CEG5304 / EE6934 (PT2)

Lecture 10:

Deep Learning Beyond Images and Texts

Video and 3D Data Analysis

© Copyright National University of Singapore. All Rights Reserved.

Admin Issues for Week 10: Exam (40%)

- Same for CEG5304/EE6934:
  - Venue TBD, 2 hours, 4 May 1pm.
  - Computer-based, using Examsoft, bring a charged laptop (that can last for 2 hours).
  - ONE A4 size help sheet allowed (double-sided allowed, only ONE).
  - Examsoft briefing will be conducted on 2nd April 6-630pm, please be presented. **Attendance will be taken for this briefing session**.
- Different for CEG5304/EE6934:
  - CEG5304: 5 questions, code-based, in Python with numpy/PyTorch.
  - EE6934: 2 questions, essay-based.

© Copyright National University of Singapore. All Rights Reserved.

Admin Issues for Week 10: Other Updates

- Slides (in .pptx form) is now uploaded to the Files tab.
- Syllabus has been updated.
- Next guest lecture: Week 13 (Tentative)

© Copyright National University of Singapore. All Rights Reserved.

Video Analysis:

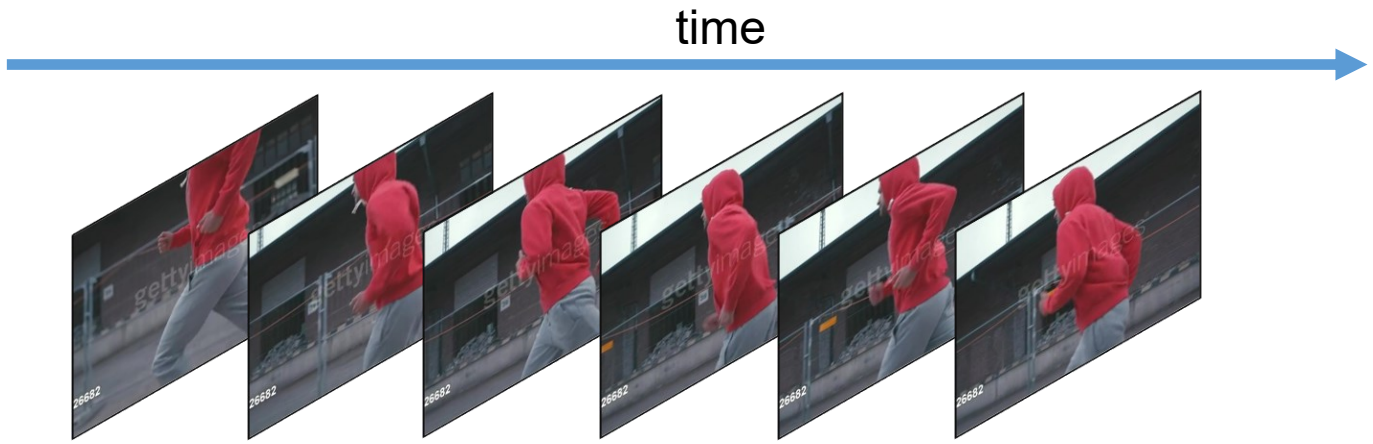
The Basic, Difference, and Development

- ✓ Overall Video Analysis Framework
- ✓ Optical Flow
- ✓ Development of DL in Action Recognition (from 2D-CNN to CLIP-based Models)

© Copyright National University of Singapore. All Rights Reserved.

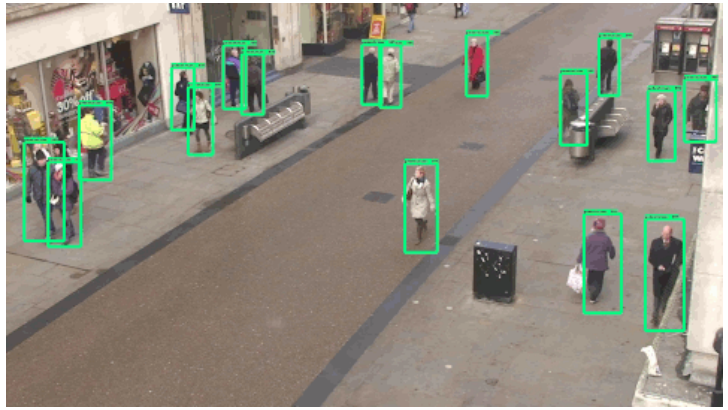
Video Fundamentals: from Images to Videos

- A video can be viewed as a **sequence** of images.
  - By default, the sequence of images are **closely related**.
- A video is more than a sequence of images:
  - Other modalities (types) of information are also included, e.g., audio.



© Copyright National University of Singapore. All Rights Reserved.

Video Analysis Real-world Applications



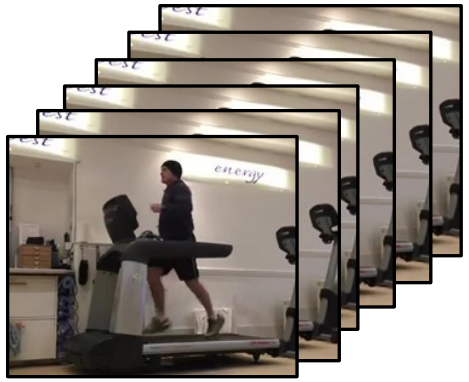
Security Surveillance

Smart Manufacturing

Autonomous Driving

© Copyright National University of Singapore. All Rights Reserved.

Characteristics of Videos



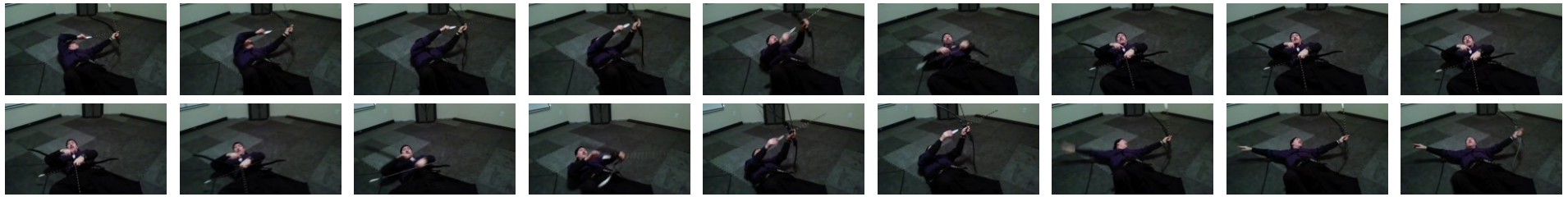
- Contains much more information than a single image:
  - Without considering information of different modalities (e.g., audio and flow-based)
  - Most videos are shot with 30 FPS (frames per second)
  - A 1-minute video usually contains 1800 images. (For 120 FPS videos, a 1-minute video contains 7200 images)
- Usually expressed as a 4D-Tensor (only consider RGB frames):
  - $T \times 3 \times H \times W$ or $3 \times T \times H \times W$ (3 for RGB channels)
  - How to start video analysis given the large amount of information?
  - What is the workflow of video analysis?

© Copyright National University of Singapore. All Rights Reserved.

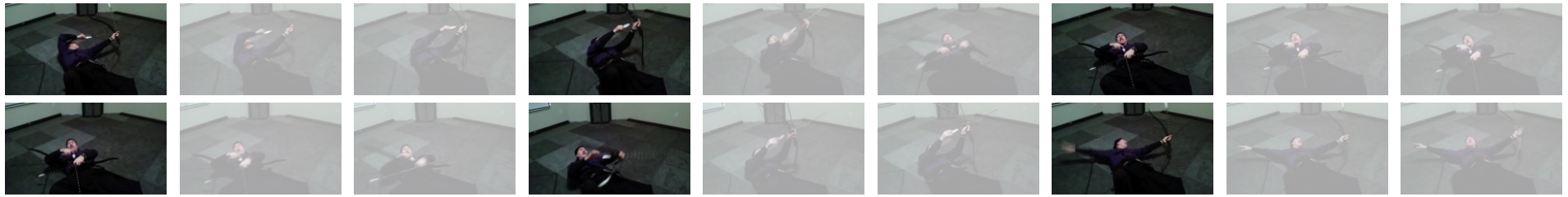
Characteristics of Video Analysis

- Problem of using raw video for video analysis:
  - Videos are too big to process and analyze (for uncompressed video: ~GB per minute).
  - Solution: Frame Sampling (Uniform Sampling; Random Sampling)

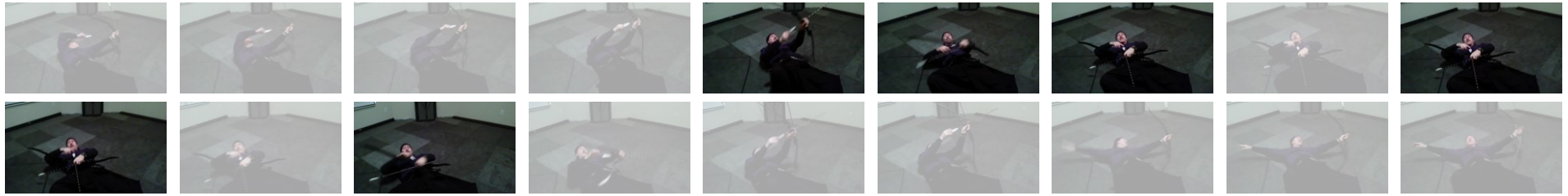
Original Video



Uniform Sampling



Random Sampling

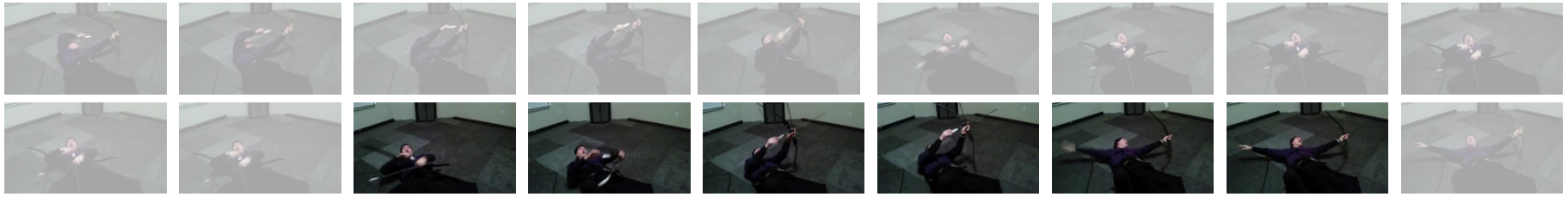


© Copyright National University of Singapore. All Rights Reserved.

Video Sampling

- More recently: learning-based motion-guided sampling (MGSampler) (Mgsampler: An explainable sampling strategy for video action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1513-1522).

