Video Understanding in Deep Learning Era

Ali Diba

Video Classification

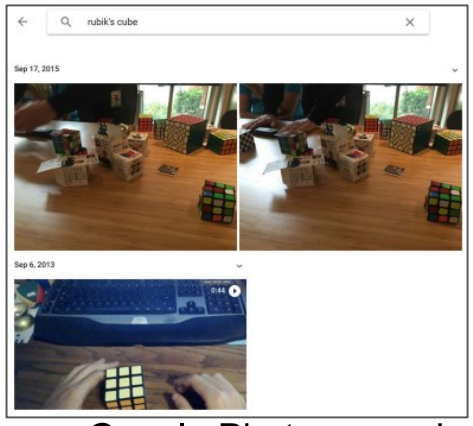
- Datasets (YouTube8M, Sport1M, Kinetics)
- 2D Convolutional Neural Network Architectures.
  - Feature Pooling Methods
- 3D Convolutional Neural Network Architectures.
- Unsupervised Methods.
- Applications

Traditional Computer vision and Machine learning

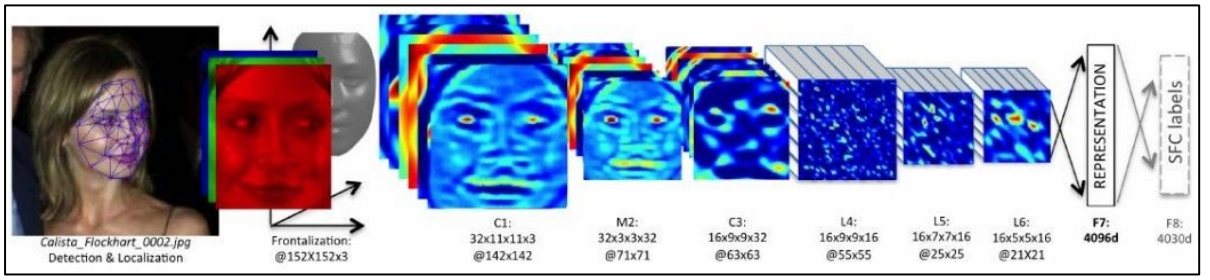
Computer vision
Less than 5 years ago

GIST, HOG, SIFT, LBP, SSIM, Line Features, Textons, Color Hist, …

Now Everywhere: Deep Learning



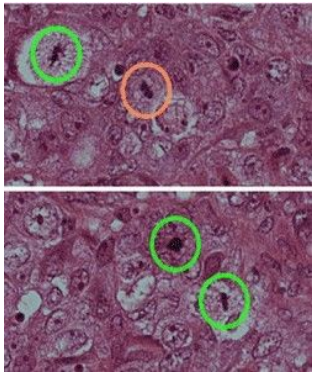
e.g. Google Photos search



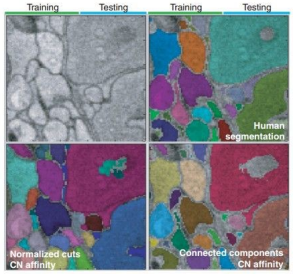
Face Verification, Taigman et al. 2014 (FAIR)

[Goodfellow et al. 2014]

Self-driving cars



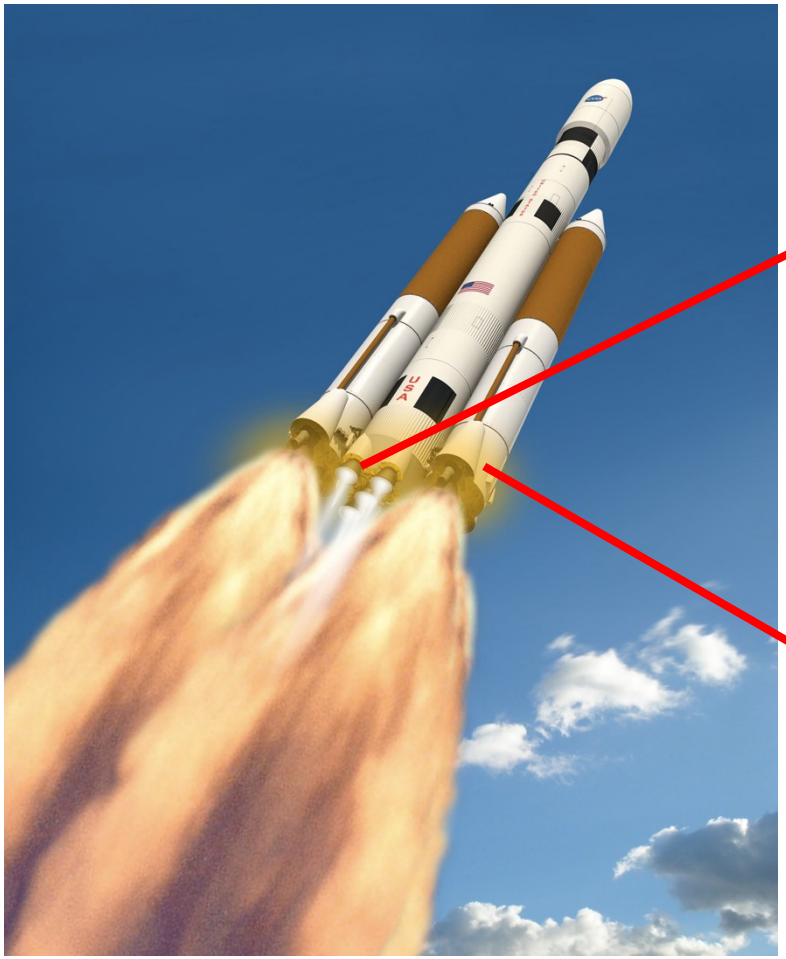
Ciresan et al. 2013



Turaga et al 2010

From Karpathy's slide

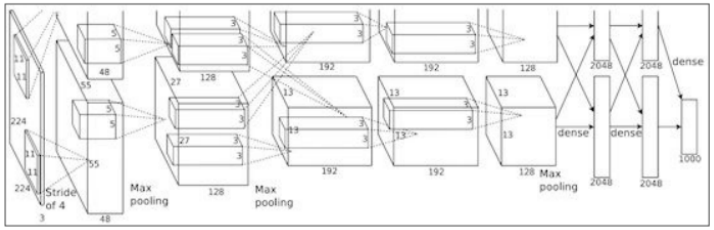
Deep Learning Era



Engine

NVIDIA et al.

"AlexNet"



Fuel

YouTube 8M

al.

YouTube

Datasets

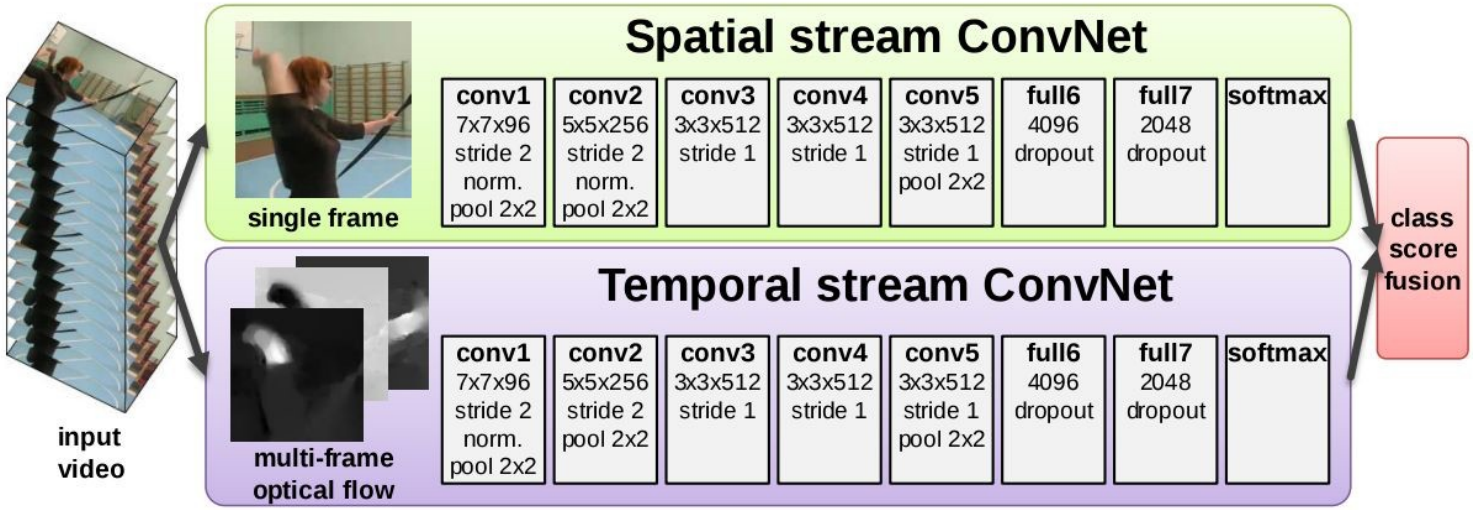
	Actions per video	Classes	Labelled instances	Total videos	Origin	Type	Temporal localization
Charades v1.0	6.8	157	67K	10K	267 Homes	Daily Activities	Yes
ActivityNet [3]	1.4	203	39K	28K	YouTube	Human Activities	Yes
UCF101 [8]	1	101	13K	13K	YouTube	Sports	No
HMDB51 [7]	1	51	7K	7K	YouTube/Movies	Movies	No
THUMOS'15 [5]	1-2	101	21K+	24K	YouTube	Sports	Yes
Sports 1M [6]	1	487	1.1M	1.1M	YouTube	Sports	No
MPII-Cooking [14]	46	78	13K	273	30 In-house actors	Cooking	Yes
ADL [25]	22	32	436	20	20 Volunteers	Ego-centric	Yes
MPII-MD [11]	Captions	Captions	68K	94	Movies	Movies	No

Datasets

	Classes	Number of Videos	Source	Type
Youtube8M	~5000 Concepts	~ 8 milion	Youtube	Any Concept
KINETICS	400	240 K	Youtube	Human Actions

Two stream Networks for Classification

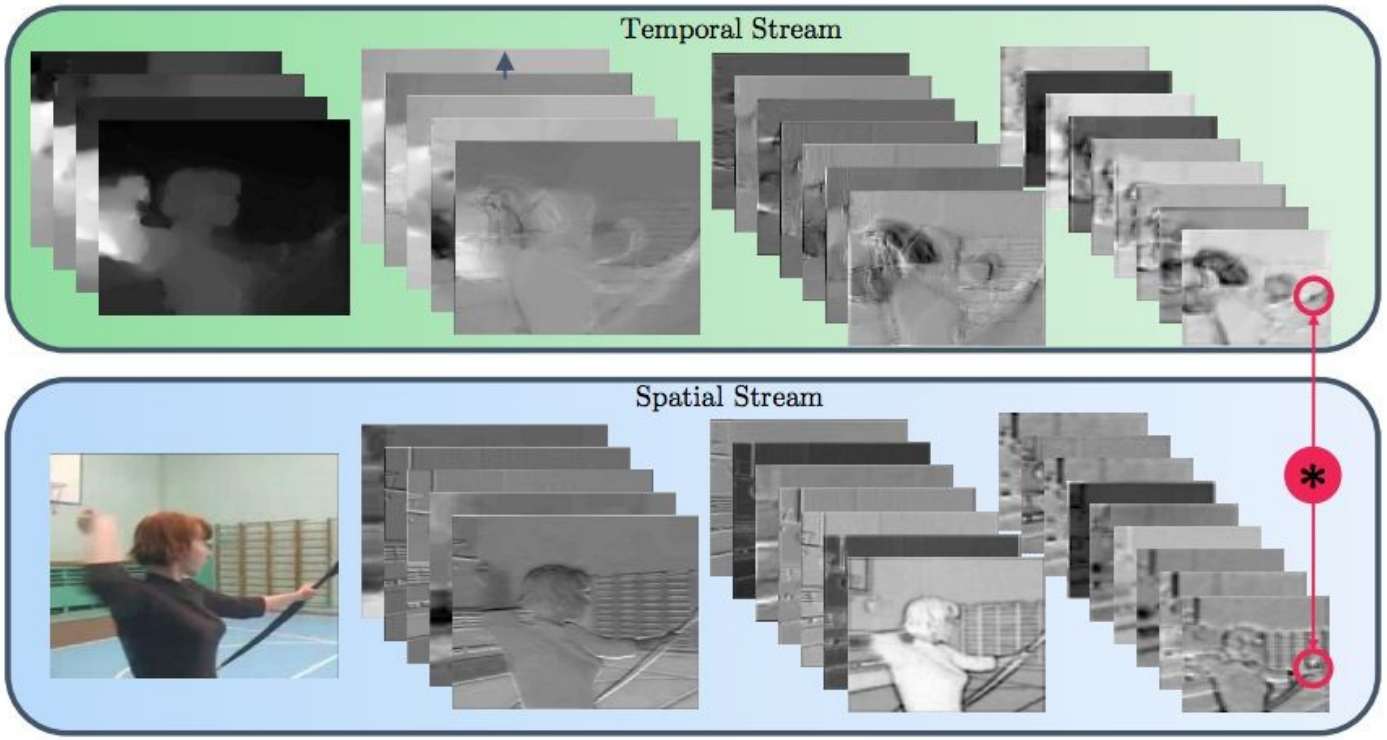
- Two Branch of CNNs
  - RGB stream
  - Optical Flow (motion) Stream
  - Pre-trained on ImageNet
- Score Fusion



Simonian, et al “Two-Stream Convolutional Networks for Action Recognition in Videos”

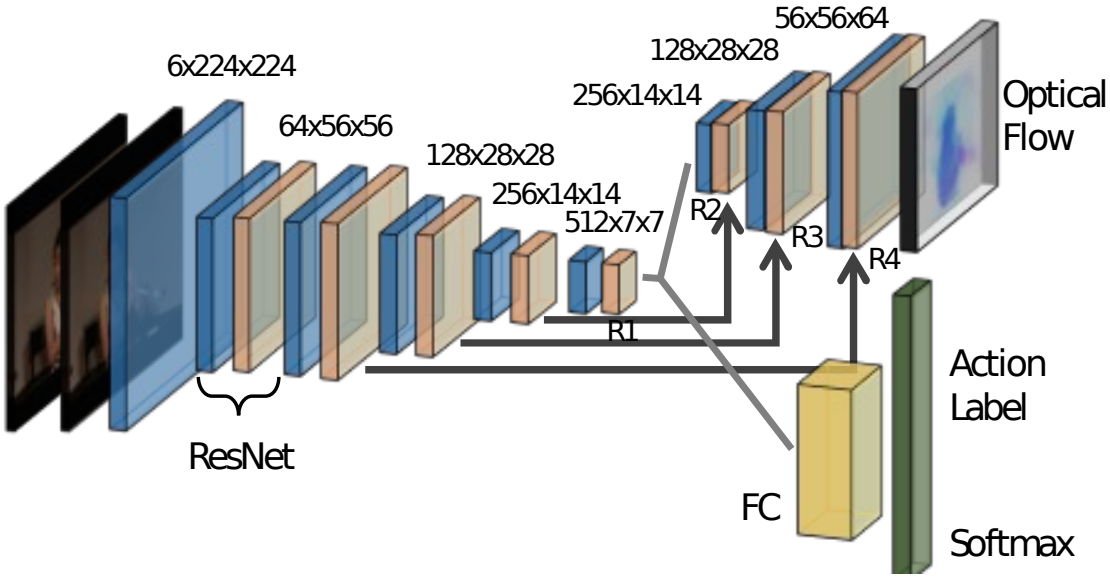
Two stream Networks for Classification

- Two Branch of CNNs
  - RGB stream
  - Optical Flow (motion) Stream
  - Pre-trained on ImageNet



