn




S. Sudhakaran and O. Lanz. "*Attention is all we need: Nailing down object-centric attention for egocentric activity recognition*." BMVC18

# Self-Supervised Task

Adding a Self-Supervised task to learn the motion together with the appearance




M. Planamente et al. "*Joint Encoding of Appearance and Motion Features with Self-supervision for First Person Action Recognition*"

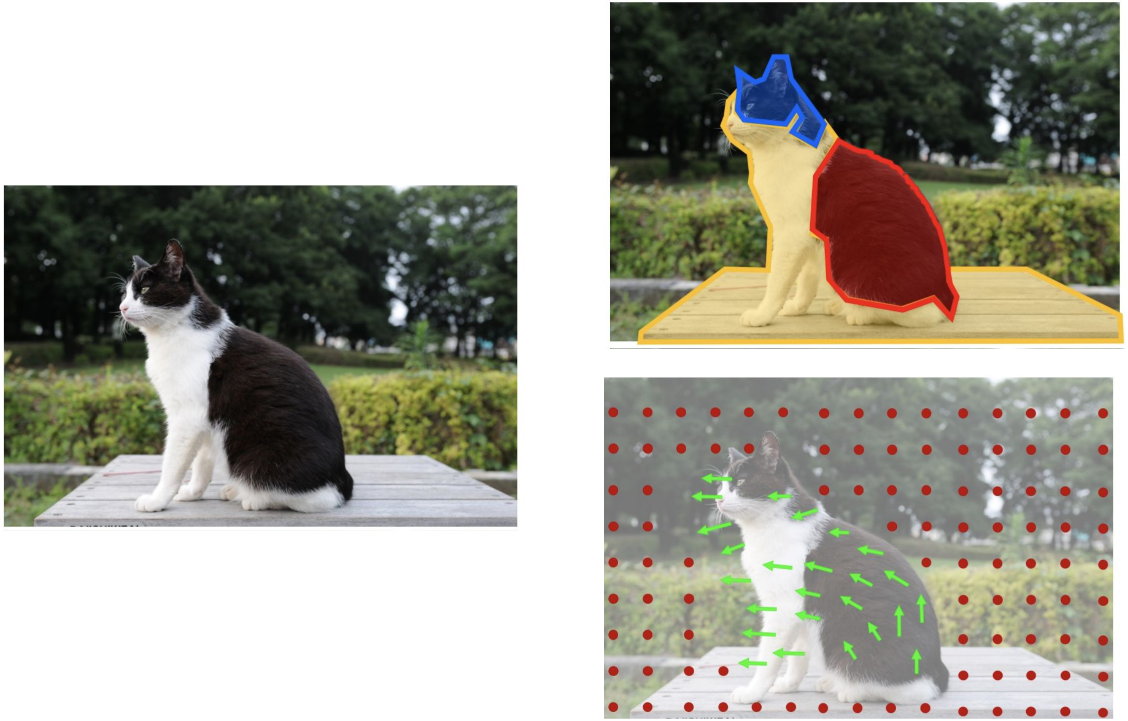
# Self-Supervised Task

Deepak Pathak, Ross B. Girshick, Piotr Dollar, Trevor Darrell, and Bharath Hariharan. *"Learning features by watching objects move."* CVPR 2017