MGSampler



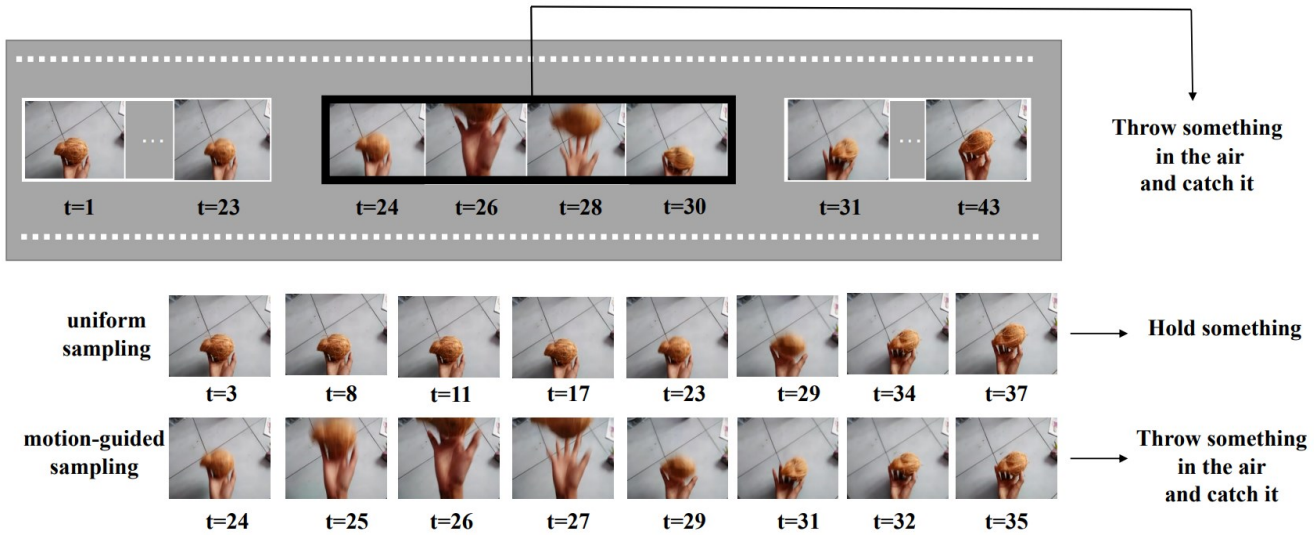
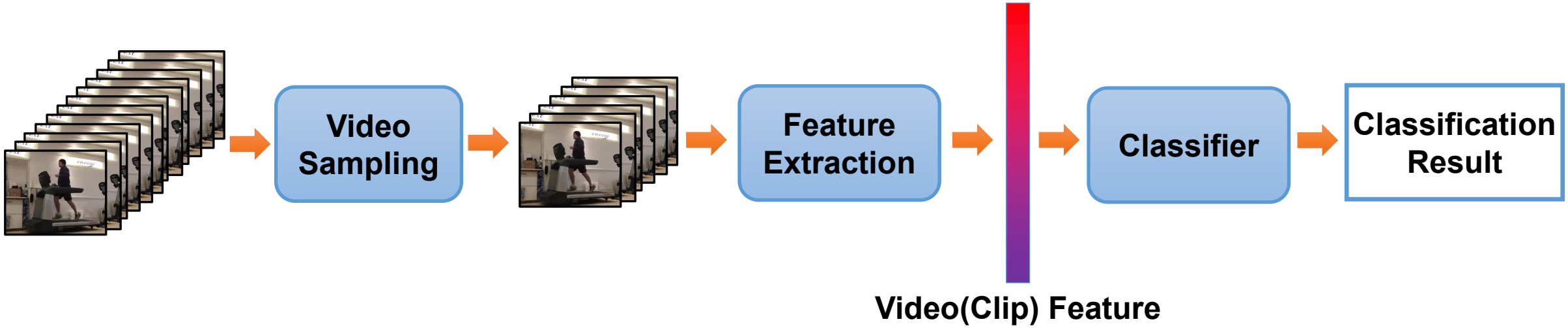


Figure 1. Sample eight frames from a video of throwing something in the air and catching it. Due to the quick moment in action, uniform sampling may miss the key information while our sampling strategy can identify and select frames with large motion magnitude.

© Copyright National University of Singapore. All Rights Reserved.

Outline of Video Analysis

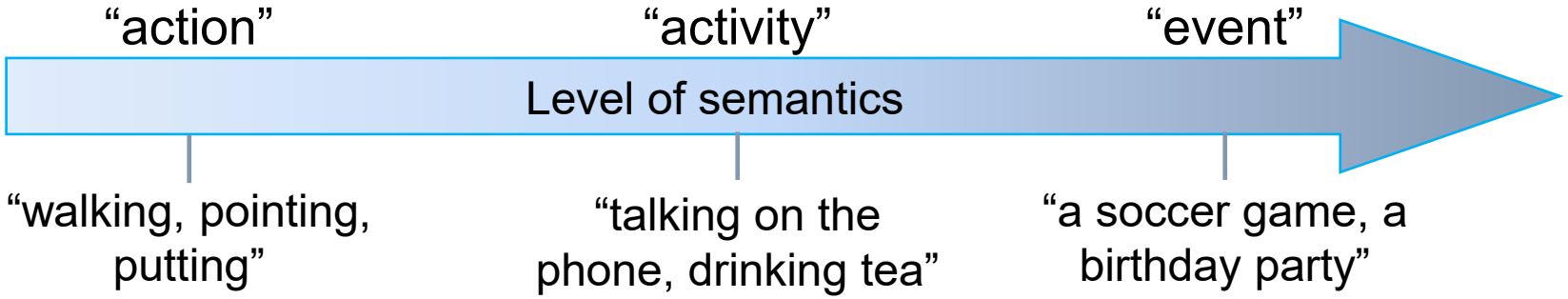


- Sample from the raw video input to form a “clip”.
- Apply feature extraction methods on the “clip”, obtain video (clip) feature.
  - Note: “Clip” and “Video” feature may be used interchangeably in literature.
- Apply a classifier (MAP decision rule/Bayes decision rule; Linear classifier; Nearest-neighbour classifier.
  - Classification is a general task category which includes (for videos): human action recognition (simplest), action segmentation, video semantic segmentation, etc.

© Copyright National University of Singapore. All Rights Reserved.

The Task: Human Action Recognition

- Given an input video, perform processing, and classify the human action in the video.



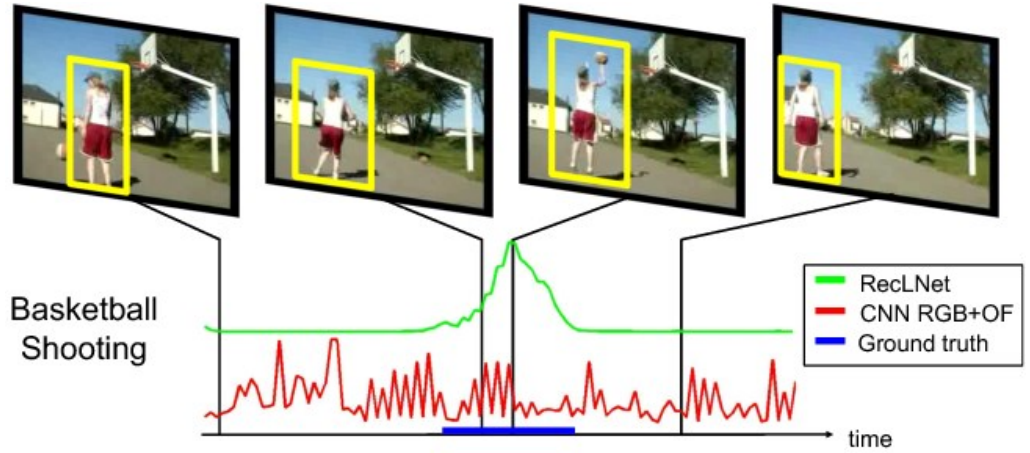
- Short clips, single human action, video captured by ordinary camera.

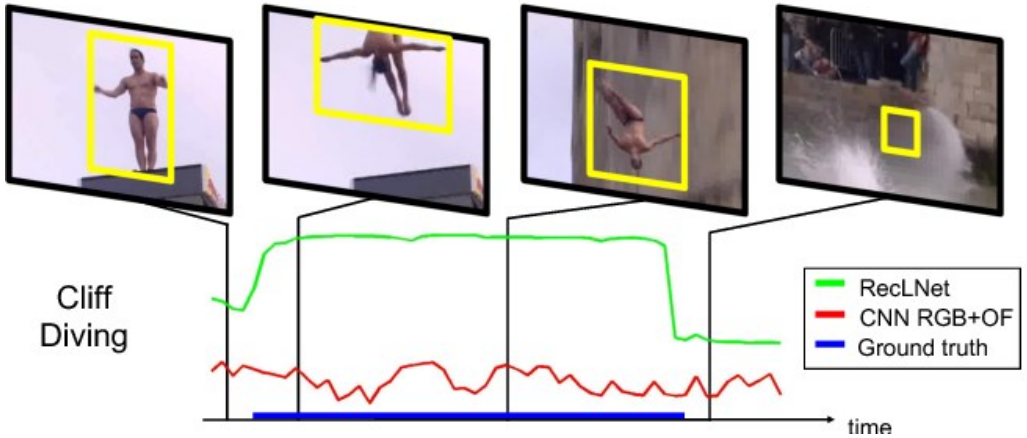
```
Norm of logits: 138.468643

Top classes and probabilities
1.0 41.8137 playing cricket
1.49716e-09 21.494 hurling (sport)
3.84312e-10 20.1341 catching or throwing baseball
1.54923e-10 19.2256 catching or throwing softball
1.13602e-10 18.9154 hitting baseball
8.80112e-11 18.6601 playing tennis
2.44157e-11 17.3779 playing kickball
1.15319e-11 16.6278 playing squash or racquetball
6.13194e-12 15.9962 shooting goal (soccer)
4.39177e-12 15.6624 hammer throw
2.21341e-12 14.9772 golf putting
1.63072e-12 14.6717 throwing discus
```

© Copyright National University of Singapore. All Rights Reserved.

Other Video-Based Tasks

Video Localization (Spatial/Temporal).

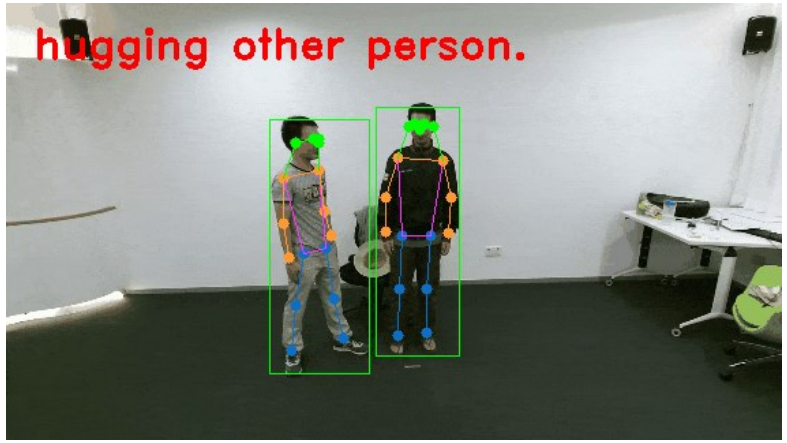




(Dense) Video Captioning.

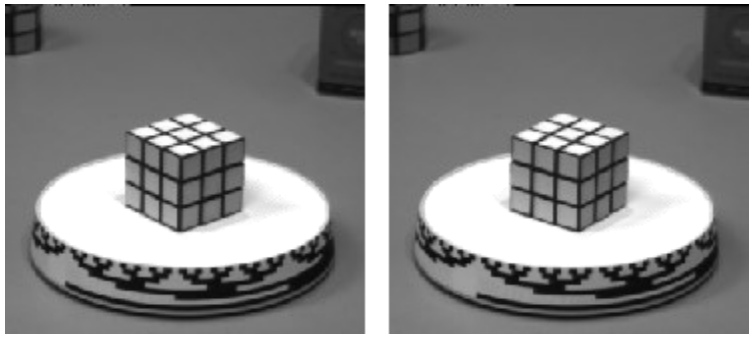
Dense video captioning	A boy is riding a bicycle. He looses his balance and falls on the ground. He gets back up and starts riding again.
Single Sentence captioning	A boy in red t-shirt is riding a bicycle.

Skeleton Action Recognition.

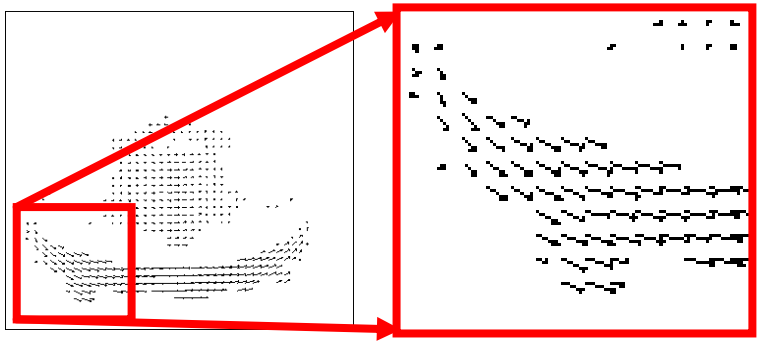


© Copyright National University of Singapore. All Rights Reserved.