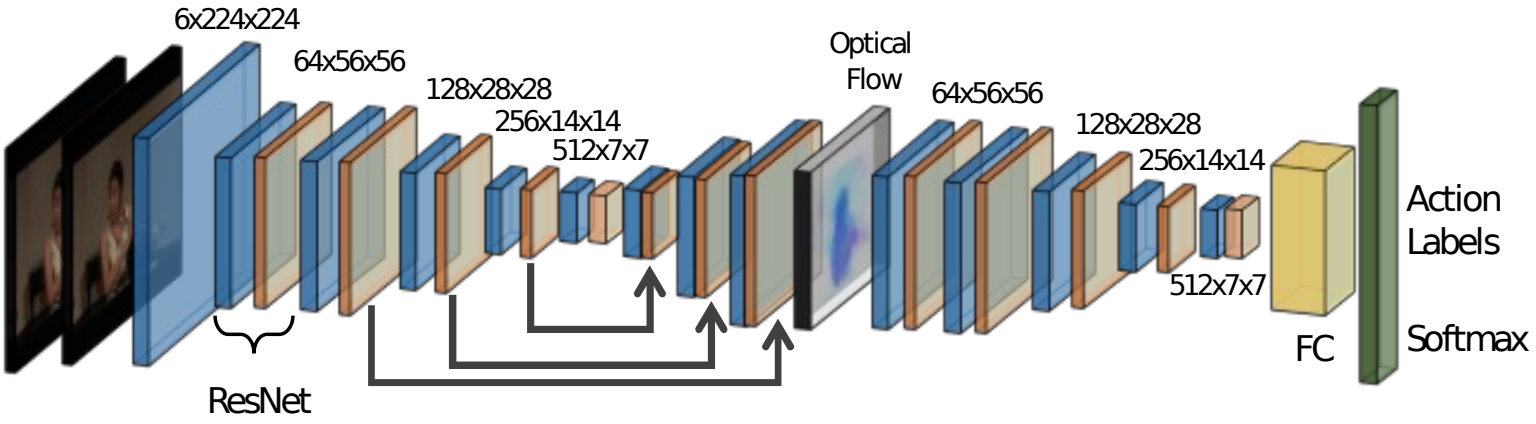
ActionFlowNet

- Learning Action and Motion together



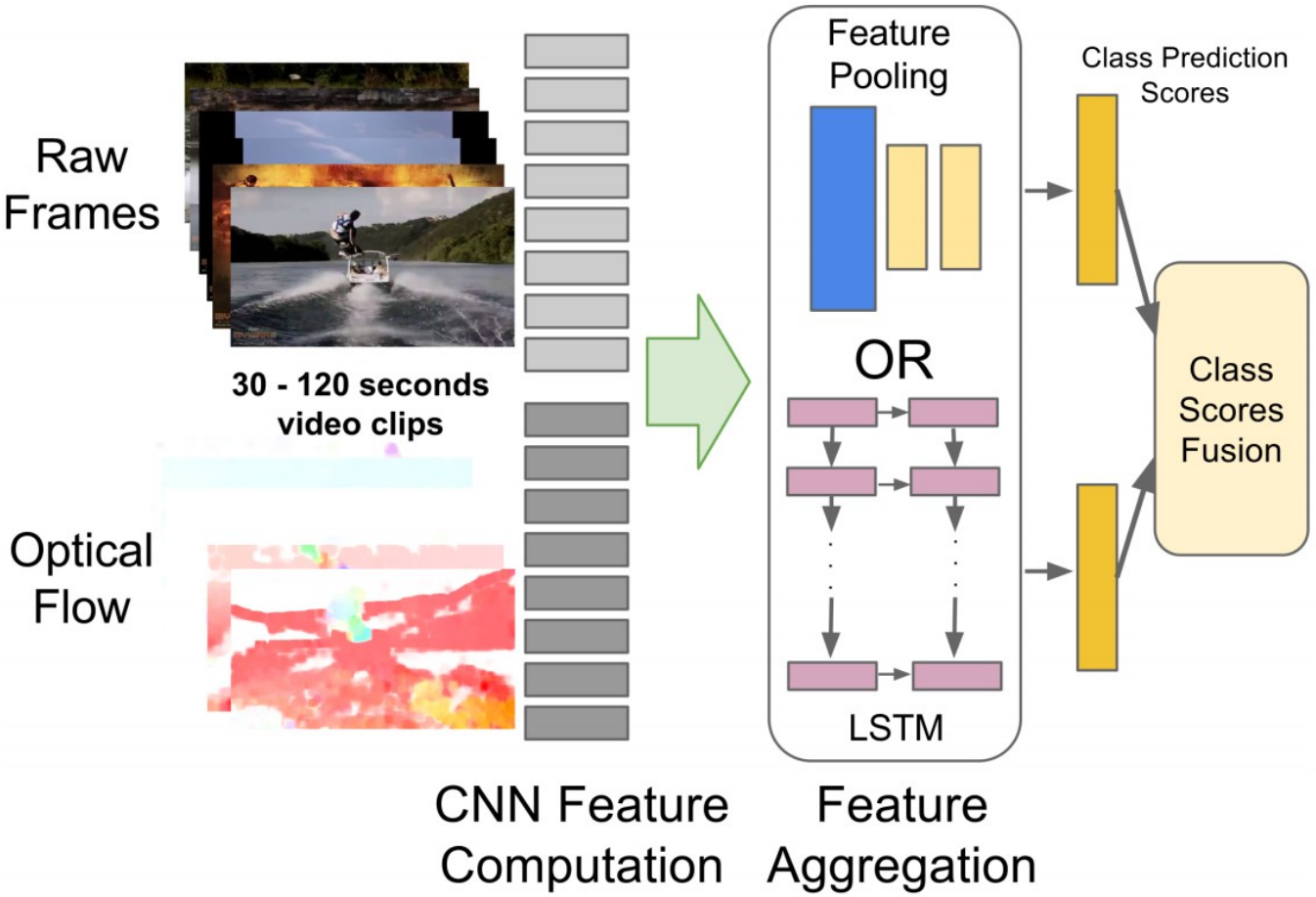
Ng, et al “ActionFlowNet: Learning Motion Representation for Action Recognition”

ActionFlowNet



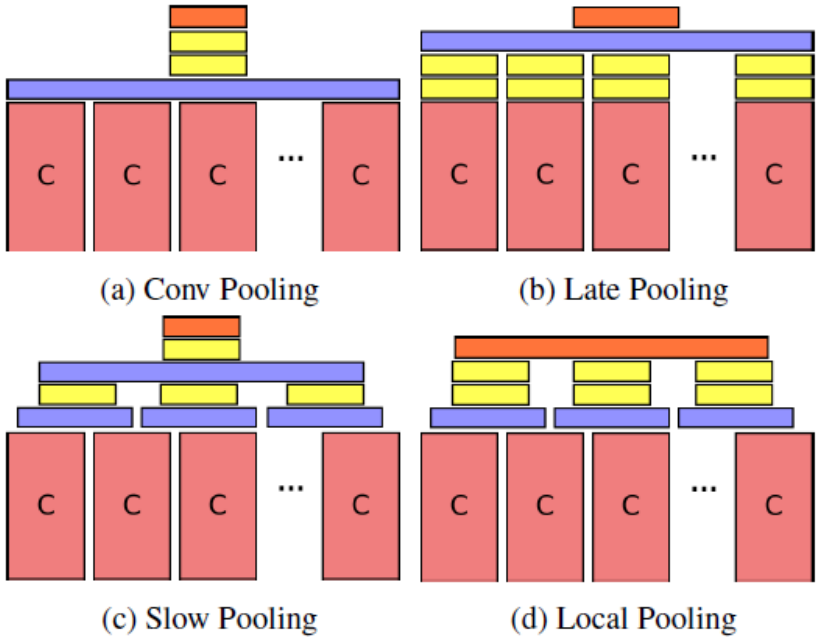
Ng, et al “ActionFlowNet: Learning Motion Representation for Action Recognition”

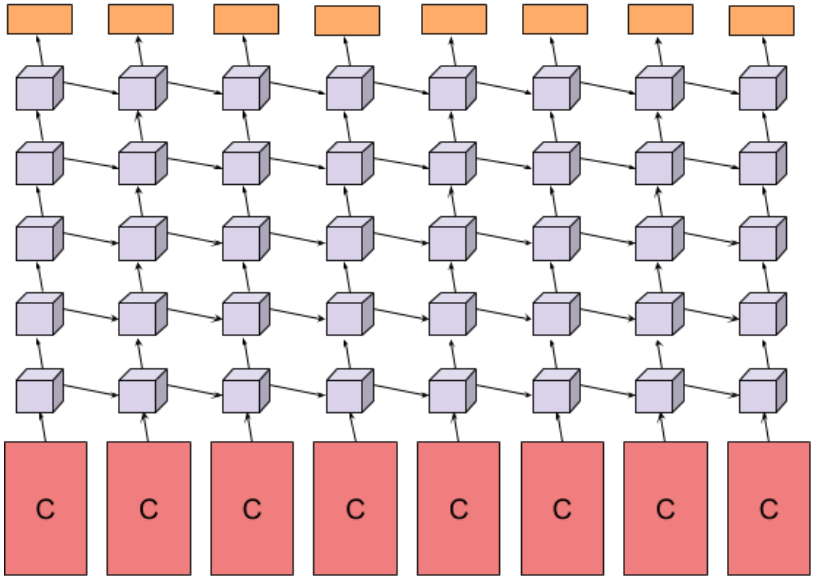
Video Clip processing



Ng, et al “Beyond Short Snippets: Deep Networks for Video Classification”

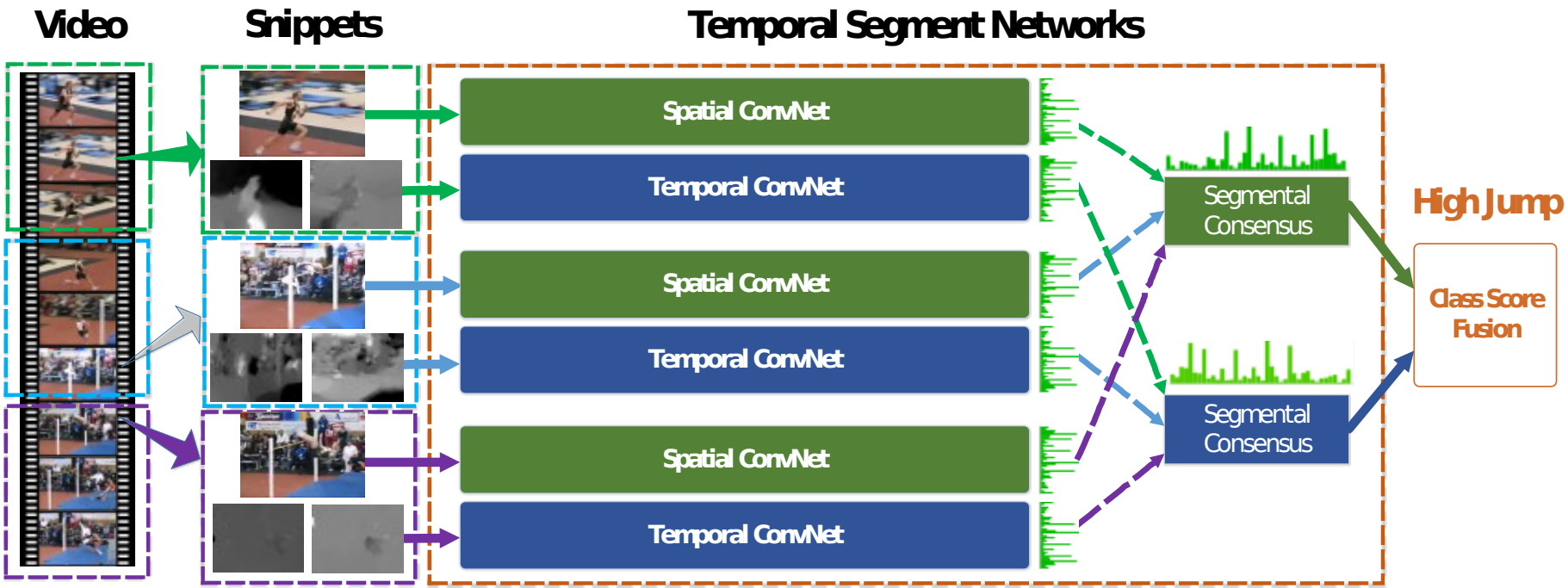
Video Clip processing





Deep Video LSTM takes input the output from the final CNN layer at each consecutive video frame. CNN outputs are processed forward through time and upwards through five layers of stacked LSTMs

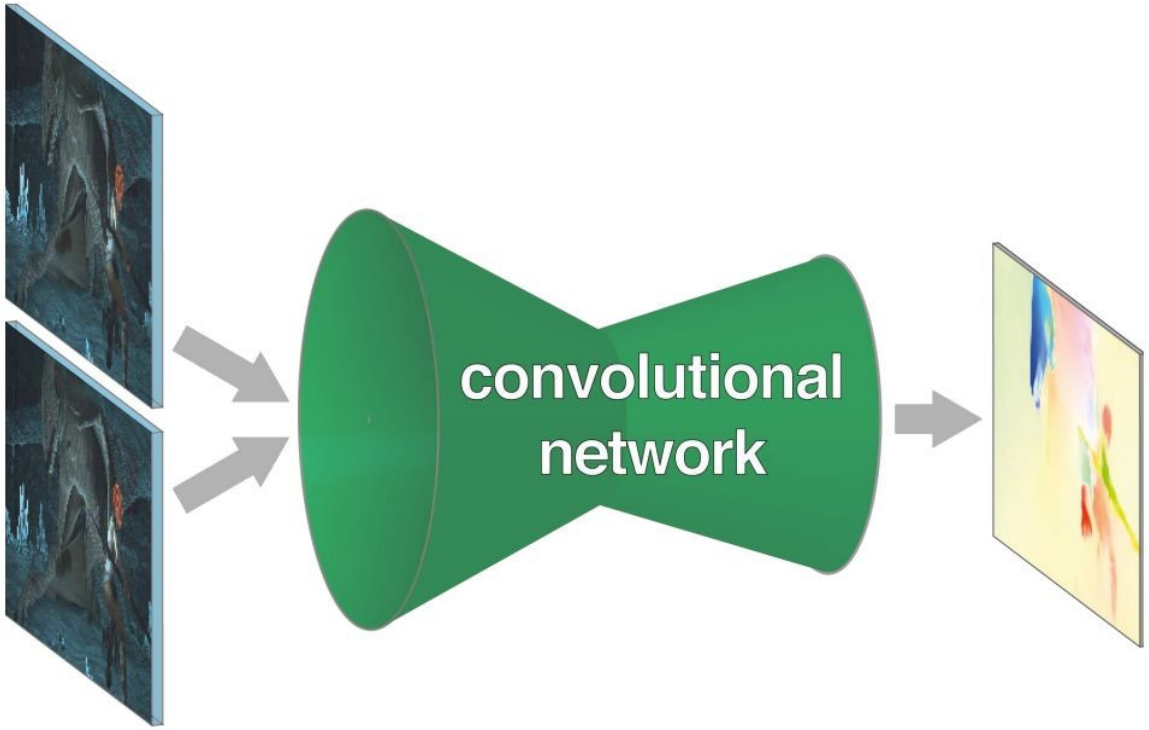
Temporal Segments Network



Wang, et al “Temporal Segment Networks: Towards Good Practices for Deep Action Recognition”