# Self-Supervised Task

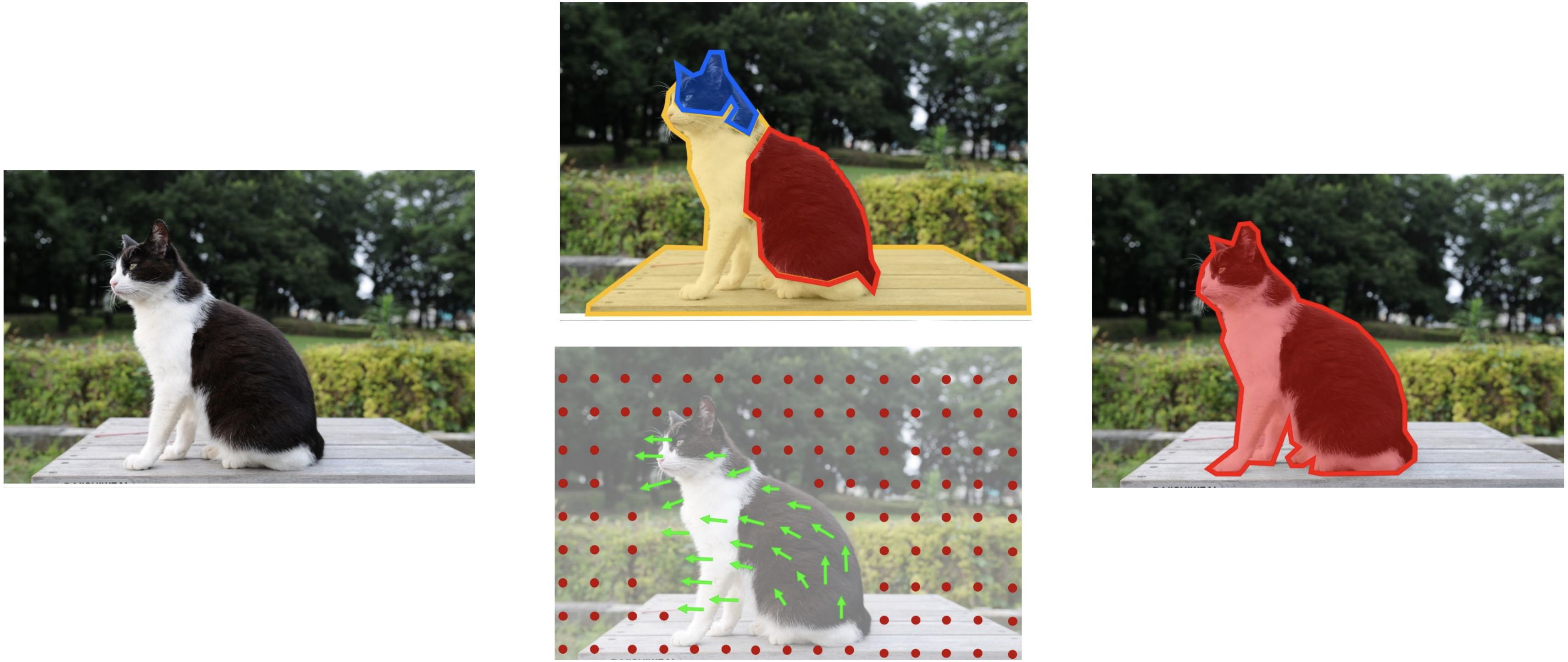
Deepak Pathak, Ross B. Girshick, Piotr Dollar, Trevor Darrell, and Bharath Hariharan. *"Learning features by watching objects move."* CVPR 2017

# Self-Supervised Task



Deepak Pathak, Ross B. Girshick, Piotr Dollar, Trevor Darrell, and Bharath Hariharan. *"Learning features by watching objects move."* CVPR 2017

# Self-Supervised Task



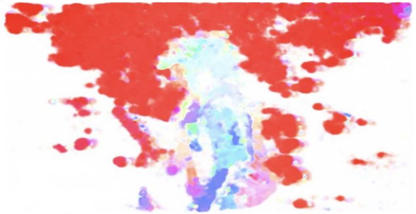
Deepak Pathak, Ross B. Girshick, Piotr Dollar, Trevor Darrell, and Bharath Hariharan. *“Learning features by watching objects move.”* CVPR 2017

# Improved Dense Trajectories (IDT)


Heng Wang and Cordelia Schmid. *"Action recognition with improved trajectories."* ICCV13