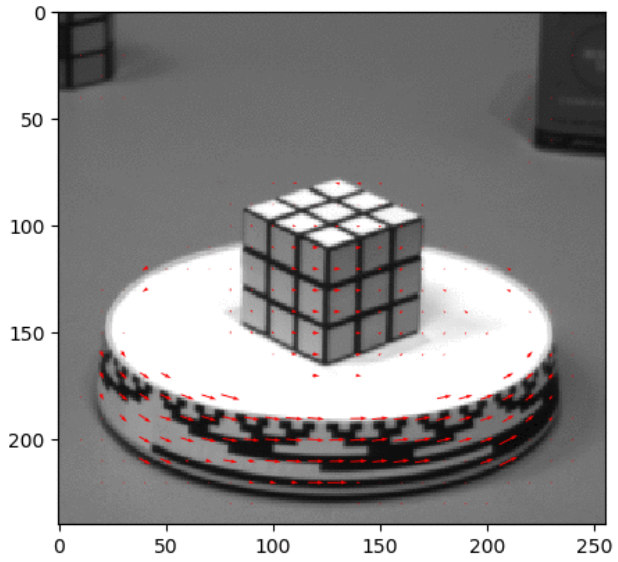
The Unique: Motion Modality



- Motion: assume how the **appearances** of frames **change** indicate how the object actually moves.
  - Obtain explicitly and implicitly.
  - Implicitly: learning-based, “temporal modality”, any method that involves computation from sequential frames.
  - Explicitly: math-based. Typical example: Optical Flow.



- Optical flow: the **apparent motion** of individual pixels caused by the movement of object or camera.
  - **2D vector field** where each vector is a displacement vector showing the movement of points across consecutive frames.
  - Apparent motion is the **appearance** of real motion from a sequence of **still images**.
  - Ideally, optical flow = or at least ≈ the actual motion field.

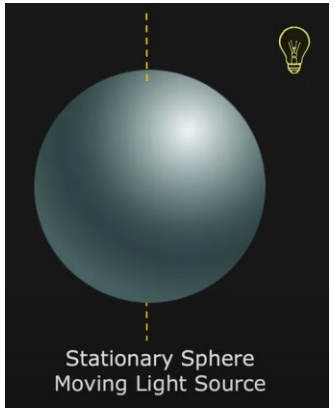


© Copyright National University of Singapore. All Rights Reserved.

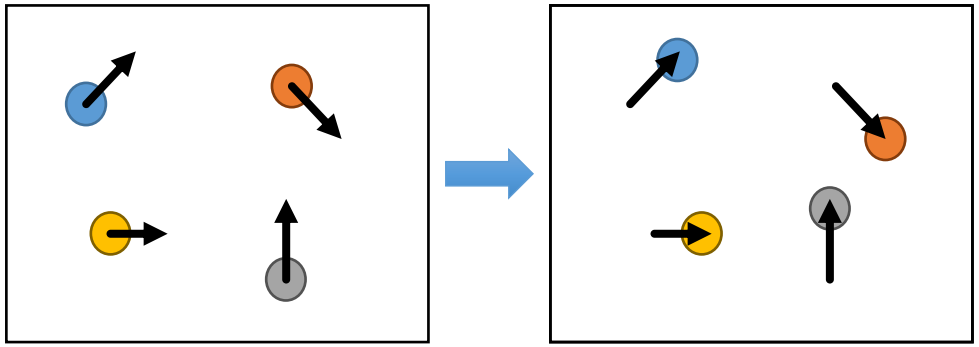
Optical Flow

- Note: apparent motion can be caused by lighting changes without any actual motion.
  - A rotating sphere with a static light source produces a static image, but a stationary sphere with a moving light source produces drifting intensities.
  - Start by estimating motion of each pixel separately, then consider motion of entire image.

v.s



- Estimating optical flow is formulated as:
  - Given two subsequent frames, estimate the apparent motion field $u(x, y)$ and $v(x, y)$ btw them.
  - At the pixel level, we need to solve the following correspondence problem.
  - Given pixel at time t, find nearby pixels of the same color at time $t + 1$

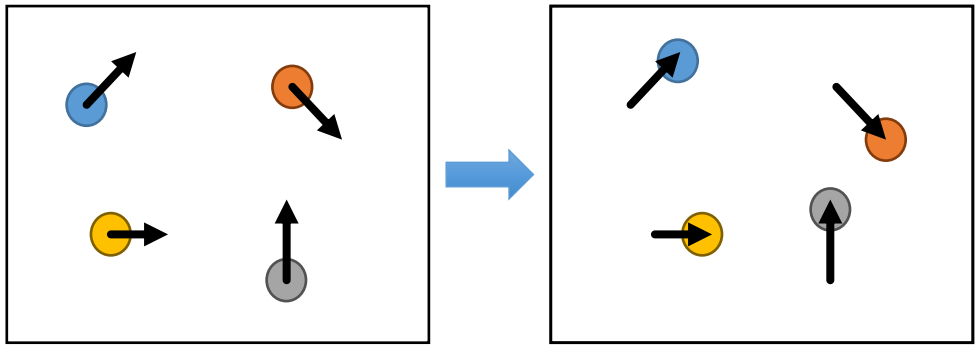


$I(x, y, t)$ $I(x, y, t + 1)$

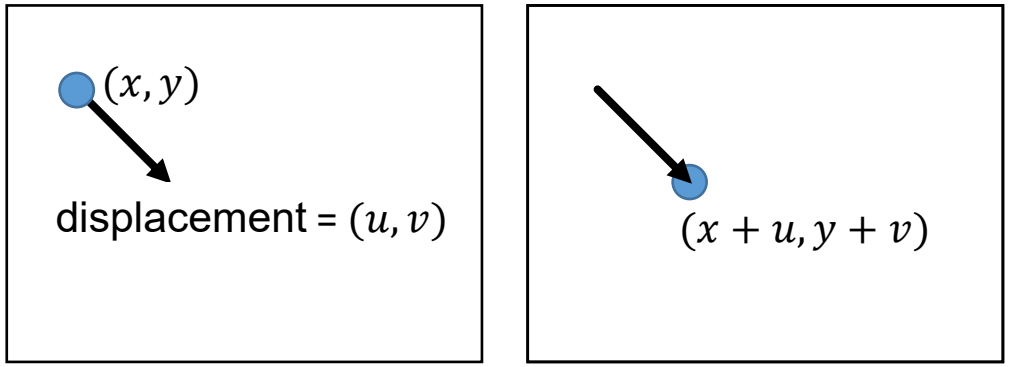
© Copyright National University of Singapore. All Rights Reserved.

Optical Flow: Key to Computation

- Key assumptions:
- **Brightness constancy**: projection of the same point looks the same in every frame.
- **Small motion**: points do not move very far.
- **Spatial coherence**: points move like their neighbors.



$I(x, y, t)$ $I(x, y, t+1)$



$I(x, y, t)$ $I(x, y, t+1)$

© Copyright National University of Singapore. All Rights Reserved.

Optical Flow: Brightness Constancy

- Brightness constancy:

$$I(x, y, t) = I(x + u, y + v, t + 1) \quad \text{Eq. B}$$

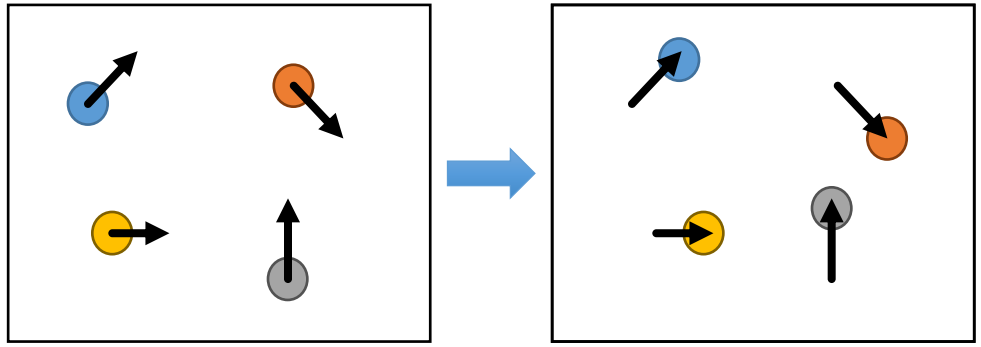
  - It is impossible to find corresponding pixels in a brute force manner.
  - Linearize the right using Taylor expansion with the 1st Taylor series approximation (I_x: gradient of I over x)

$$I(x, y, t) \approx I(x, y, t + 1) + I_x u + I_y v$$

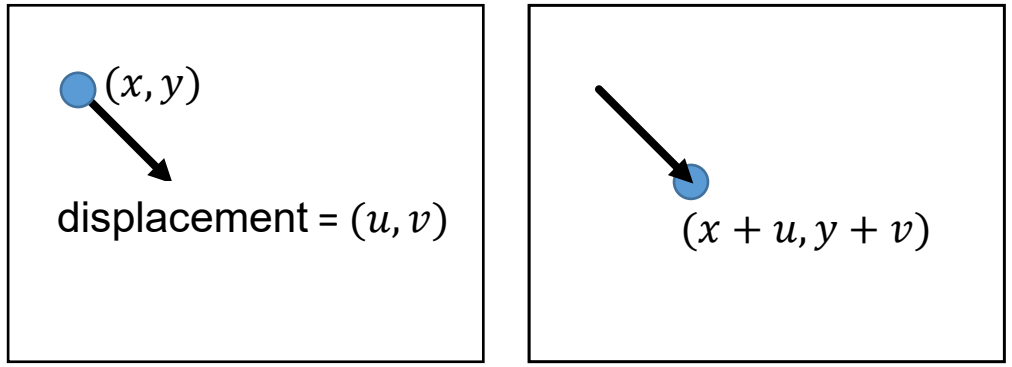
$$\boxed{I(x, y, t + 1) - I(x, y, t)} + I_x u + I_y v = 0$$

Essentially I_t

$$I_x u + I_y v + I_t = 0 \text{ or } \nabla I \cdot (u, v) + I_t = 0$$



$I(x, y, t)$ $\quad$ $I(x, y, t + 1)$



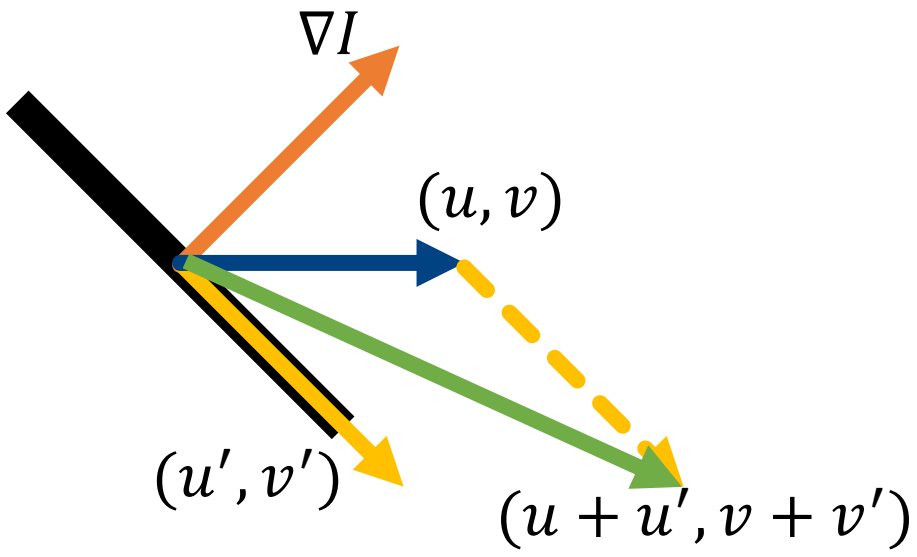
$I(x, y, t)$ $\quad$ $I(x, y, t + 1)$

$$f(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \frac{f^{(3)}(a)}{3!}(x - a)^3 + \ldots + \frac{f^{(n)}(a)}{n!}(x - a)^n + \ldots.$$

© Copyright National University of Singapore. All Rights Reserved.