Optical Flow Learning

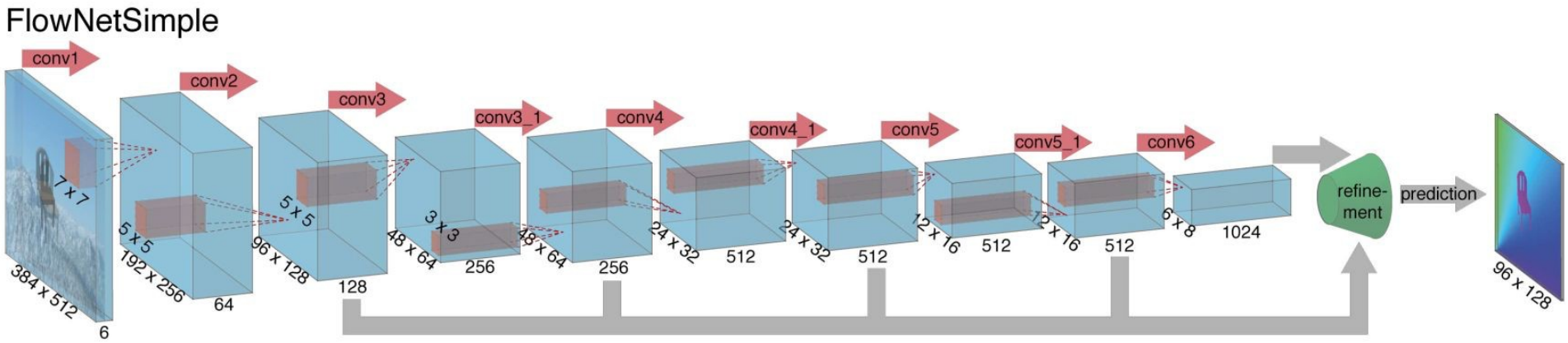
- Supervised OpticalFlow Deep CNN:



Dosovitskiy, et al “FlowNet: Learning Optical Flow With Convolutional Networks ”

Optical Flow Learning

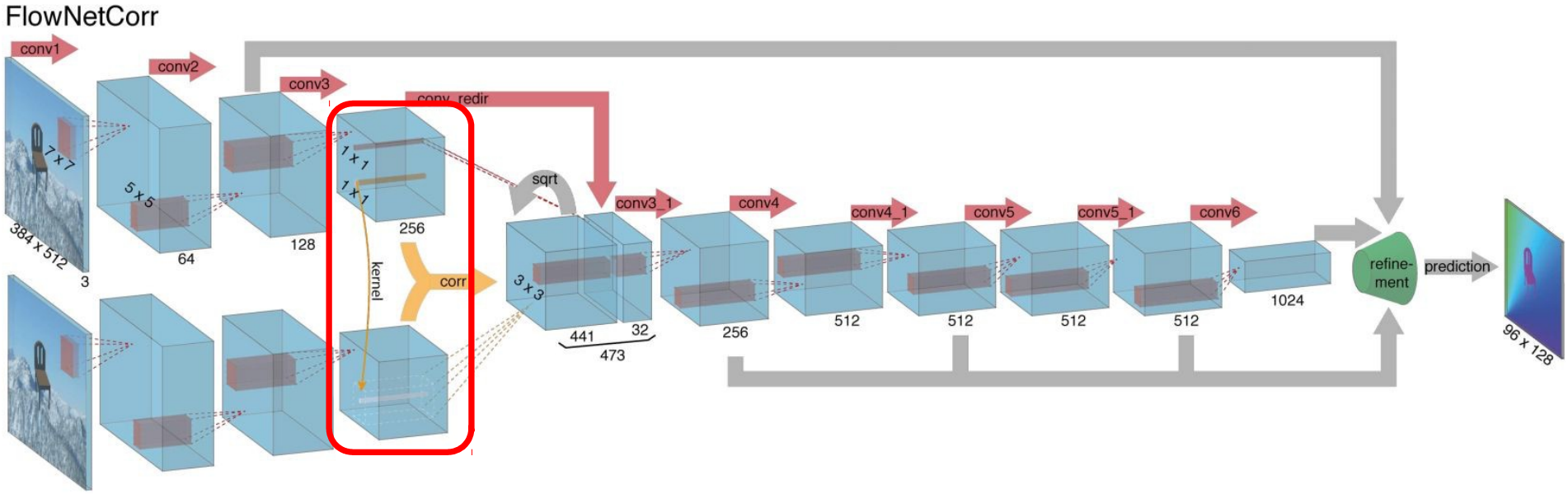
- Supervised OpticalFlow Deep CNN:



Dosovitskiy, et al “FlowNet: Learning Optical Flow With Convolutional Networks ”

Optical Flow Learning

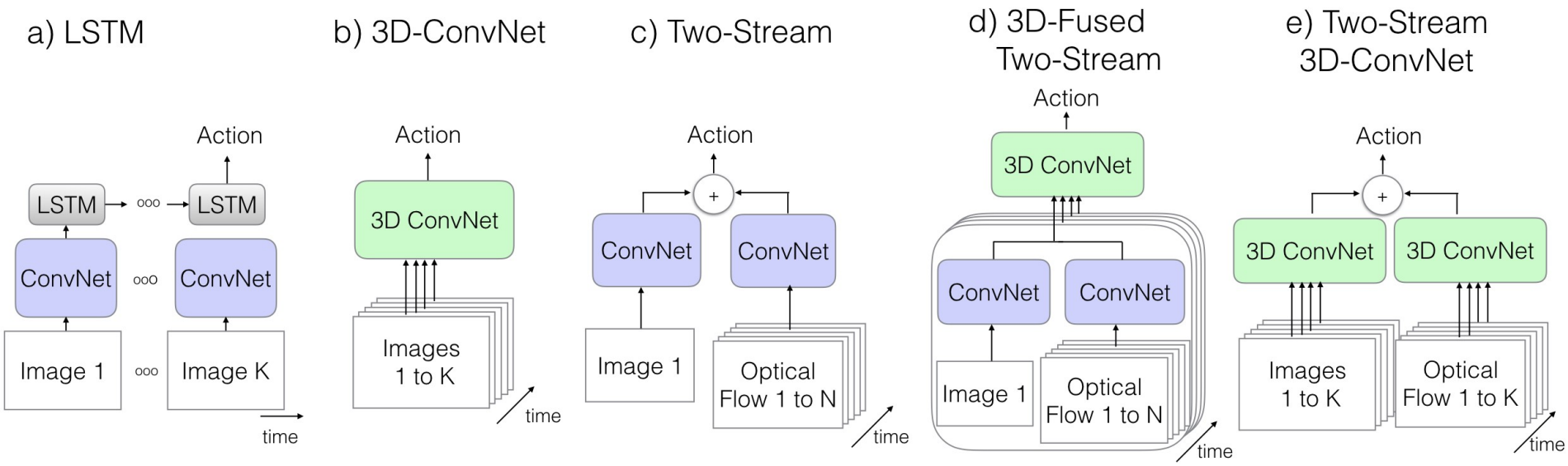
- Supervised OpticalFlow Deep CNN:



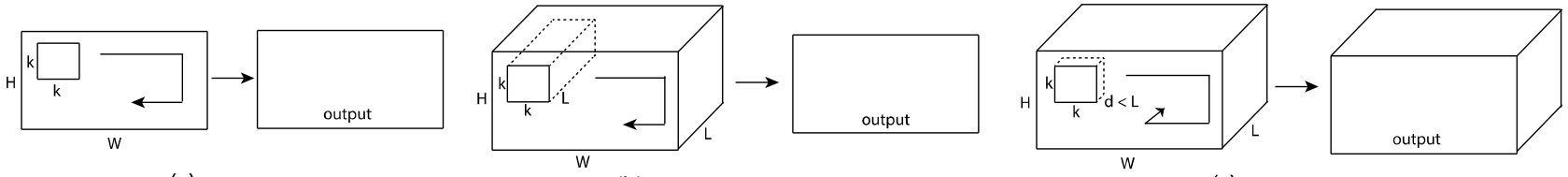
Correlation Convolutional Layer

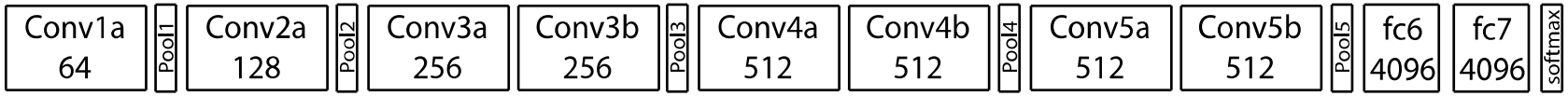
Dosovitskiy, et al “FlowNet: Learning Optical Flow With Convolutional Networks ”

3D Convolutional Networks



C3D

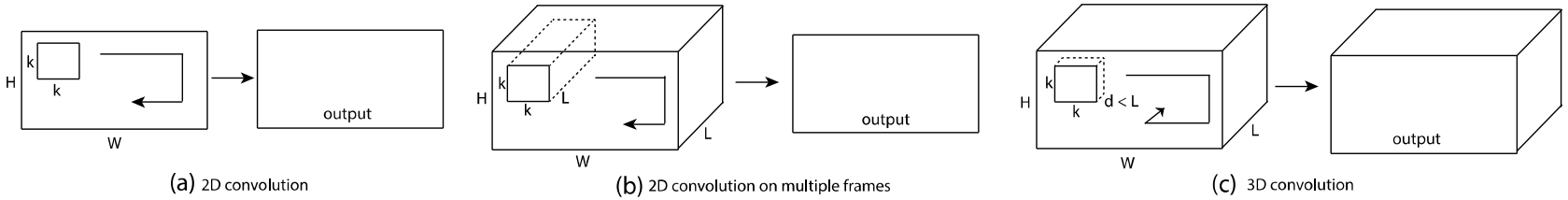


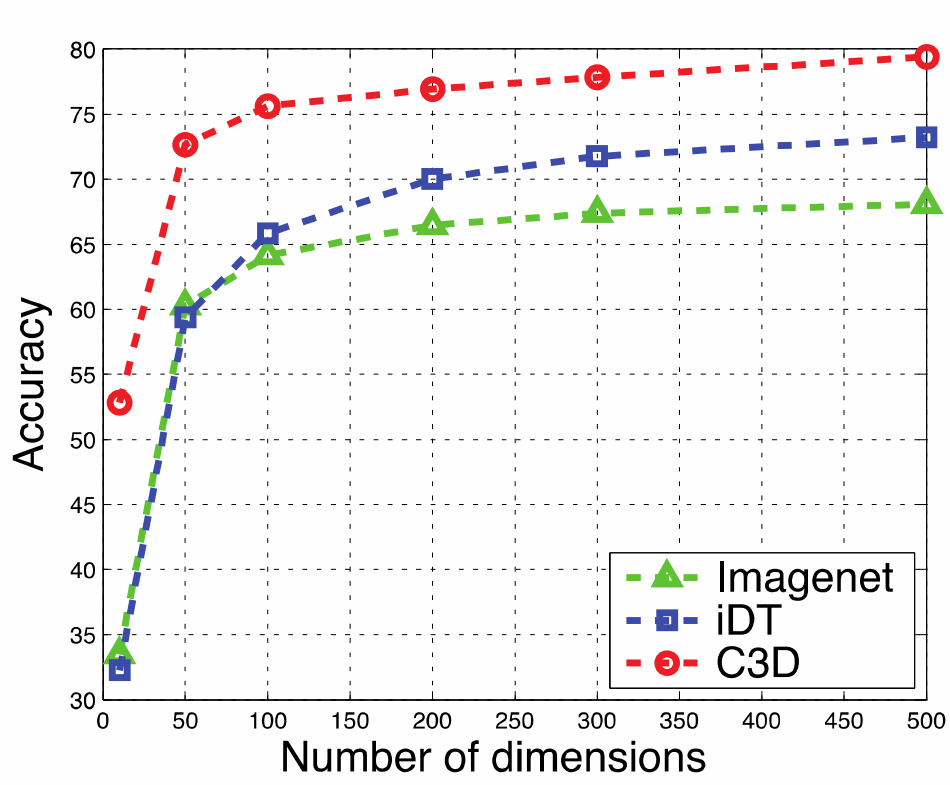


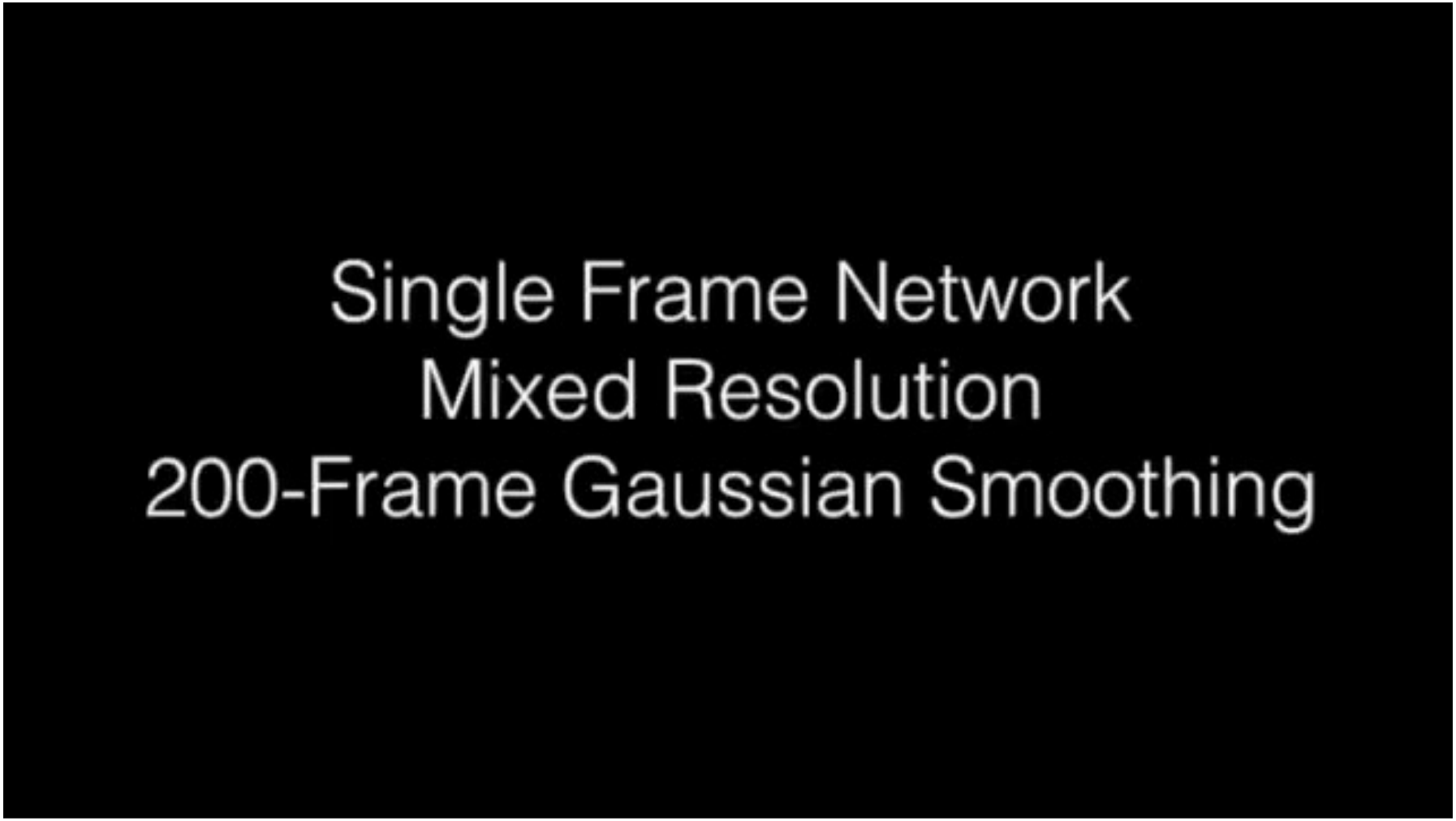
C3D: VGG style 3D Network

Tran, et al “Learning Spatiotemporal Features with 3D Convolutional Networks”