# Improved Dense Trajectories (IDT)

Optical Flow




Heng Wang and Cordelia Schmid. *“Action recognition with improved trajectories.”* ICCV13

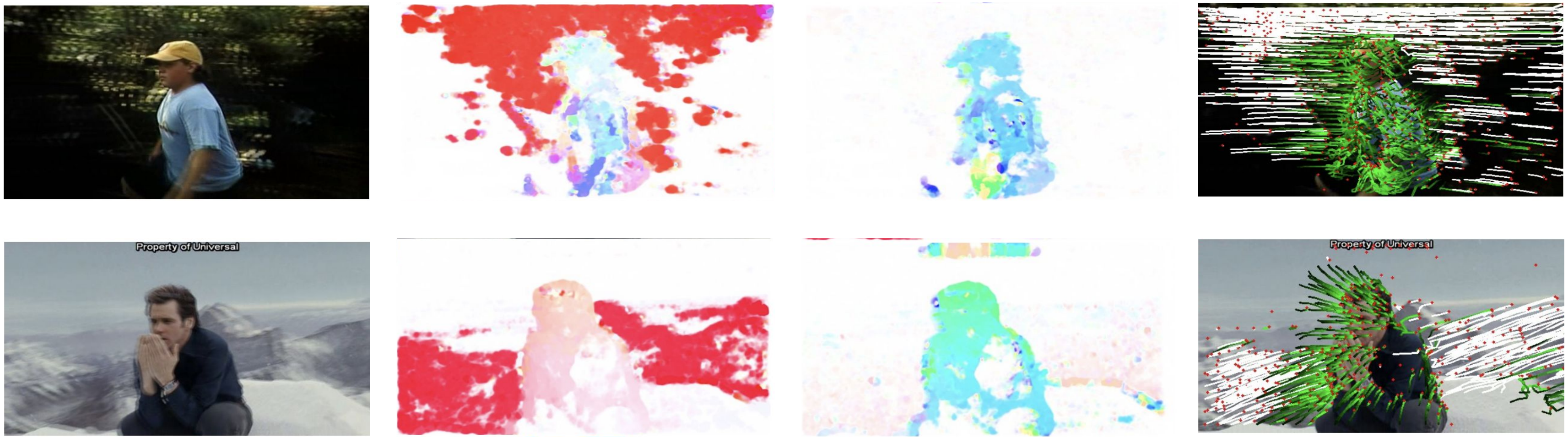
# Improved Dense Trajectories (IDT)

Optical Flow Warp Flow


Heng Wang and Cordelia Schmid. *"Action recognition with improved trajectories."* ICCV13

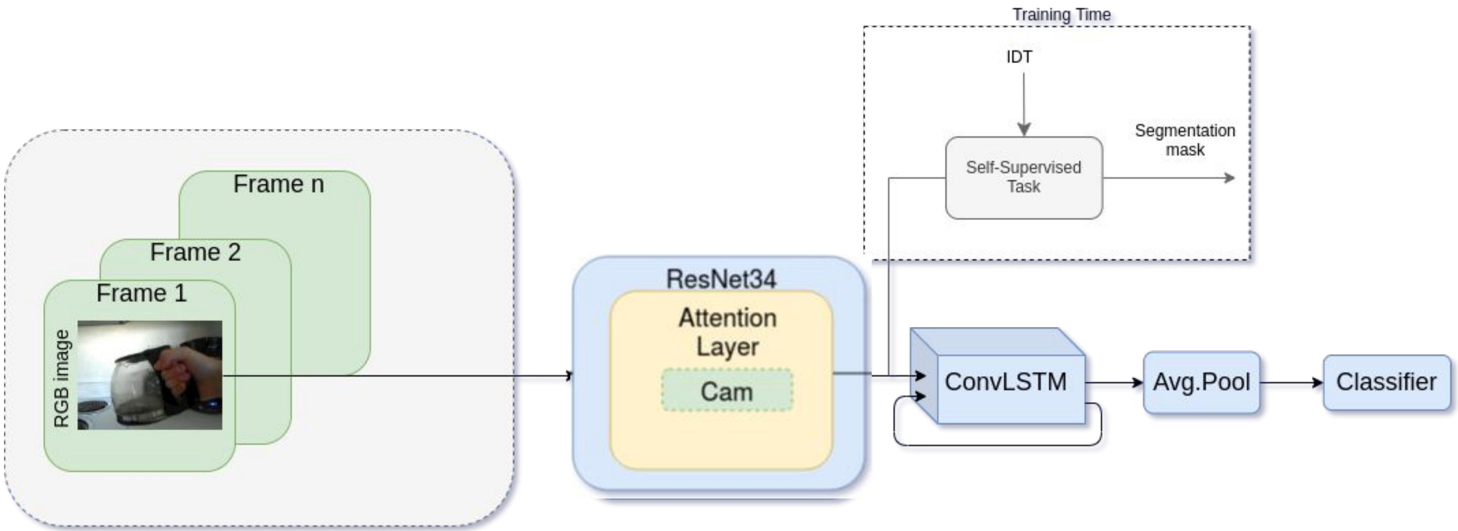
# Improved Dense Trajectories (IDT)