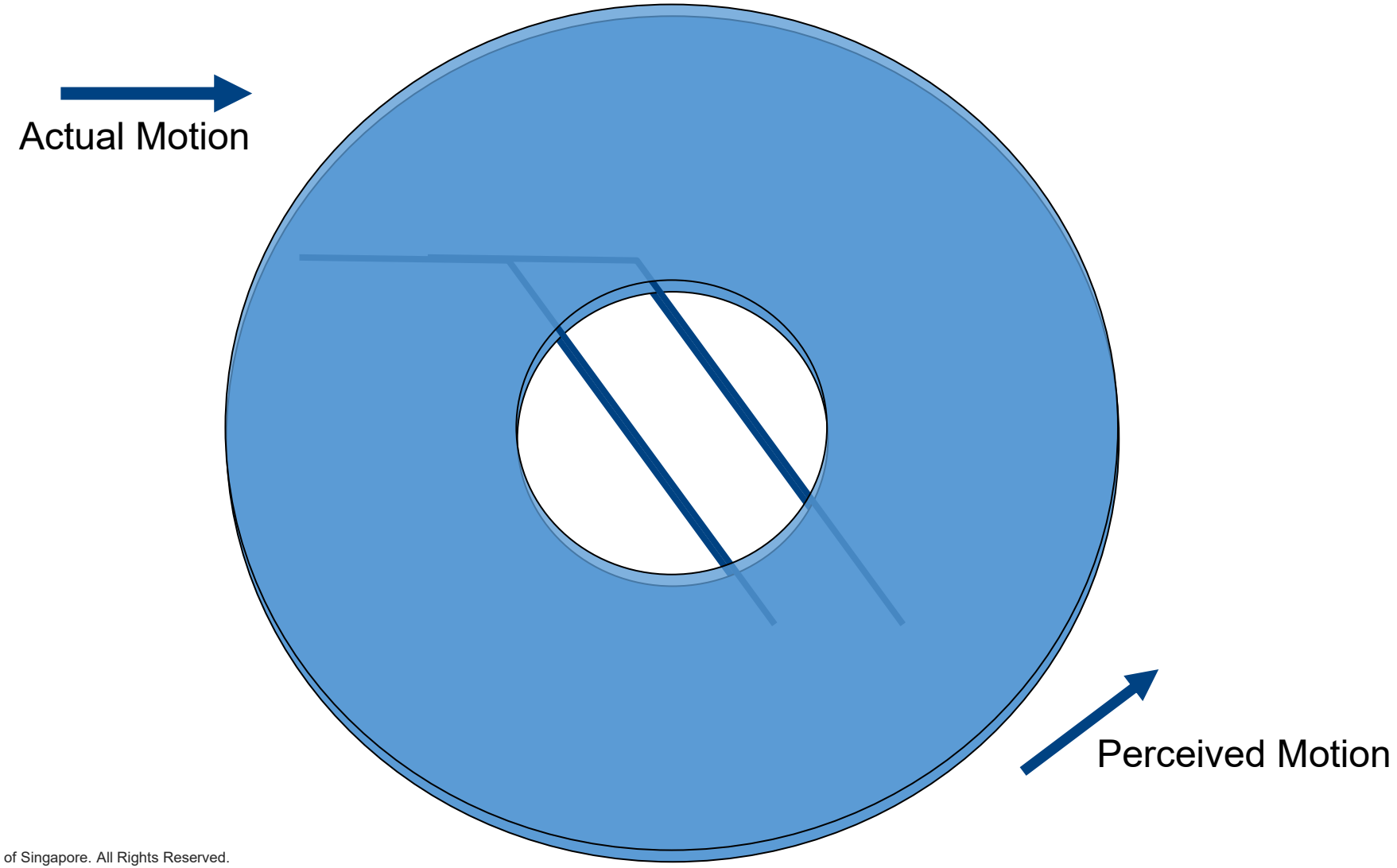
Optical Flow: Brightness Constancy

- What does the equation $I_x u + I_y v + I_t = \nabla I \cdot (u, v) + I_t = 0$ indicates?
- A single equation but two unknowns – trivial solutions $\nabla I \cdot (u, v) = 0$.
- Brightness constancy constraint can only identify the motion along the gradient but **not** the motion perpendicular to the gradient.
- If (u, v) satisfies brightness constancy constraint, so does $(u + u', v + v')$ if $\nabla I \cdot (u', v') = 0$.



© Copyright National University of Singapore. All Rights Reserved.

Optical Flow: Aperture Problem



© Copyright National University of Singapore. All Rights Reserved.

Optical Flow: Aperture Problem

- How to solve the aperture problem?
  - Prevent trivial solutions by getting more equations for a set of unknowns.
  - Spatial coherence constraint: assume the pixel's neighbors have the same (u, v) – **Lucas-Kanade** optical flow.
  - E.g., 5×5 window gives 25 new equations.

$$I_t + I_x u + I_y v = 0$$

$$\begin{bmatrix} I_x(p_1) & I_y(p_1) \\ \vdots & \vdots \\ I_x(p_{25}) & I_y(p_{25}) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = -\begin{bmatrix} I_t(p_1) \\ \vdots \\ I_t(p_{25}) \end{bmatrix}$$

$$p_1 = (x_1, y_1)$$

© Copyright National University of Singapore. All Rights Reserved.

Lucas-Kanade Optical Flow

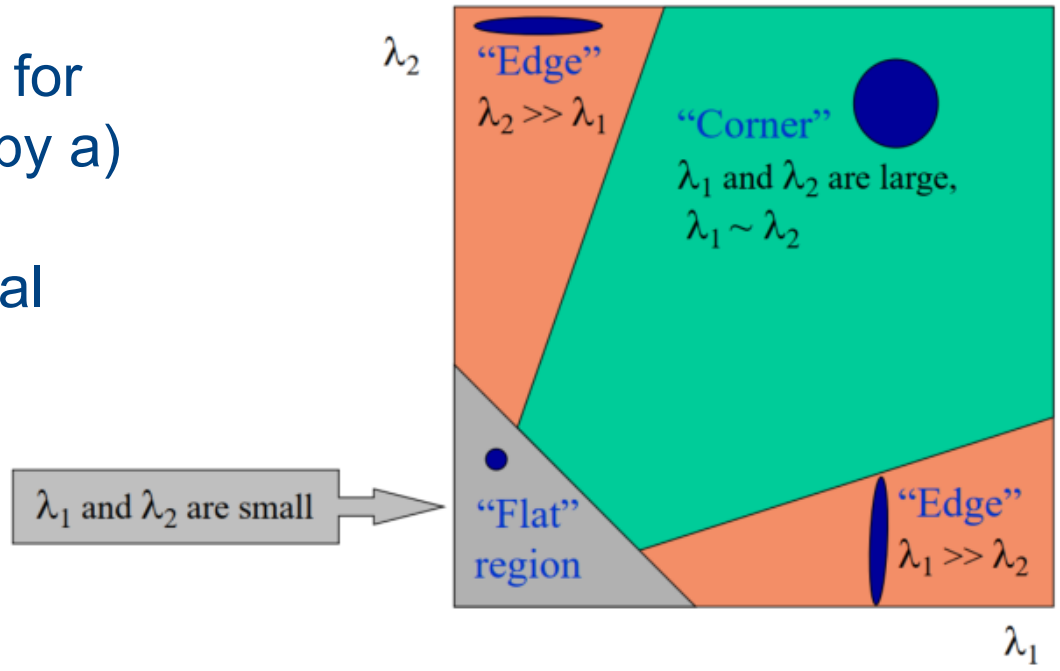
- $\begin{bmatrix} I_x(p_1) & I_y(p_1) \\ \vdots & \vdots \\ I_x(p_{25}) & I_y(p_{25}) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = -\begin{bmatrix} I_t(p_1) \\ \vdots \\ I_t(p_{25}) \end{bmatrix}$ is essentially in the form of: $\underset{25\times 2}{\boldsymbol{A}} \; \underset{2\times 1}{\boldsymbol{d}} = \underset{25\times 1}{\boldsymbol{b}}$
- The solution is given by $(\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{d} = \boldsymbol{A}^T\boldsymbol{b} \rightarrow \boldsymbol{d} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}$.
- In matrix form, we have:
- $\begin{bmatrix} \sum \boldsymbol{I_x I_x} & \sum \boldsymbol{I_x I_y} \\ \sum \boldsymbol{I_x I_y} & \sum \boldsymbol{I_y I_y} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = -\begin{bmatrix} \sum \boldsymbol{I_x I_t} \\ \sum \boldsymbol{I_y I_t} \end{bmatrix}$
- $\boldsymbol{A}^T\boldsymbol{A}$ $\qquad\qquad$ $\boldsymbol{A}^T\boldsymbol{b}$
  - (summation over all pixels in the window)
- The next question becomes when can we find the solution (u, v)

© Copyright National University of Singapore. All Rights Reserved.

Lucas-Kanade Optical Flow

- Define $\boldsymbol{M} = \boldsymbol{A}^T\boldsymbol{A} = \begin{bmatrix} \sum \boldsymbol{I_x I_x} & \sum \boldsymbol{I_x I_y} \\ \sum \boldsymbol{I_x I_y} & \sum \boldsymbol{I_y I_y} \end{bmatrix}$, which is a second moment matrix, and can be decomposed (through eigenvector decomposition) as $\boldsymbol{M} = \boldsymbol{R}^{-1} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \boldsymbol{R}$.

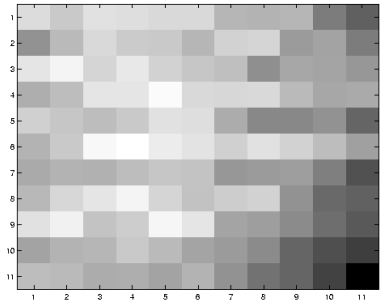
- Estimation of optical flow is well-conditioned for regions with high "cornerness", conditioned by a) $\boldsymbol{\lambda_1}$ **and** $\boldsymbol{\lambda_2}$ **are large**, b) $\boldsymbol{\lambda_1 \sim \lambda_2}$, and c) M is invertible (therefore cannot be precisely equal brightness → Hight texture region.
- λ_1 and λ_2 are eigenvalues.

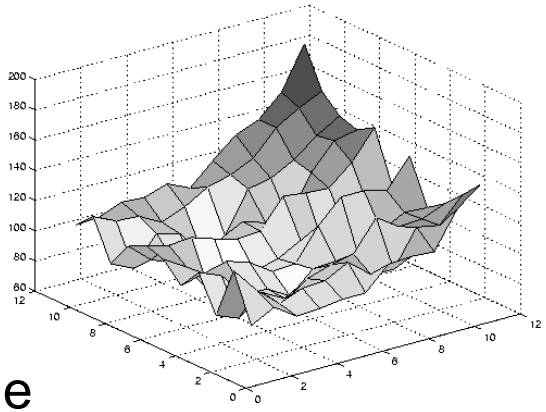


© Copyright National University of Singapore. All Rights Reserved.