C3D

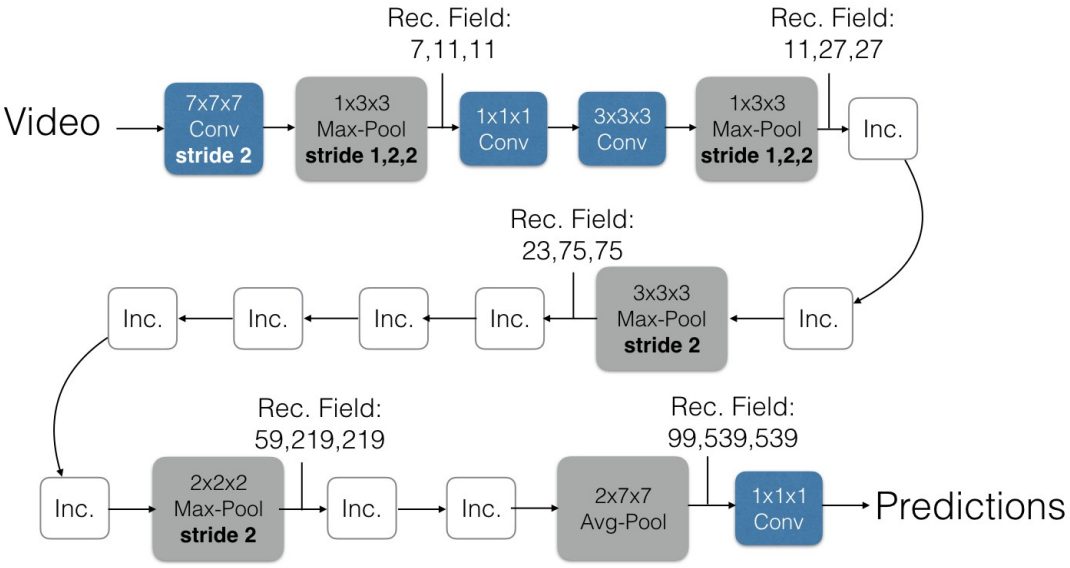

H
k
k
W
output
(a) 2D convolution
H
k
k
L
W
L
output
(b) 2D convolution on multiple frames
H
k
k
d < L
W
L
output
(c) 3D convolution


Accuracy
Number of dimensions
Imagenet
iDT
C3D
80
75
70
65
60
55
50
45
40
35
30
0
50
100
150
200
250
300
350
400
450
500

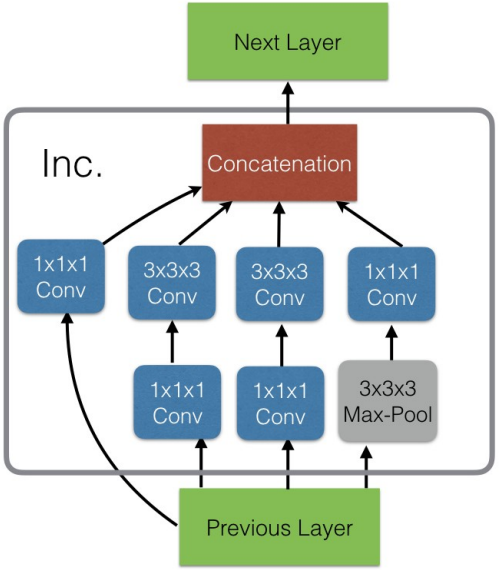

Single Frame Network
Mixed Resolution
200-Frame Gaussian Smoothing

Inflated 3D ConvNet (I3D)

Inflated Inception-V1



Inception Module (Inc.)



Carreira, et al “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset”

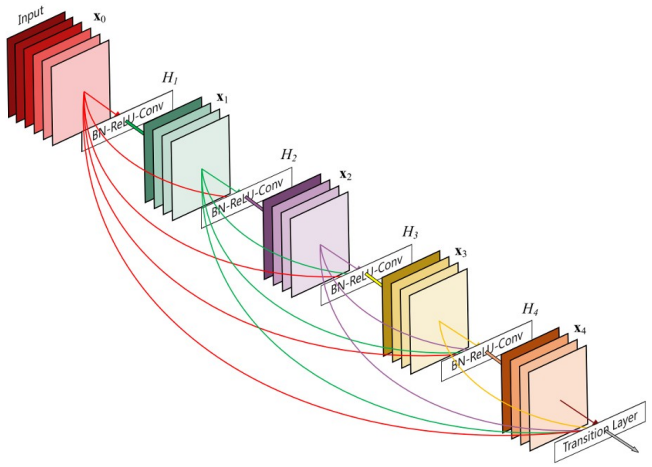
Inflated 3D ConvNet (I3D)

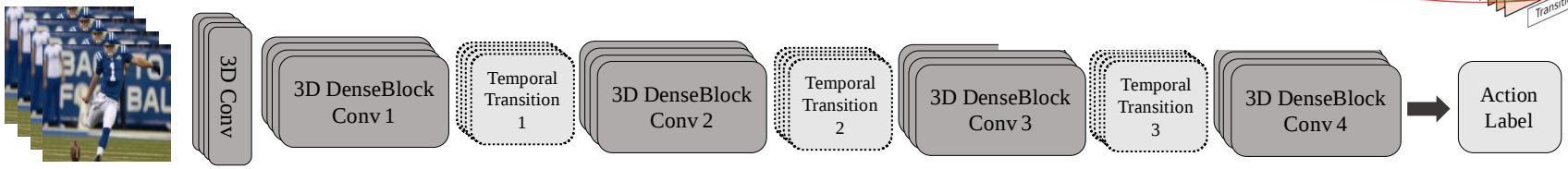
Method	#Params	Training		Testing	
		#Input Frames	Temporal Footprint	#Input Frames	Temporal Footprint
ConvNet+LSTM	9M	25 rgb	5s	50 rgb	10s
3D-ConvNet	79M	16 rgb	0.64s	240 rgb	9.6s
Two-Stream	12M	1 rgb, 10 flow	0.4s	25 rgb, 250 flow	10s
3D-Fused	39M	5 rgb, 50 flow	2s	25 rgb, 250 flow	10s
Two-Stream I3D	25M	64 rgb, 64 flow	2.56s	250 rgb, 250 flow	10s

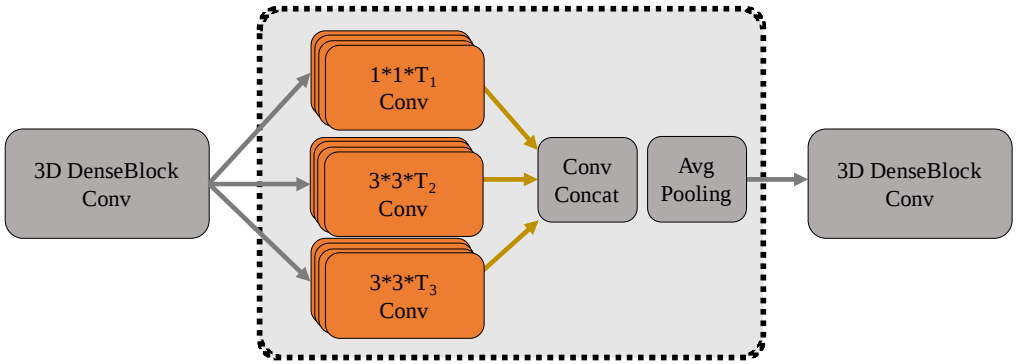
Architecture	UCF-101			HMDB-51			miniKinetics		
	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow
(a) LSTM	81.0	–	–	36.0	–	–	69.9	–	–
(b) 3D-ConvNet	51.6	–	–	24.3	–	–	60.0	–	–
(c) Two-Stream	83.6	85.6	91.2	43.2	56.3	58.3	70.1	58.4	72.9
(d) 3D-Fused	83.2	85.8	89.3	49.2	55.5	56.8	71.4	61.0	74.0
(e) Two-Stream I3D	**84.5**	**90.6**	**93.4**	**49.8**	**61.9**	**66.4**	**74.1**	**69.6**	**78.7**

Carreira, et al “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset”

Temporal 3D ConvNets



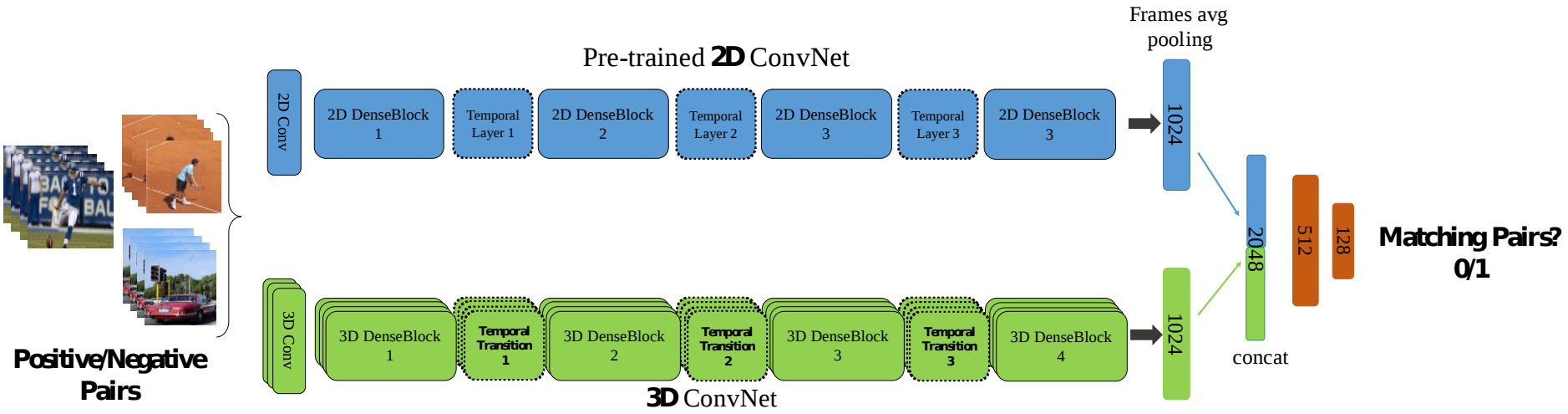




3D Temporal Transition Layer

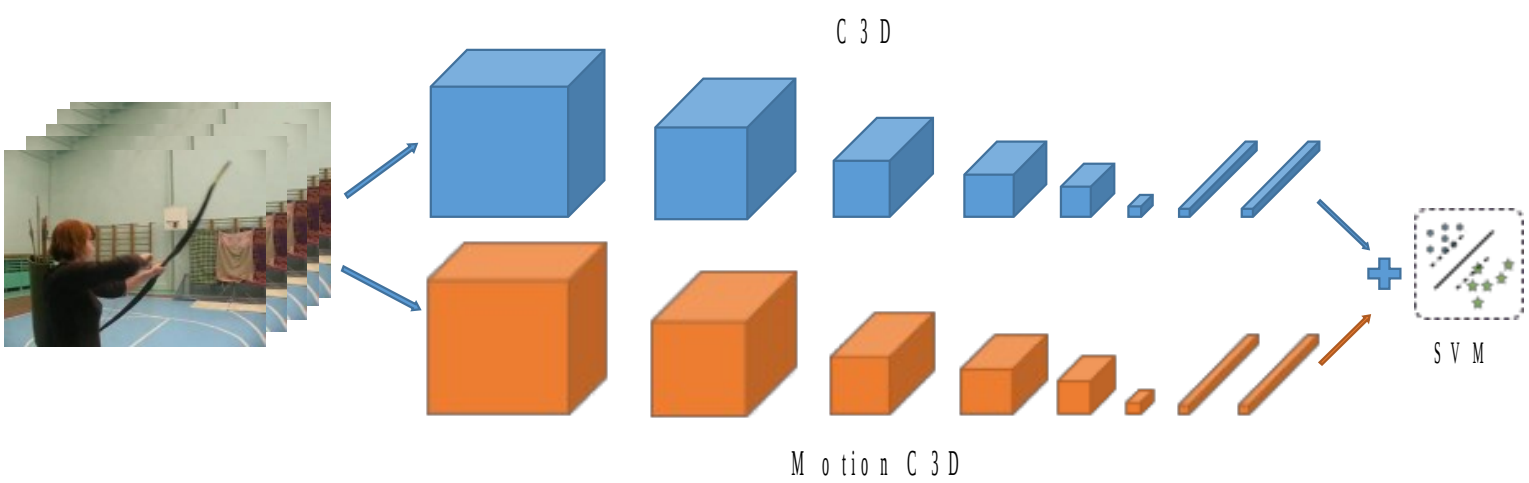
Diba, et al “Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification”