Optical Flow | Warp Flow | IDT




Heng Wang and Cordelia Schmid. *“Action recognition with improved trajectories.”* ICCV13

# Training




<u>Loss</u> = Loss Classification + Loss Secondary task

# Project Description

# Dataset

Gtea61:

- 61 actions
- 4 users:
    - S1, S2, S3, S4

Training sets = S1 S3 S4 (labeled)

Validation set = test set = S2.

***Hint**: the dataLoaders of the original repository of Ego-RNN require some modifications to work with the data that are provided at this link (data are in .png format)*

Dataset: https://drive.google.com/drive/folders/1_NAcoR0UGH1eLsiWMOx_Py8yeAocknA2?usp=sharing

# Project Description

**Step 1:**

- Ego-RNN

**Step 2:**

- Self-supervised task

**Step 3:**

- Propose your own variation

# Step 1

- Complete the table below using the same parameters of [1] (run these experiments with fewer than 25 frames, for example 7 - 16)

| **Configurations** | **Accuracy%** 7 Frames | **Accuracy%** 16 Frames |
| --- | --- | --- |
| ConvLSTM | | |
| ConvLSTM-attention | | |
| Temporal-warp flow | | |
| two-stream (joint train) | | |

# Step 2

- Implement the self-supervised task proposed in [2] and integrate it on the second stage of the Ego-RNN network (with the same training procedure proposed in [1]).

- CAM visualization like in Ego-Rnn paper when the self-supervised task is added.

- Implement the same self-supervised task as a regression problem. This variant requires you to choose a correct activation function for the regressor and a suitable loss function for the problem.

- Run an appropriate hyperparameter optimization for the self-supervised task: for each method you have to choose at least 3 different sets of hyperparameters.

# Step 3

It is your turn to propose an improvement!

# Questions you should be able to answer at the end of the project.

What are the differences between RGB and Optical Flow network?

How does it work the attention mechanism?

Why in this work [1] do they use a ConvLSTM and how does it work?

Which is the motivation behind the use of the self-supervised task in [2]?

What are the differences between Optical Flow and Warp Flow?

# References

[1] Attention is All We Need: Nailing Down Object-centric Attention for Egocentric Activity Recognition BMVC18

[2] Joint Encoding of Appearance and Motion Features with Self-supervision for First Person Action Recognition

[3] Heng Wang and Cordelia Schmid. *“Action recognition with improved trajectories.”* ICCV13

[4] Deepak Pathak, Ross B. Girshick, Piotr Dollar, Trevor Darrell, and Bharath Hariharan. *“Learning features by watching objects move.”* CVPR 2017