Lucas-Kanade Optical Flow: Failed Case

- Failed case 1: Low-texture regions.





$$\begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} = \sum \nabla \mathrm{I} (\nabla \mathrm{I})^{\mathrm{T}}$$

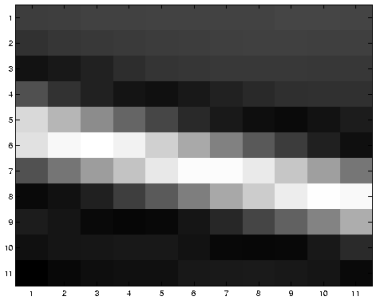
– Gradients have small magnitude

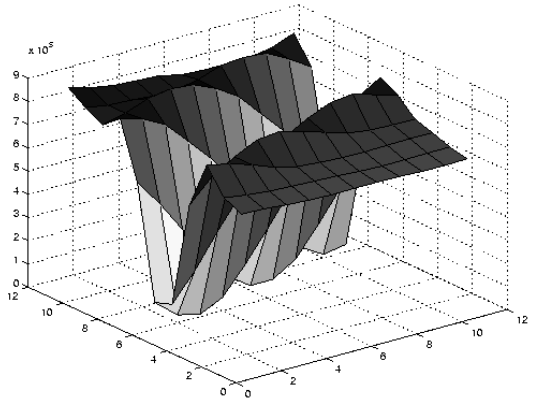
– Small λ_1, small λ_2

© Copyright National University of Singapore. All Rights Reserved.

Lucas-Kanade Optical Flow: Failed Case

- Failed case 2: Edges.





$$\begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} = \sum \nabla \mathrm{I} (\nabla \mathrm{I})^{\mathrm{T}}$$

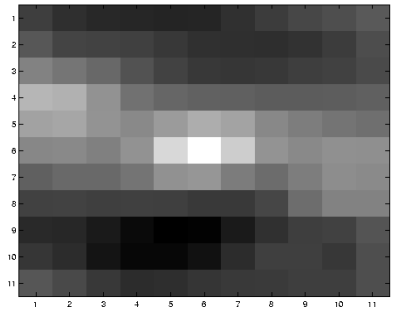
– Large gradients, along a certain direction

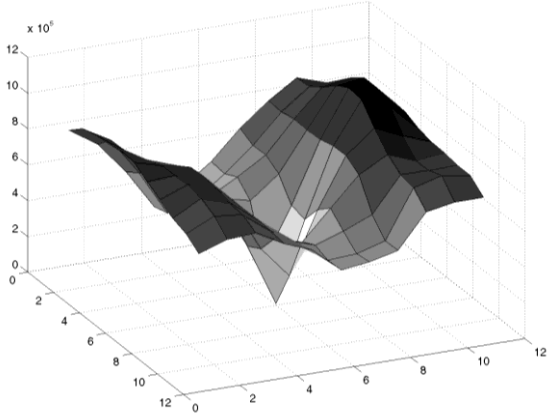
– Large λ_1, small λ_2, or vice versa

© Copyright National University of Singapore. All Rights Reserved.

Lucas-Kanade Optical Flow: Success Case

- Success case: High-texture regions.





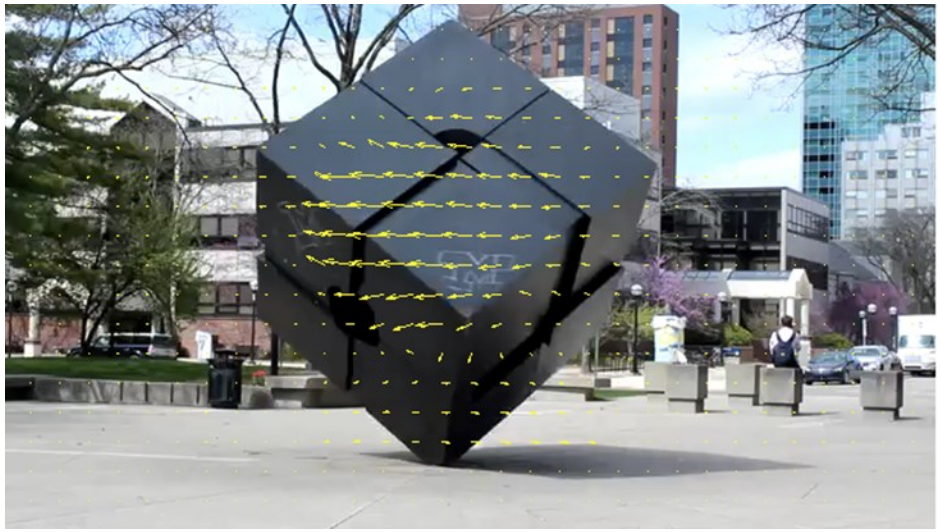
$$\begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} = \sum \nabla \mathrm{I} (\nabla \mathrm{I})^{\mathrm{T}}$$

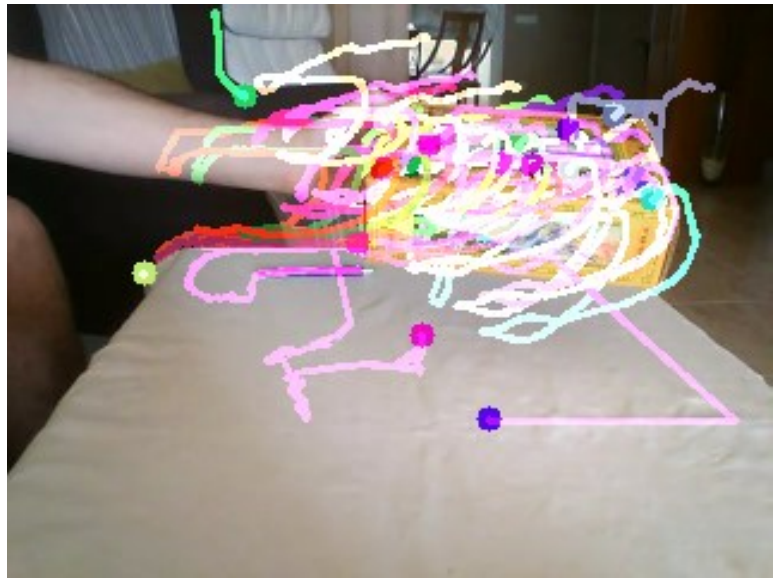
– Gradients are different across different directions, large magnitudes

– Large λ_1, large λ_2

© Copyright National University of Singapore. All Rights Reserved.

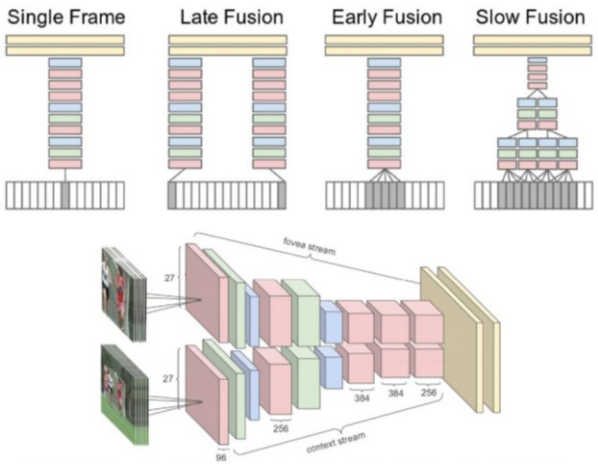
Lucas-Kanade Optical Flow





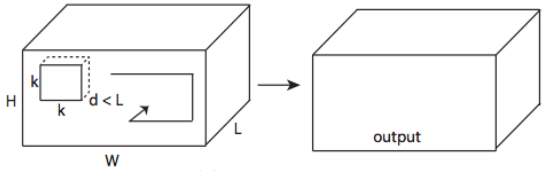
© Copyright National University of Singapore. All Rights Reserved.

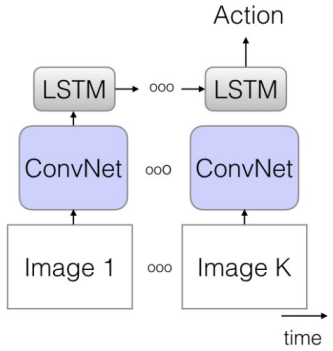
Deep Learning-based Video Analysis: Outline