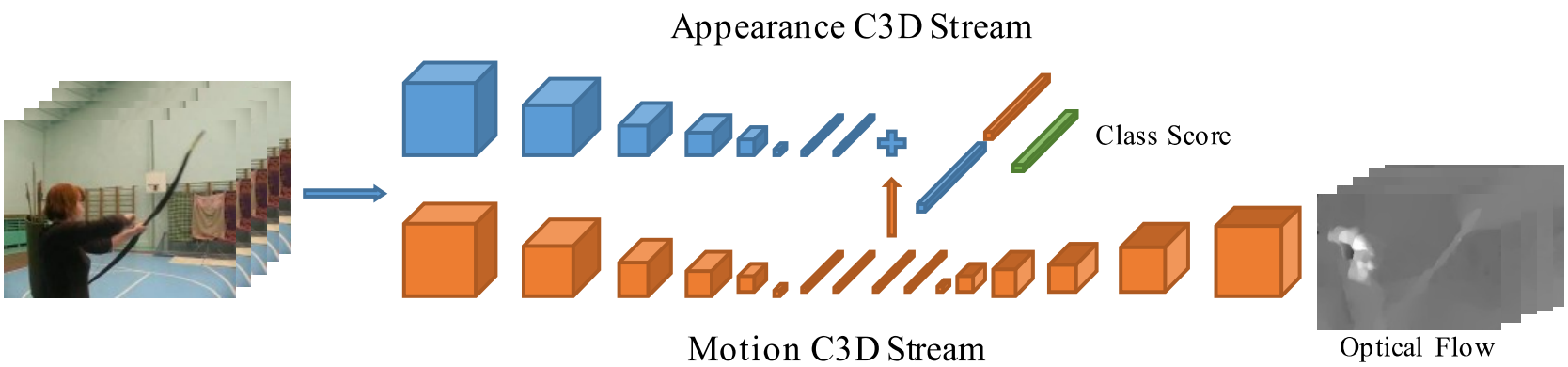
Transfer Learning from 2D to 3D



Diba, et al “Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification”

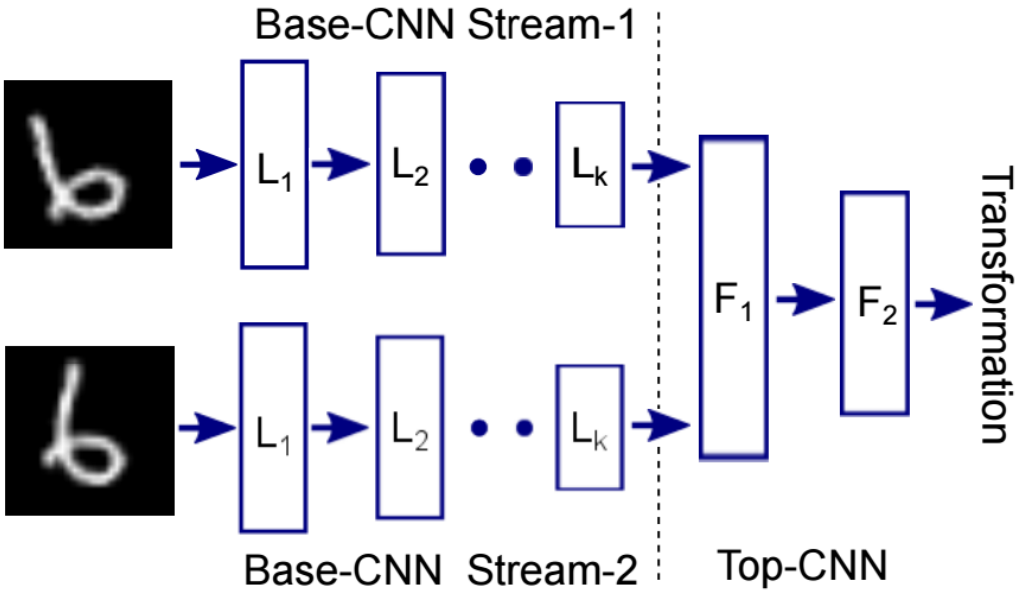
Action recognition with two stream 3D CNN

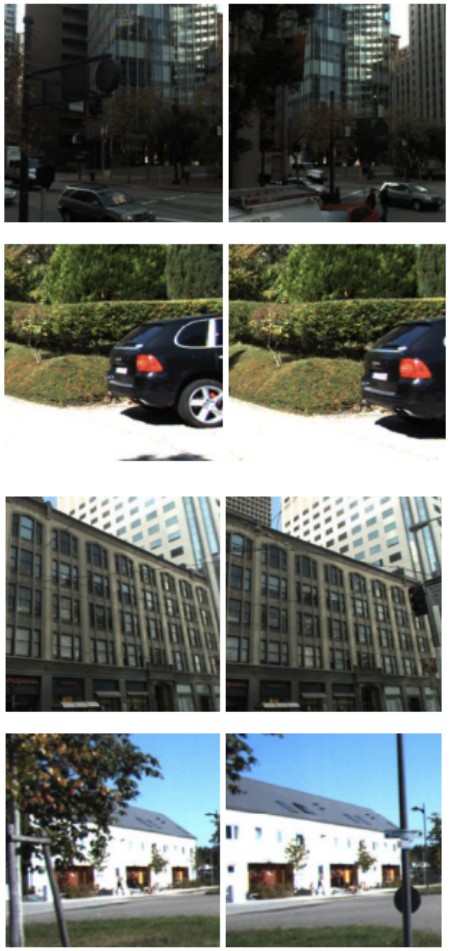




Ali Diba, et al. “Efficient Two-Stream Motion and Appearance 3D CNNs for Video

Unsupervised visual learning from motion



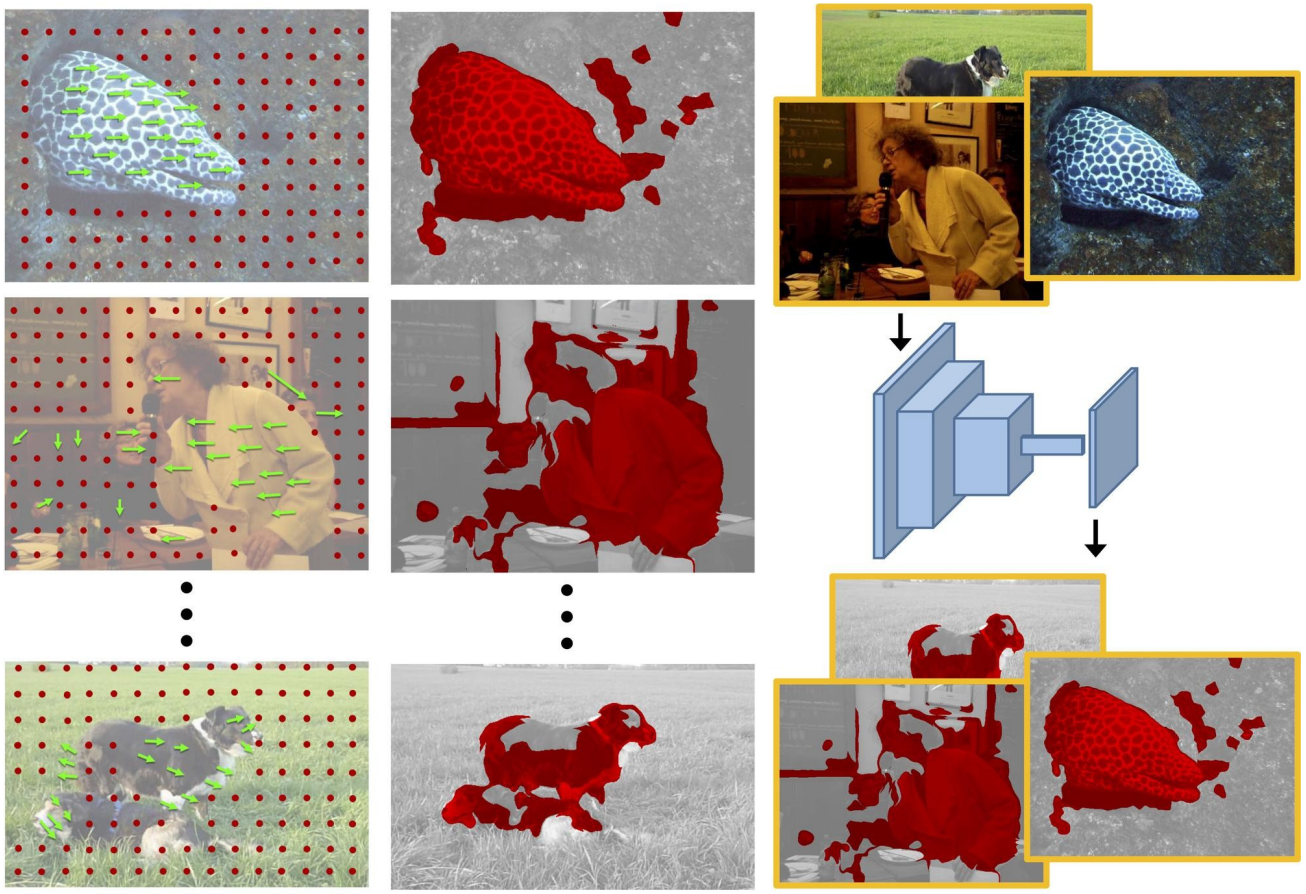


$$L(x_{t_1}, x_{t_2}, W) =$$

$$\begin{cases} D(x_{t_1}, x_{t_2}) & \text{if } |t_1 - t_2| \leq T \\ 1 - max\big(0, m - D(x_{t_1}, x_{t_2})\big) & \text{if } |t_1 - t_2| > T \end{cases}$$

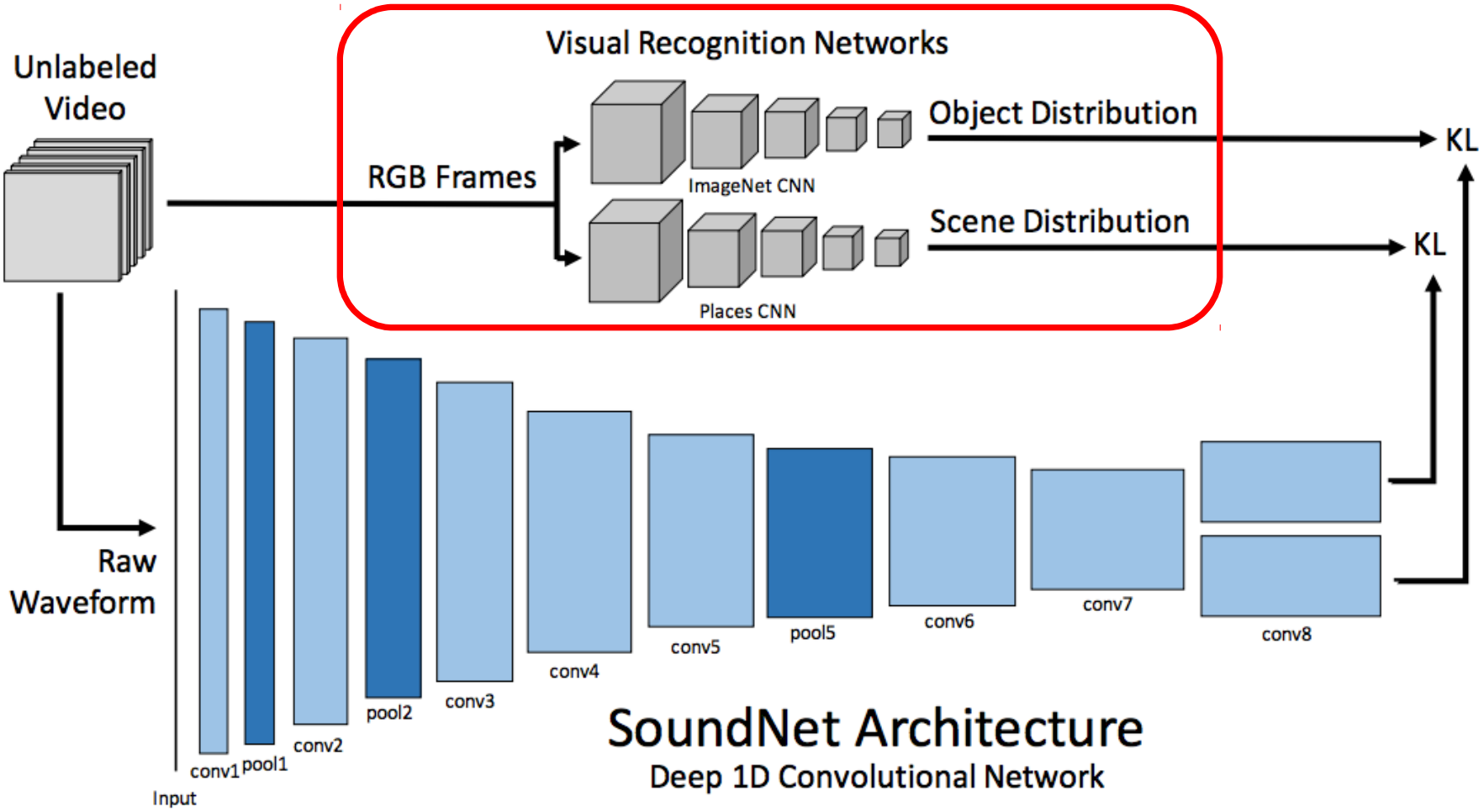
Pulkit Agrawal, et al. “Learning to See by Moving”

Learning Features by Watching Objects Move



1. Collect videos 2. Segment using motion 3. Train ConvNet

Sound and Vision



Yusuf Aytar, et al. “SoundNet: Learning Sound Representations from Unl

Sound and Vision

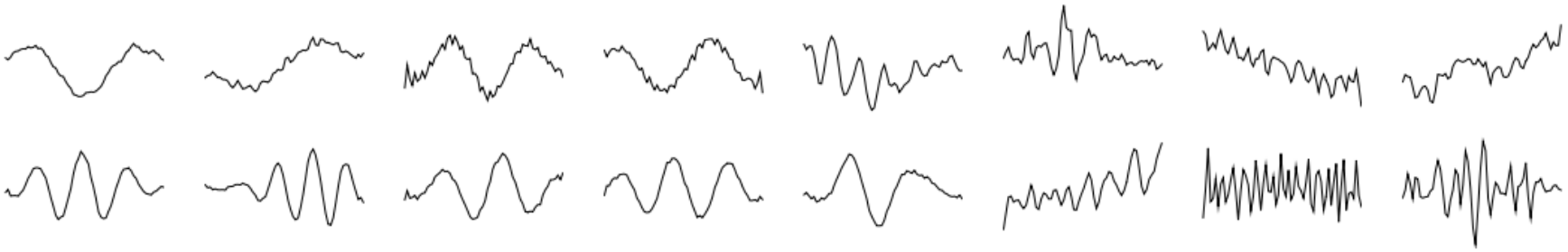


Figure 5: **Learned filters in conv1:** We visualize the filters for raw audio in the first layer of the deep convolutional network.

Baby Talk | Bubbles | Cheering | Bird Chirps

Figure 6: **What emerges in sound hidden units?** We visualize some of the hidden units in the last hidden layer of our sound representation by finding inputs that maximally activate a hidden unit. Above, we illustrate what these units capture by showing the corresponding video frames. No vision is used in this experiment; we only show frames for visualization purposes only.