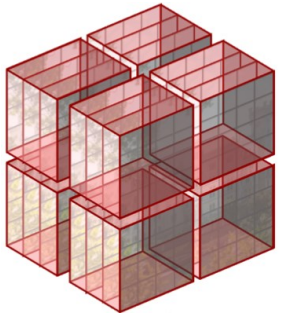
2D CNN

3D CNN

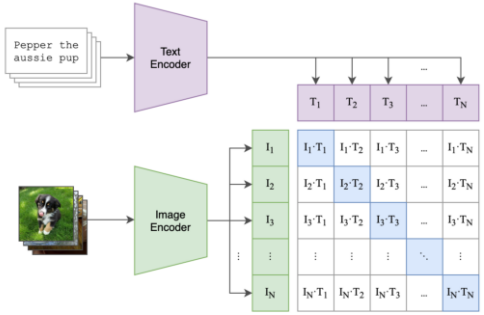




2D CNN with RNN



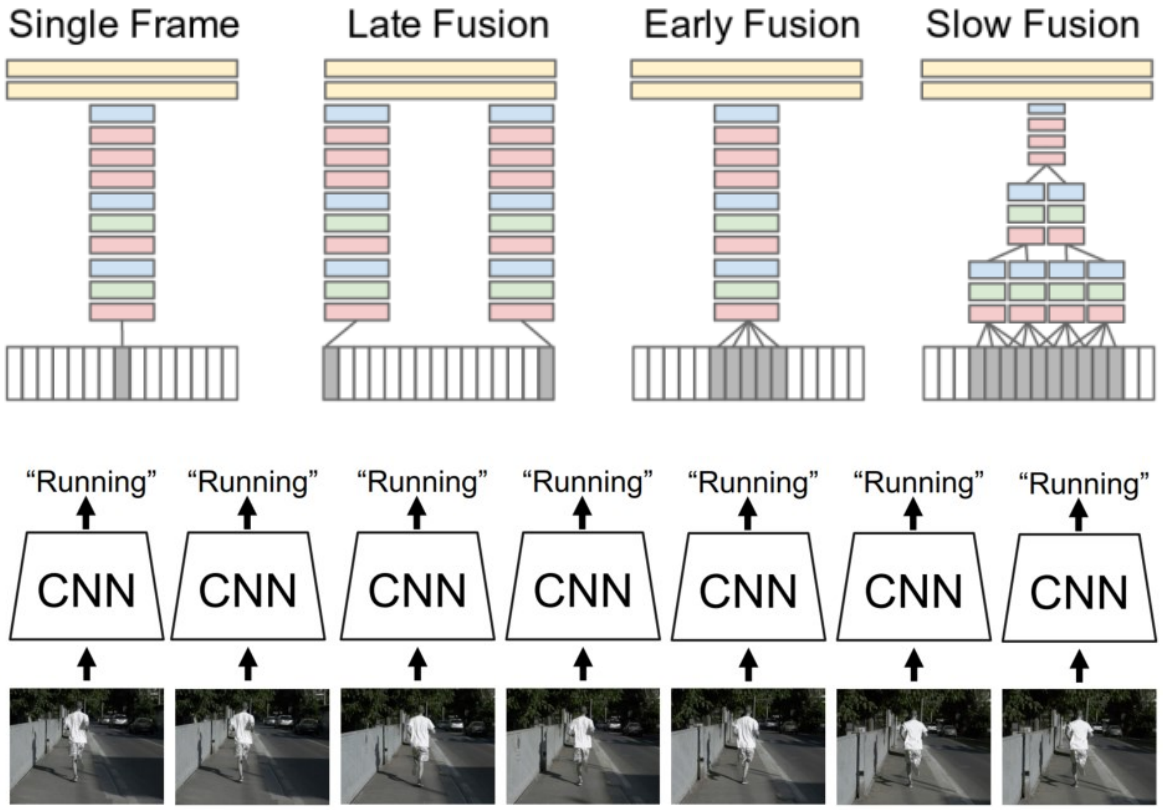
Video Transformer

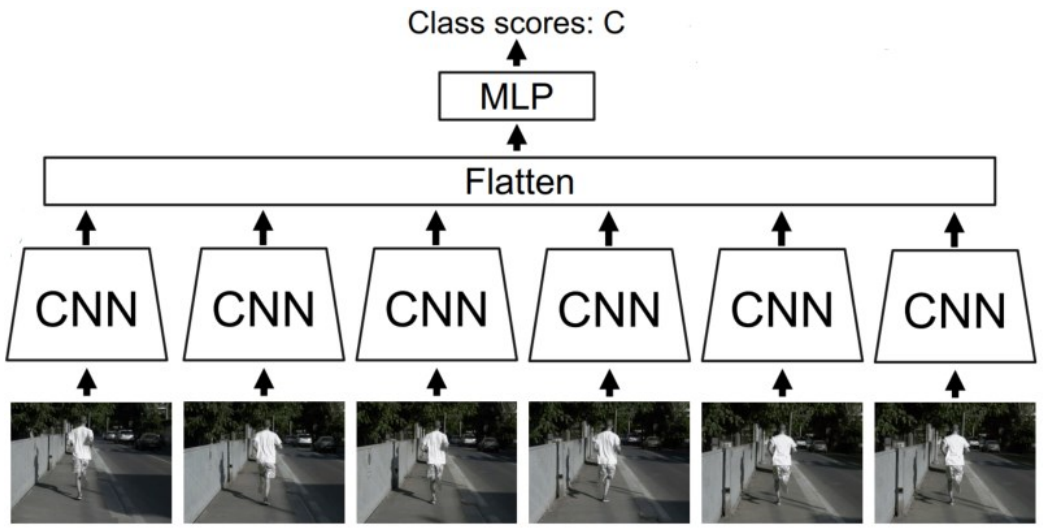


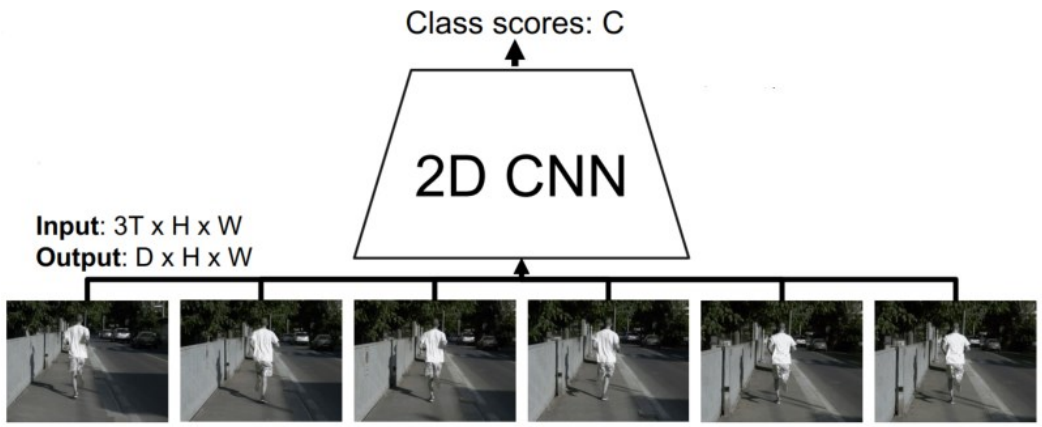
Large Language-Vision Model

© Copyright National University of Singapore. All Rights Reserved.

2D CNN: Fusion of Image Features



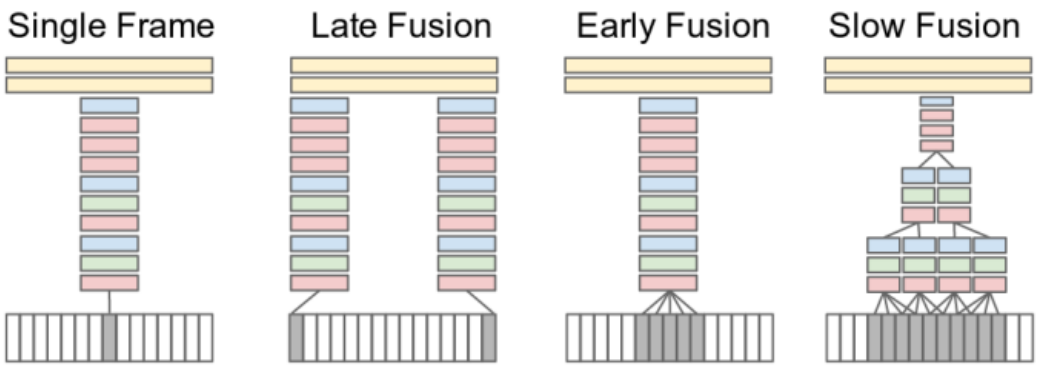


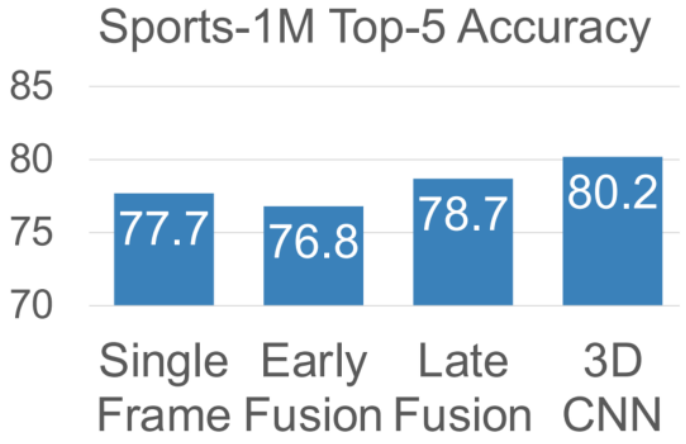


© Copyright National University of Singapore. All Rights Reserved.

2D CNN: Fusion of Image Features

- Advantages and limitations of 2D CNN – fusion of image features:
  - Implementation? Scenarios?
  - Accuracy? Computation Cost?
- The core shortcoming of limitations?
  - Cannot deal with **temporal** information explicitly!





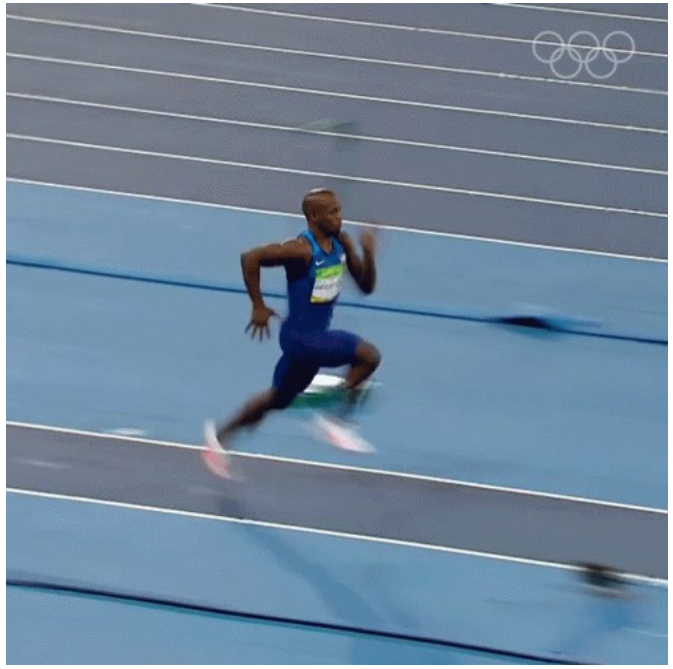
Group	mAP from scratch	mAP fine-tune top 3	mAP fine-tune top
Human-Object Interaction	0.26	0.55	0.52
Body-Motion Only	0.32	0.57	0.52
Human-Human Interaction	0.40	0.68	0.65
Playing Musical Instruments	0.42	0.65	0.46
Sports	**0.57**	**0.79**	**0.80**
All groups	0.44	0.68	0.66

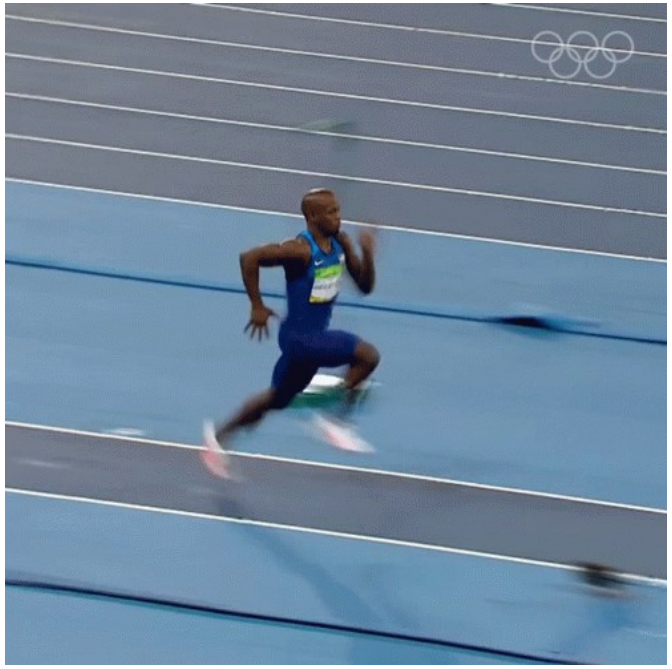
Table 4: Mean Average Precision of the Slow Fusion network on UCF-101 classes broken down by category groups.

© Copyright National University of Singapore. All Rights Reserved.

2D CNN: Fusion of Image Features

- The core shortcoming of disadvantages?
- Cannot deal with temporal information explicitly!

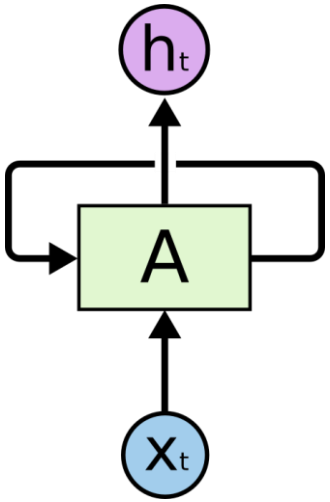


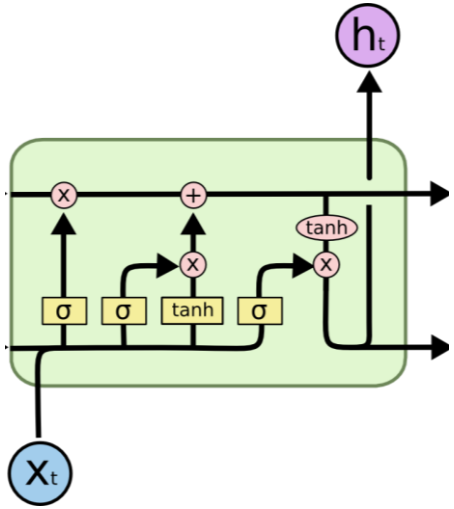


© Copyright National University of Singapore. All Rights Reserved.

Dealing With Temporal Information via RNN / LSTM

- What temporal information/motion information contains essentially?
  - A sequence of **change of appearance**.
  - Both “sequence” and “change of appearance” can be monitored **globally** or **locally**.





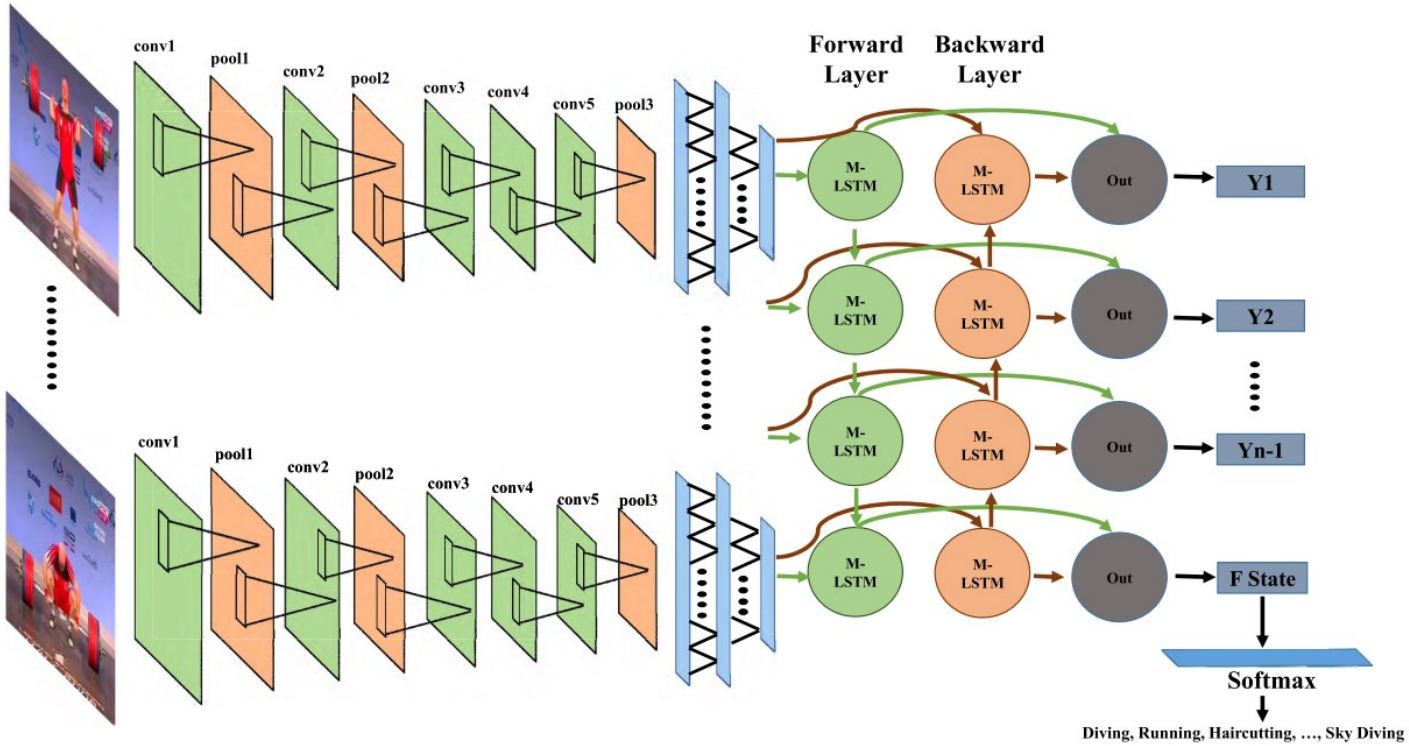


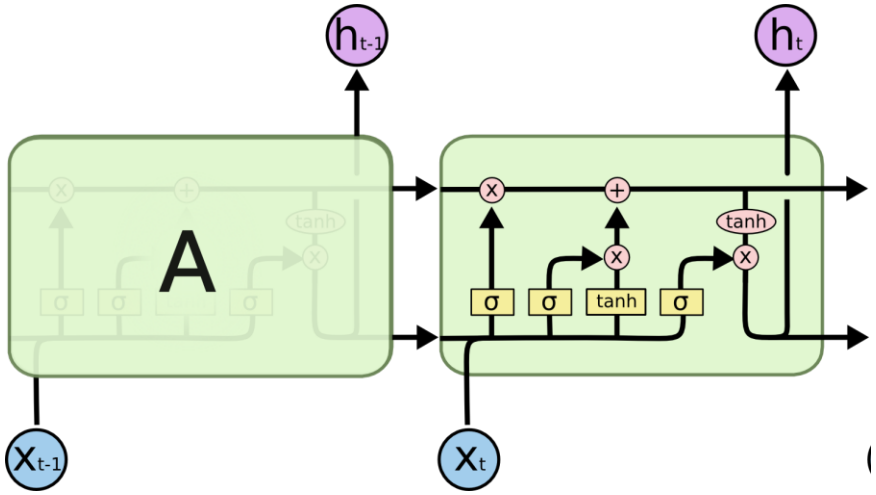
FIGURE 1. **Framework of the proposed DB-LSTM for action recognition.**

© Copyright National University of Singapore. All Rights Reserved.

Dealing With Temporal Information via RNN / LSTM

- Discussion: advantages and limitations of 2D CNN with LSTM:
  - a) Accuracy? b) Computation Cost? c) Implementation?
- Why LSTM are less commonly used for deep learning task?
  - Cannot proceed to next LSTM block w/o the previous block – Very slow training. (Cannot compute in parallel – GPU advantage)

Method	YouTube	HMDB51	UCF101
Multiresolution CNNs [21]	-	-	65.4%
Proposed DB-LSTM	**92.84**	**87.64**	**91.21**



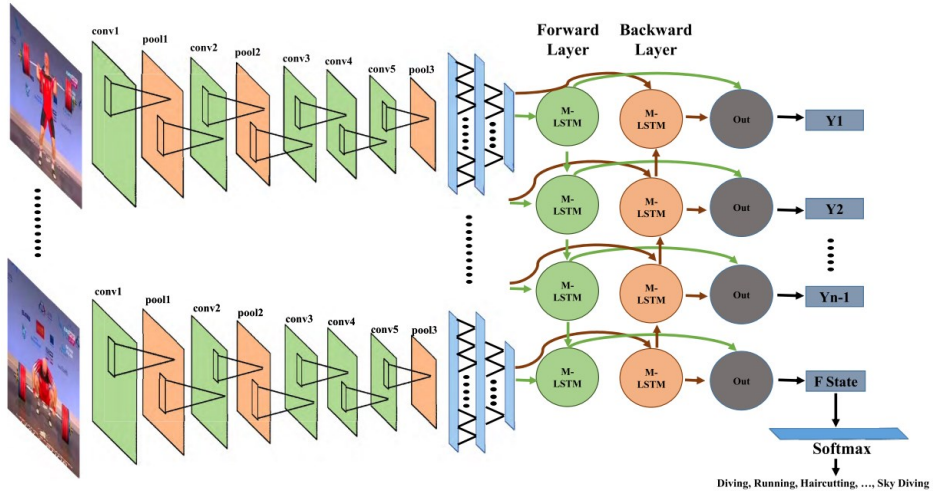
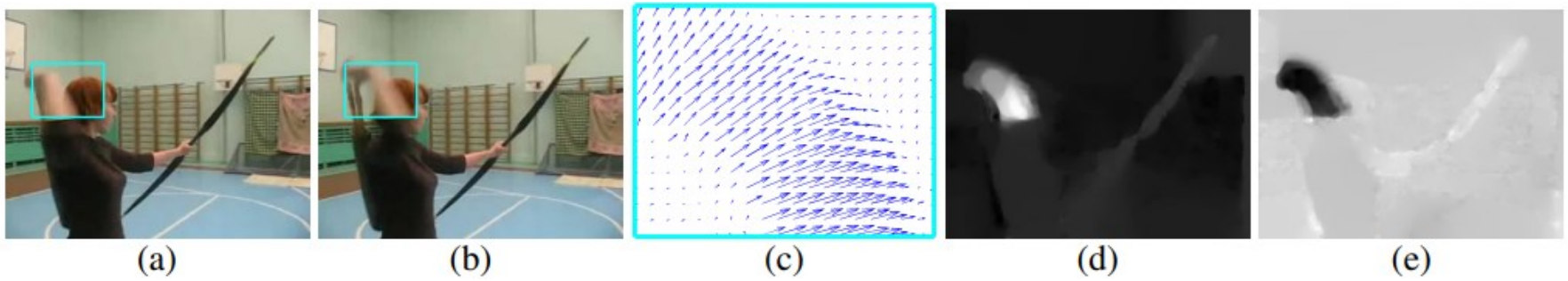


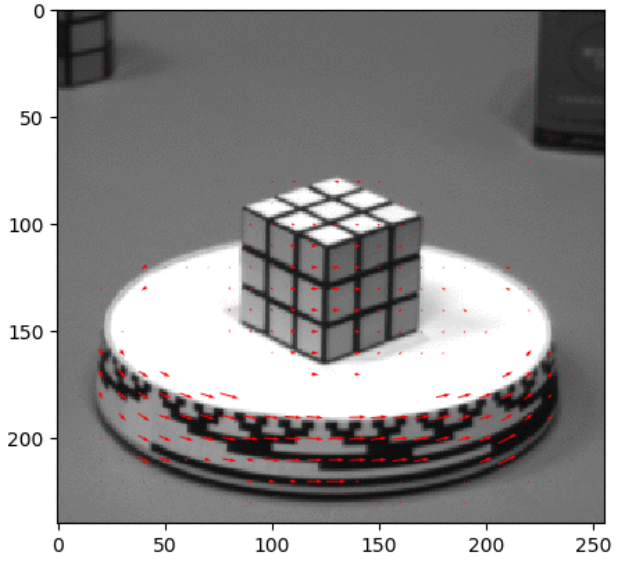
FIGURE 1. Framework of the proposed DB-LSTM for action recognition.

© Copyright National University of Singapore. All Rights Reserved.

Temporal Information from Optical Flow

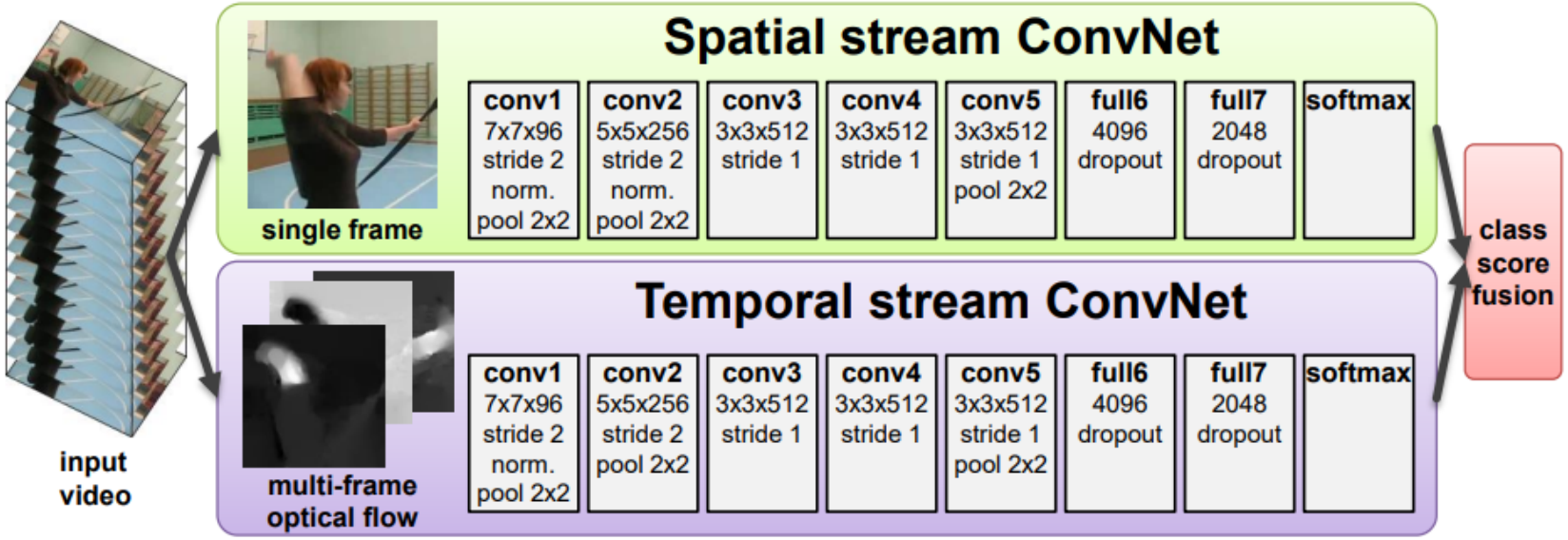
- Other methods to obtain temporal information?
  - Motion information, essentially temporal information, can be obtained from optical flow.
- Intuitive way to deal with temporal information with 2D CNN:
  - Spatial information from a single frame with 2D CNN; temporal information from optical flow frames with 2D CNN.