Sound and Vision

Method	Accuracy
RG [29]	69%
LTT [21]	72%
RNH [30]	77%
Ensemble [34]	78%
SoundNet	**88%**

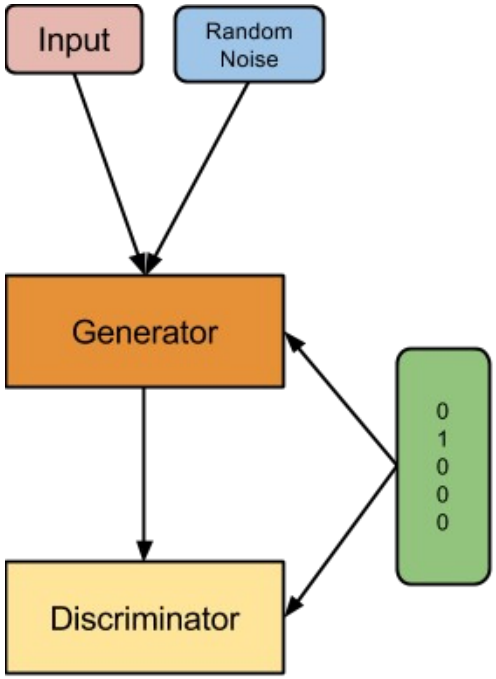
Table 3: **Acoustic Scene Classification on DCASE:** We evaluate classification accuracy on the DCASE dataset. By leveraging large amounts of unlabeled video, SoundNet generally outperforms hand-crafted features by 10%.

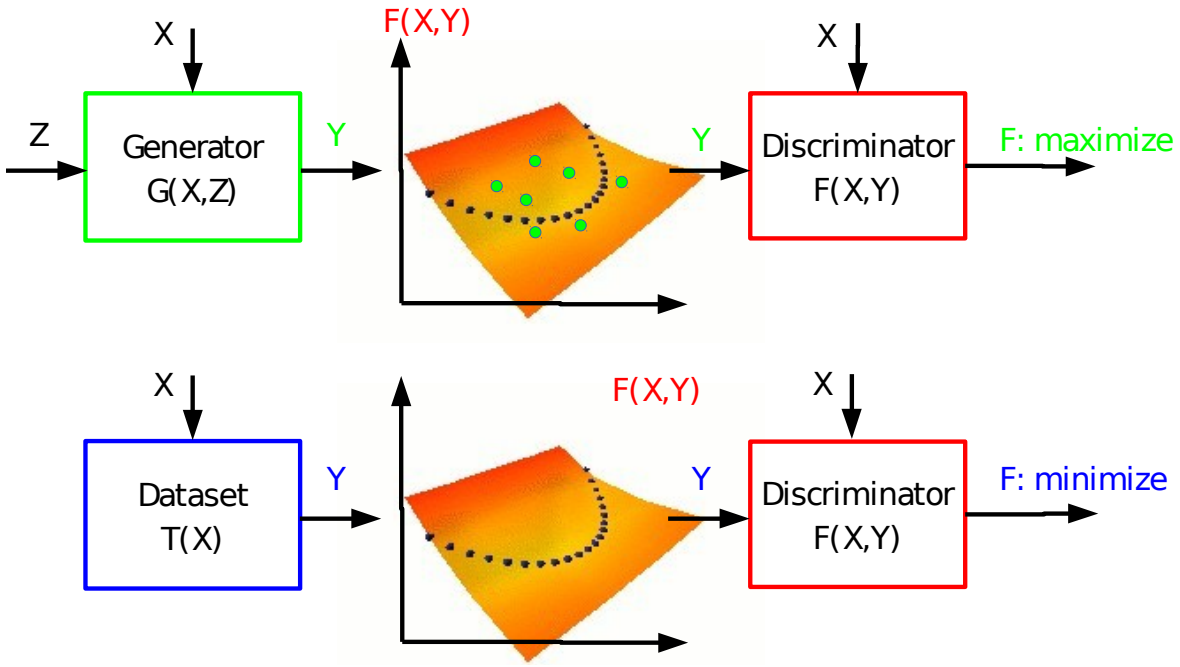
	Accuracy on	
Method	ESC-50	ESC-10
SVM-MFCC [28]	39.6%	67.5%
Convolutional Autoencoder	39.9%	74.3%
Random Forest [28]	44.3%	72.7%
Piczak ConvNet [27]	64.5%	81.0%
SoundNet	**74.2%**	**92.2%**
Human Performance [28]	81.3%	95.7%

Table 4: **Acoustic Scene Classification on ESC-50 and ESC-10:** We evaluate classification accuracy on the ESC datasets. Results suggest that deep convolutional sound networks trained with visual supervision on unlabeled data outperforms baselines.

Adversarial Training

GANS: Images and Unsupervised Learning





Ian Goodfellow, et al. “Generative Adversarial Ne

GANS: Making Videos

Generating Videos with Scene Dynamics

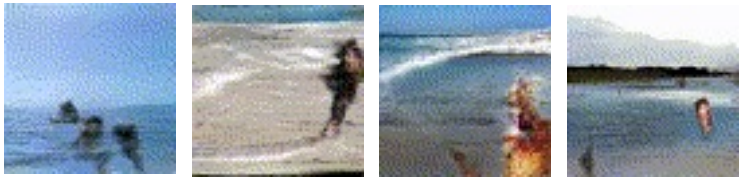
Carl Vondrick
MIT
vondrick@mit.edu

Hamed Pirsiavash
University of Maryland Baltimore County
hpirsiav@umbc.edu

Antonio Torralba
MIT
torralba@mit.edu

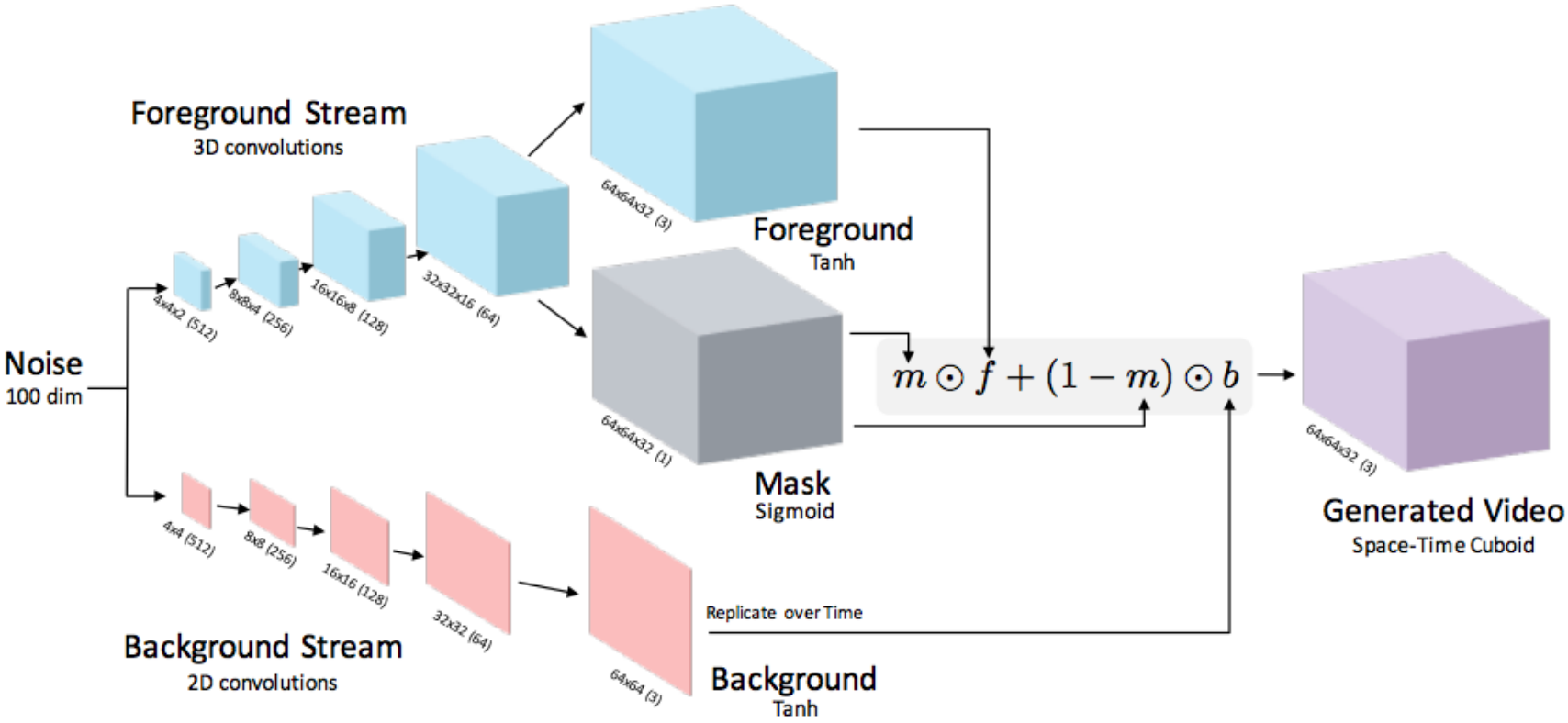
Train

Golf

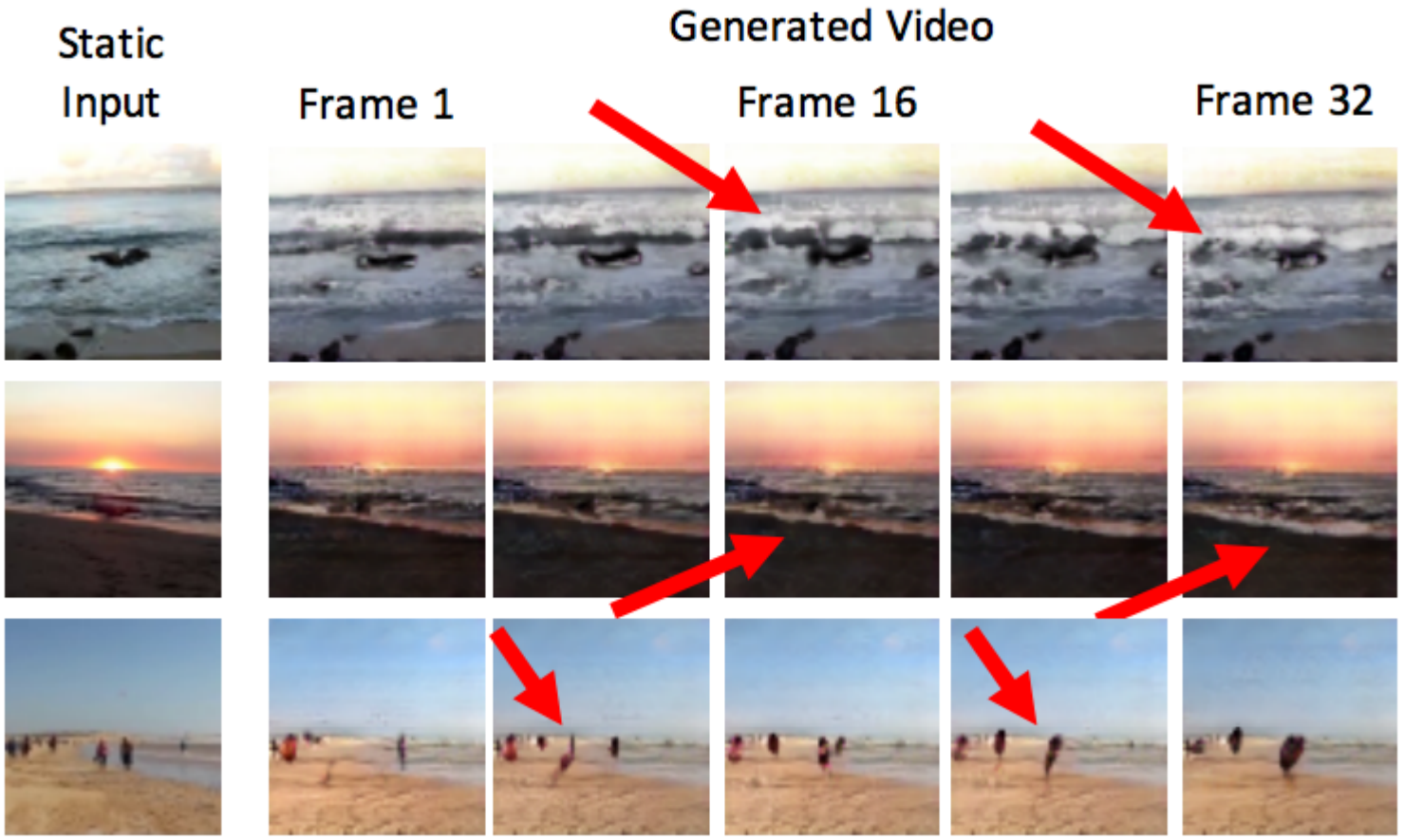


Beach

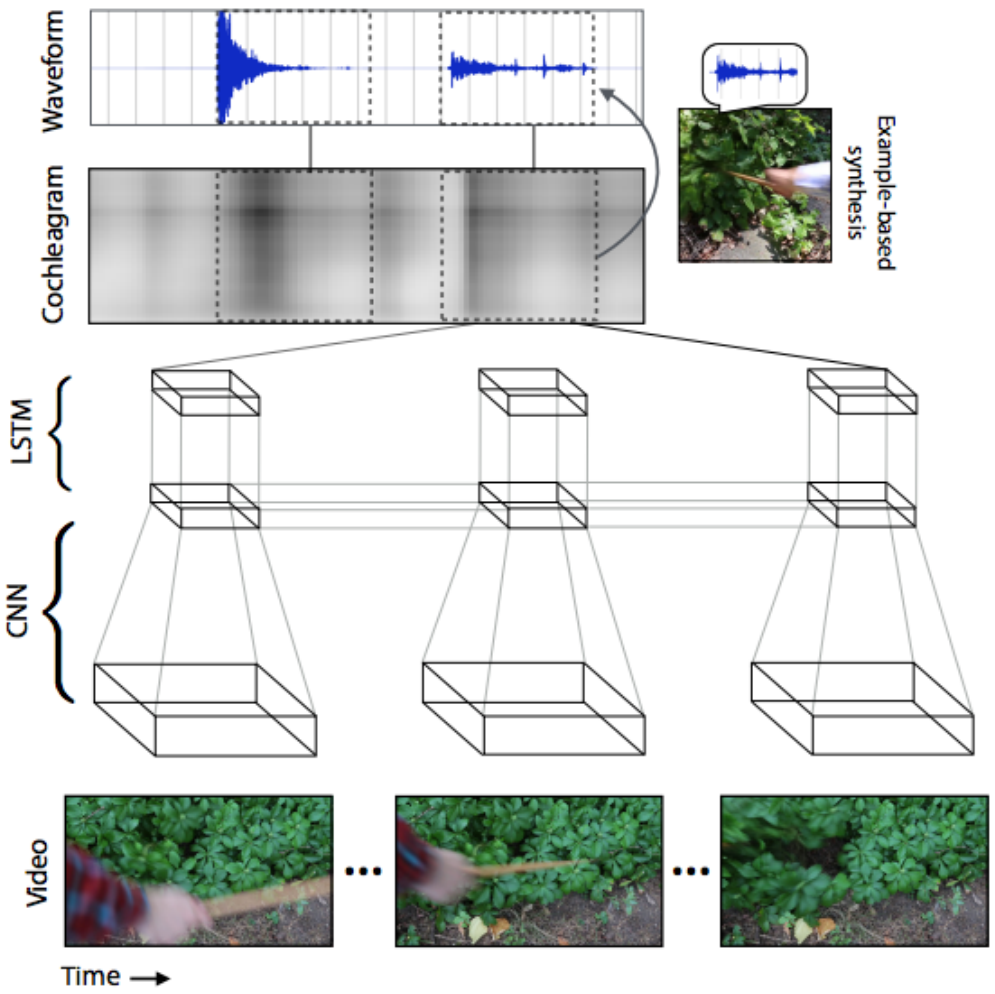
GANS: Making Videos



GANS: Making Videos

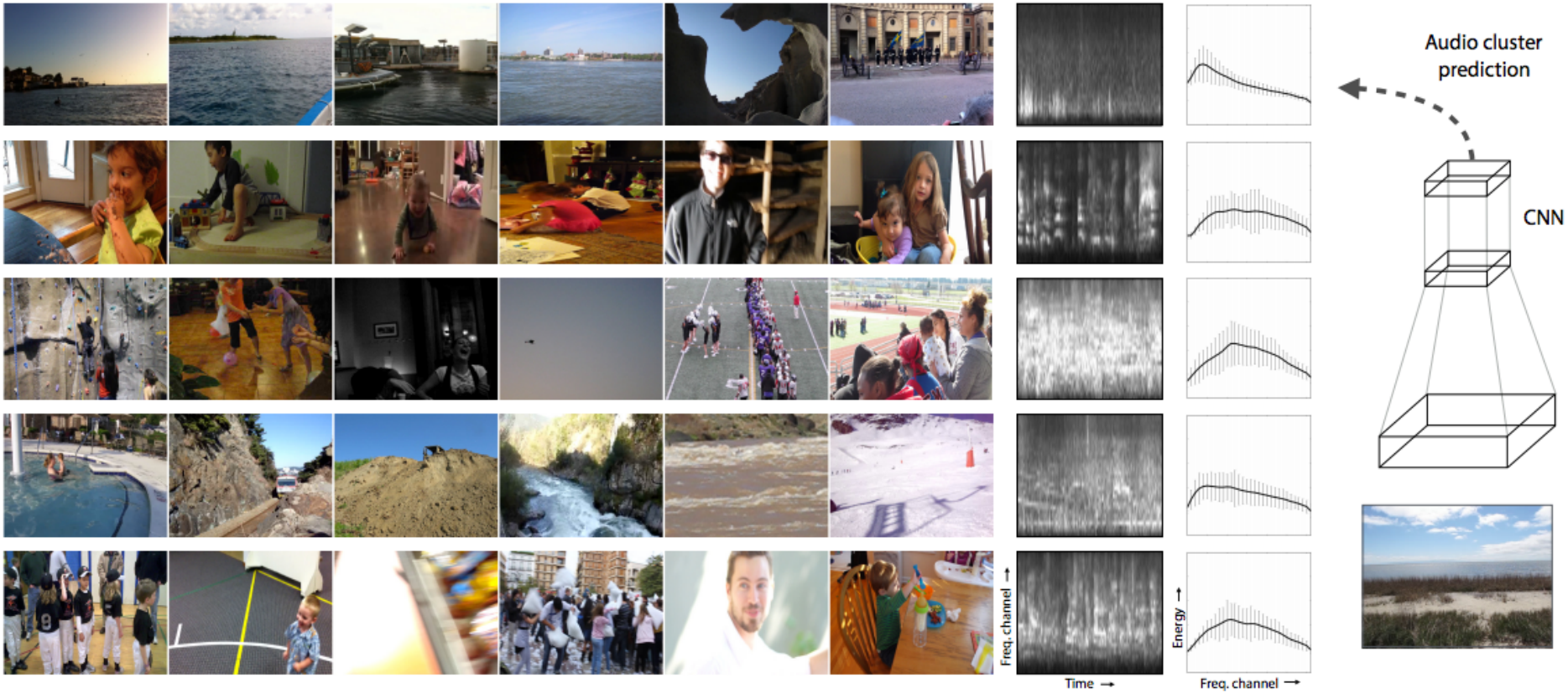


Ambient Sound Supervision



Andrew Owens, et al. “Visually Indicated Sound

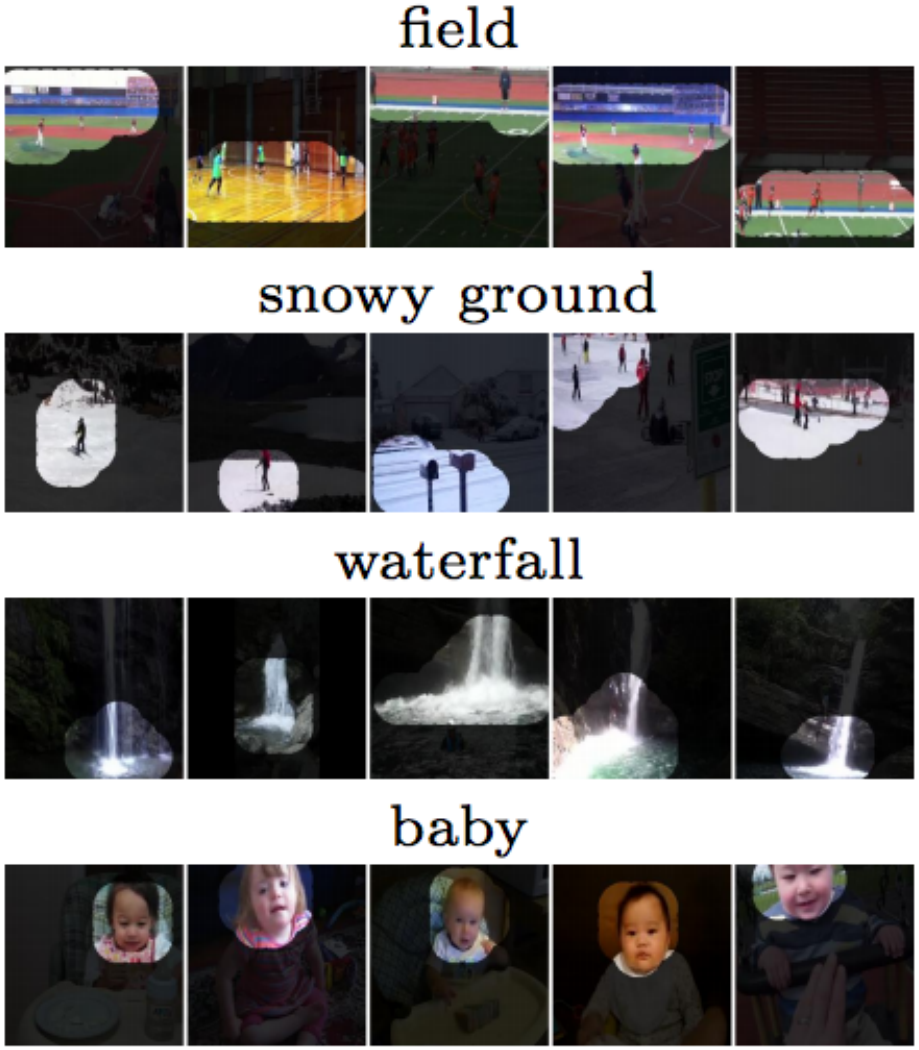
Ambient Sound Supervision



(a) Images grouped by audio cluster (b) Clustered audio stats. (c) CNN model

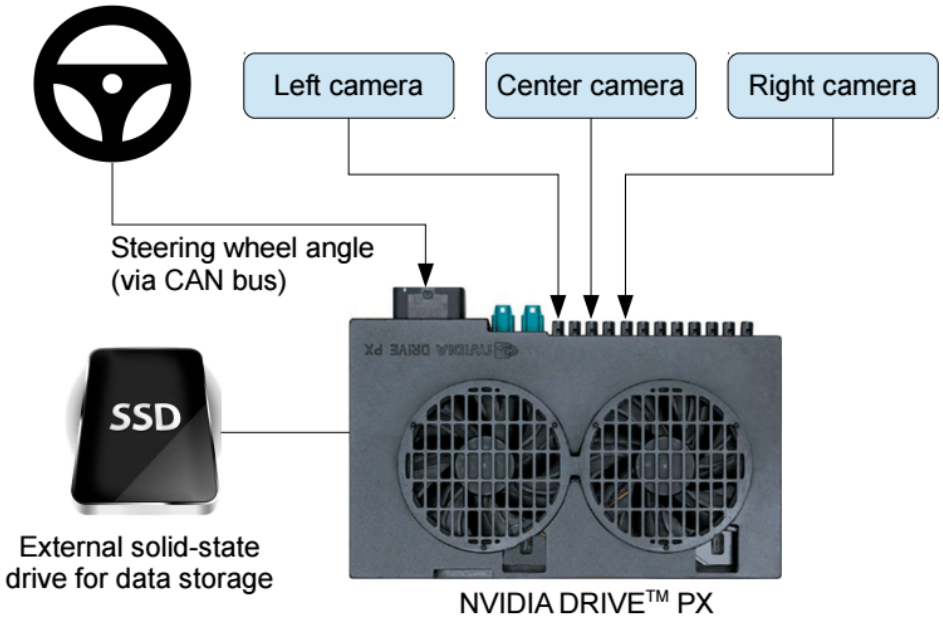
Andrew Owens, et al. “Ambient Sound Provides Supervision for Visua

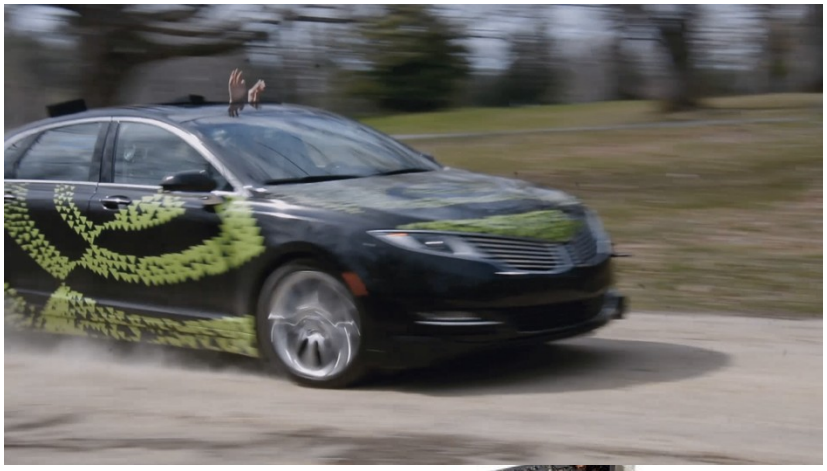
Ambient Sound Supervision

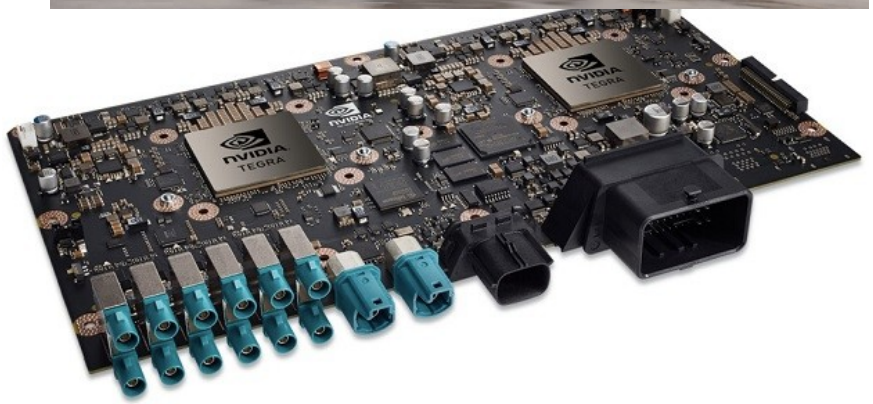


Autonomous Cars

End to End Learning for Self-Driving Cars

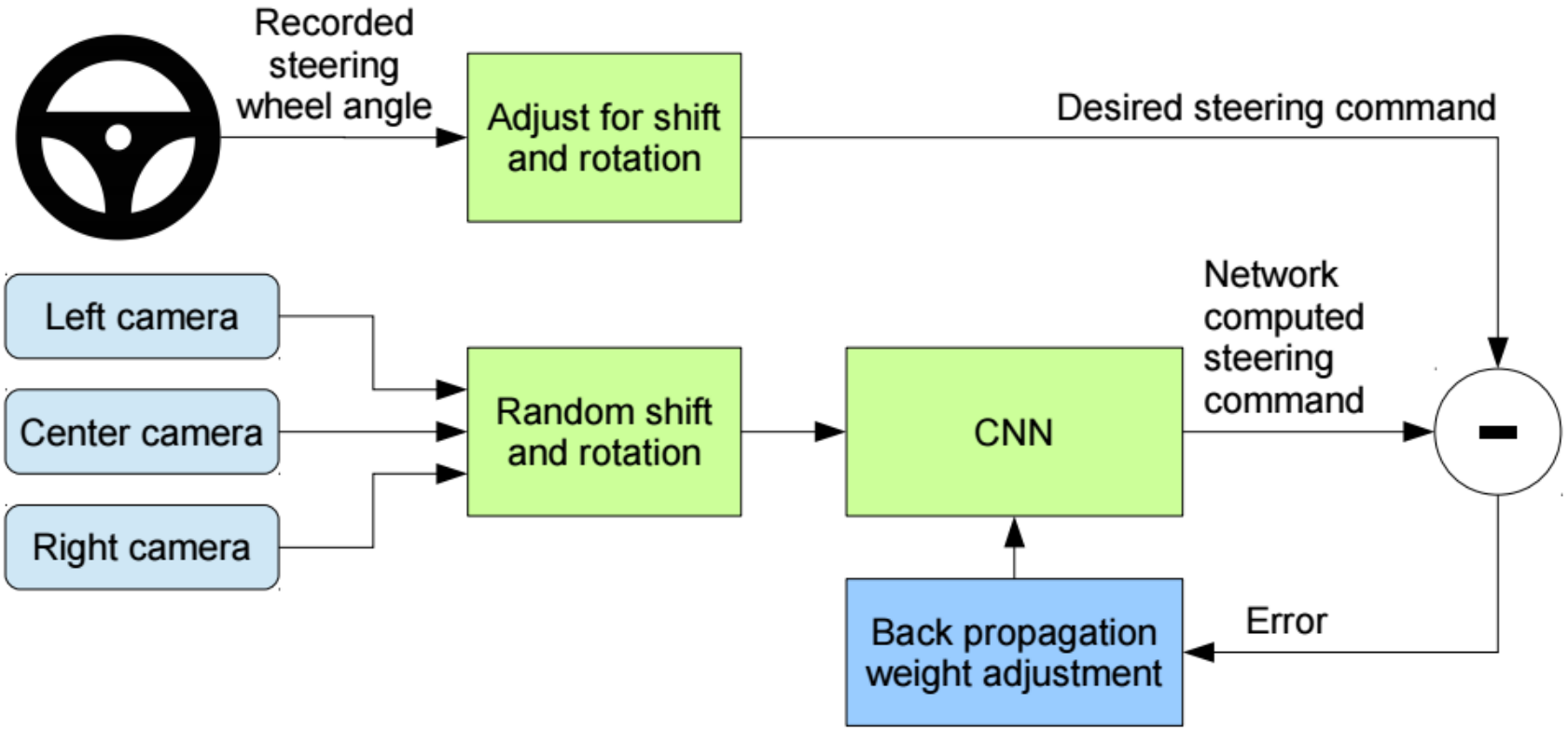


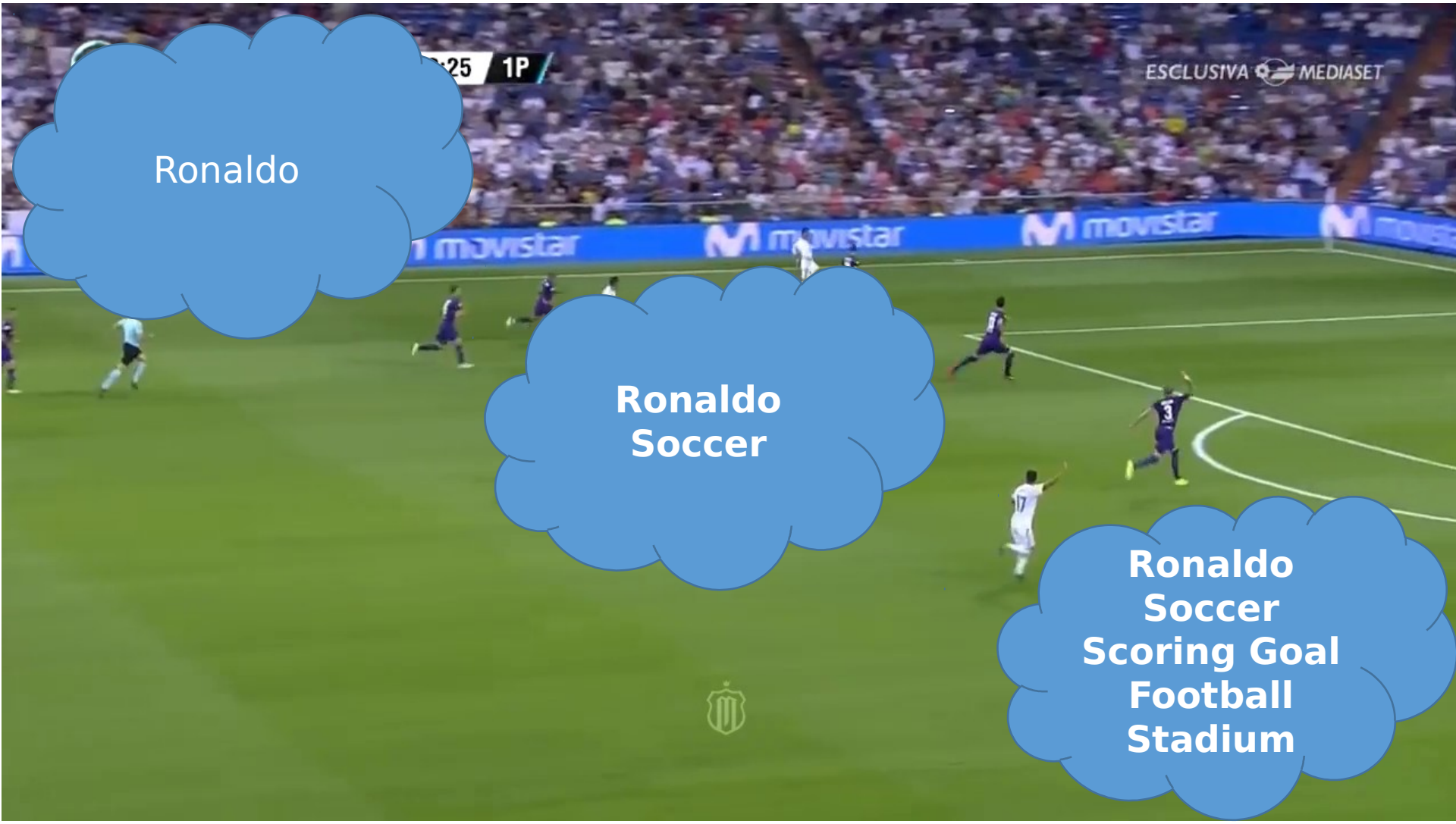


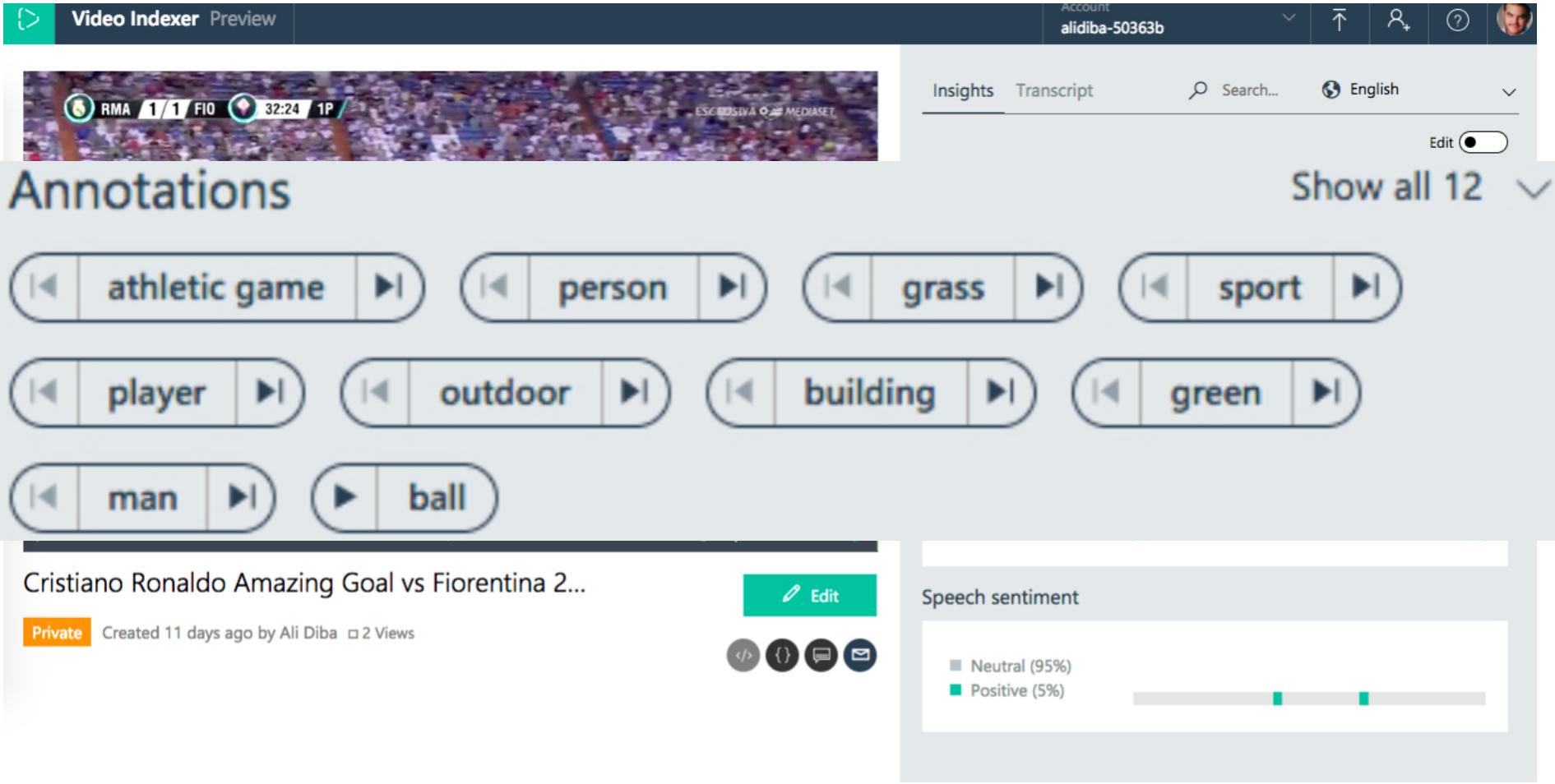


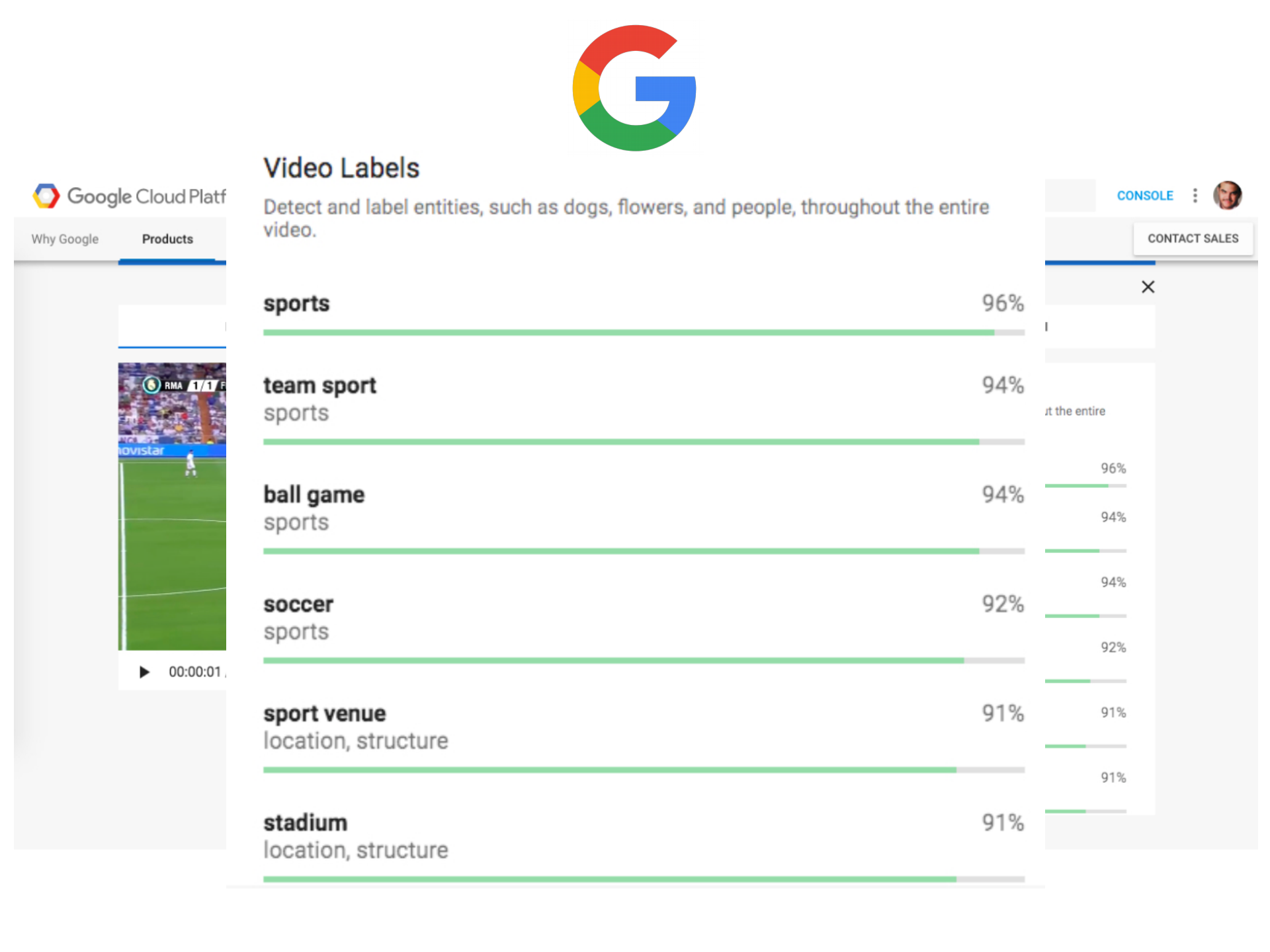
NVIDIA Autonomous car system

Autonomous Cars




Ronaldo
Ronaldo Soccer
Ronaldo Soccer Scoring Goal Football Stadium
ESCLUSIVA MEDIASET
movistar

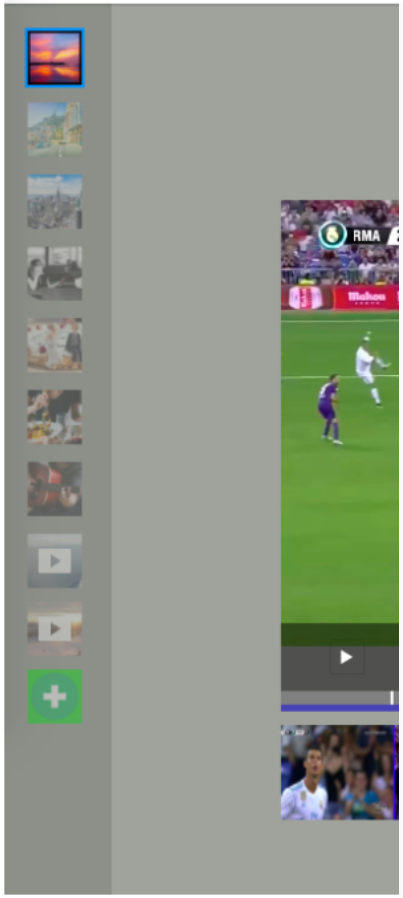

Video Indexer Preview
Account
alidiba-50363b
RMA 1 1 FIO 32:24 1P
Insights
Transcript
Search...
English
Edit
Annotations
Show all 12
athletic game
person
grass
sport
player
outdoor
building
green
man
ball
Cristiano Ronaldo Amazing Goal vs Fiorentina 2...
Edit
Private
Created 11 days ago by Ali Diba
2 Views
Speech sentiment
Neutral (95%)
Positive (5%)


Google Cloud Platf
Why Google
Products
CONSOLE
CONTACT SALES
00:00:01
Video Labels
Detect and label entities, such as dogs, flowers, and people, throughout the entire video.
sports 96%
team sport 94%
sports
ball game 94%
sports
soccer 92%
sports
sport venue 91%
location, structure
stadium 91%
location, structure

Home Contact Sales

sensifai

Logout



PREDICTED CONCEPT	PROBABILITY
Player	1.00
Stadium	0.79
Sports >association football	0.42
Action >shooting goal (soccer)	1.00
Soccer	0.77
Soccer-specific stadium	0.70

NAME	FREQUENCY
Cristiano Ronaldo	7

Thanks!

Diba@sensifai.com