© Copyright National University of Singapore. All Rights Reserved.

Temporal Information from Optical Flow

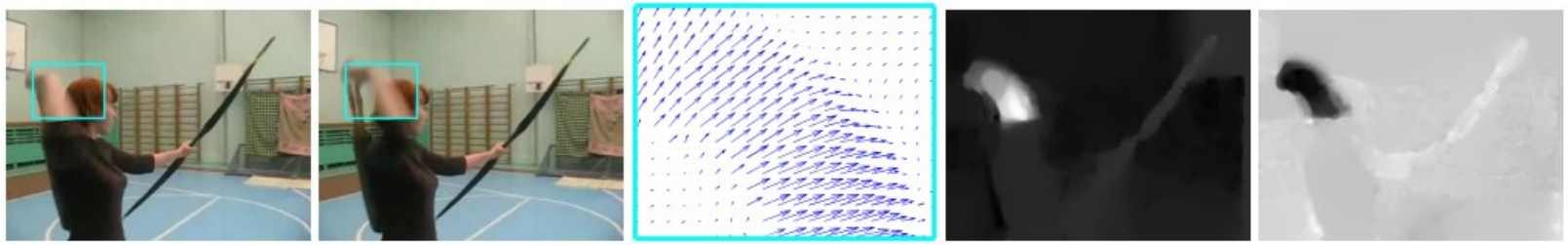


Method	UCF-101	HMDB-51
Improved dense trajectories (IDT) [26, 27]	85.9%	57.2%
IDT with higher-dimensional encodings [20]	**87.9%**	61.1%
IDT with stacked Fisher encoding [21] (based on Deep Fisher Net [23])	-	**66.8%**
Spatio-temporal HMAX network [11, 16]	-	22.8%
"Slow fusion" spatio-temporal ConvNet [14]	65.4%	-
Spatial stream ConvNet	73.0%	40.5%
Temporal stream ConvNet	83.7%	54.6%
Two-stream model (fusion by averaging)	86.9%	58.0%
Two-stream model (fusion by SVM)	**88.0%**	**59.4%**

© Copyright National University of Singapore. All Rights Reserved.

Temporal Information from Optical Flow

- Discussion: advantages and limitations of 2D CNN with LSTM:
  - a) Accuracy? b) Implementation? c) Computation Cost?
- What do we look forward in Deep Learning?
  - End-to-end process.
  - A balance between computation/memory cost and accuracy.

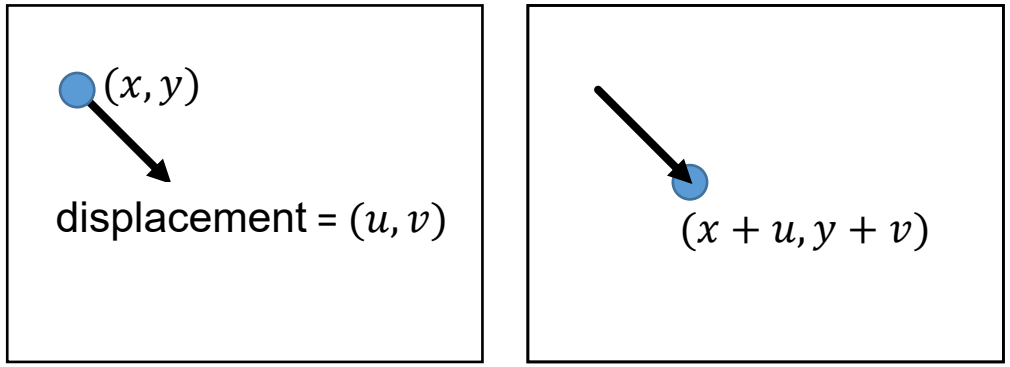


Method	UCF-101	HMDB-51
Improved dense trajectories (IDT) [26, 27]	85.9%	57.2%
IDT with higher-dimensional encodings [20]	**87.9%**	61.1%
IDT with stacked Fisher encoding [21] (based on Deep Fisher Net [23])	-	**66.8%**
Spatio-temporal HMAX network [11, 16]	-	22.8%
"Slow fusion" spatio-temporal ConvNet [14]	65.4%	-
Spatial stream ConvNet	73.0%	40.5%
Temporal stream ConvNet	83.7%	54.6%
Two-stream model (fusion by averaging)	86.9%	58.0%
Two-stream model (fusion by SVM)	**88.0%**	**59.4%**

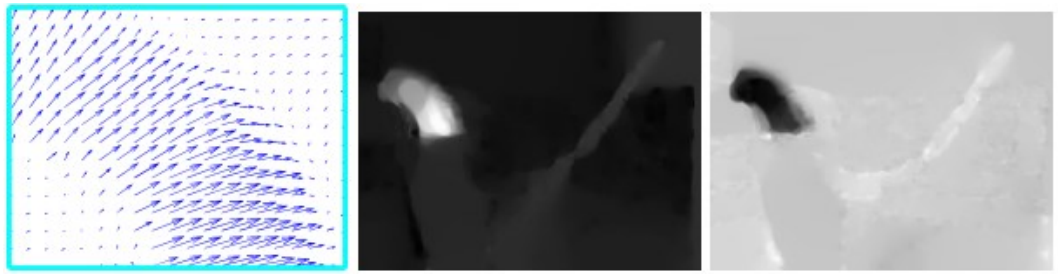
© Copyright National University of Singapore. All Rights Reserved.

Temporal Information from Optical Flow

- Why is Optical Flow still leveraged even until today?
- What are the limitations of Optical Flow?
  - NOT end-to-end process.
  - High memory cost. (1MB of RGB → about 3MB of Optical Flow)



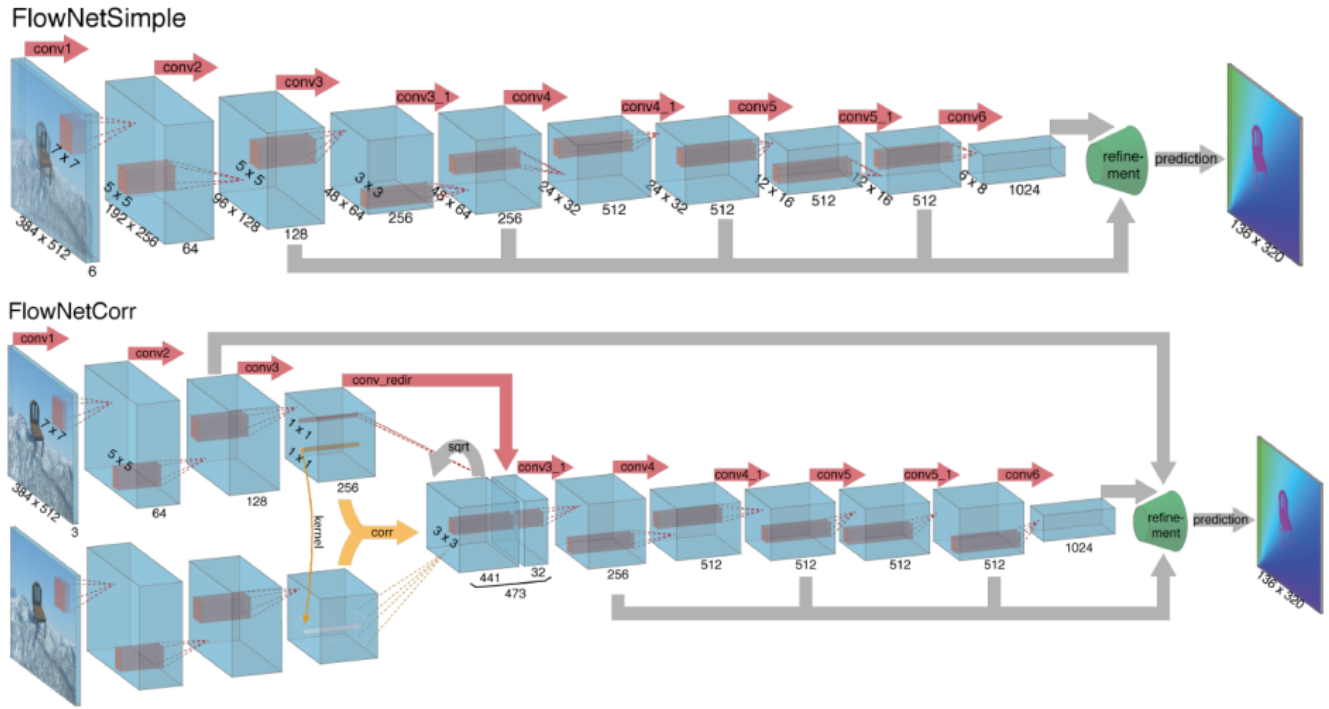
$I(x, y, t)$ $I(x, y, t + 1)$



© Copyright National University of Singapore. All Rights Reserved.

Estimating Optical Flow with Deep Learning

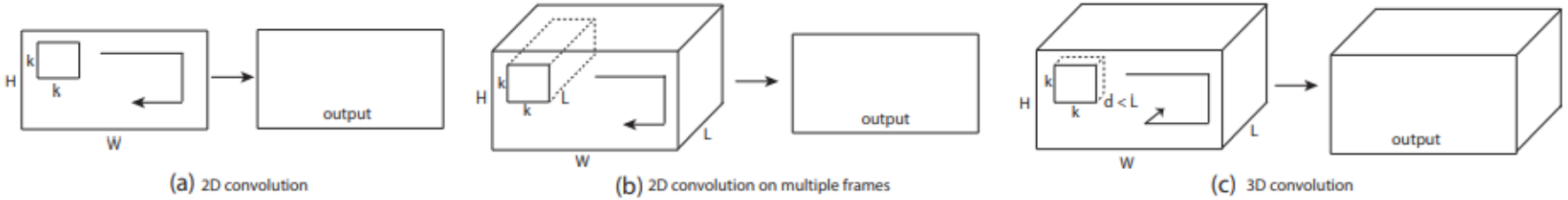
- Merging optical flow estimation with deep learning structure:
  - End-to-end optical flow with neural networks
  - Note: optical flow is obtained from AT LEAST two video frames.



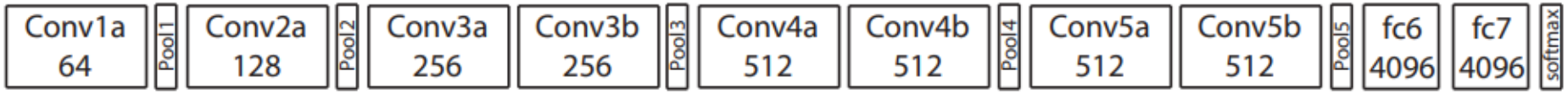
© Copyright National University of Singapore. All Rights Reserved.

Joint Modeling of Spatial and Temporal Info

- What do we look forward in Deep Learning?
  - End-to-end process – model spatial and temporal information jointly?
  - A balance between computation/memory cost and accuracy.
- 3D CNN: expand 2D Conv kernels to the temporal axis. (C3D)



(a) 2D convolution (b) 2D convolution on multiple frames (c) 3D convolution

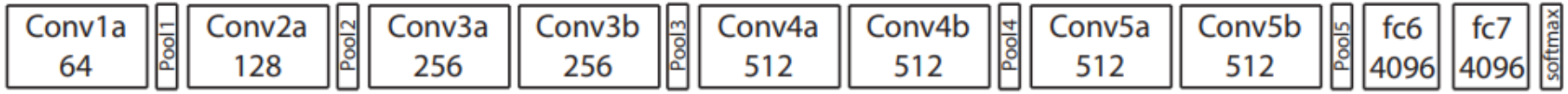


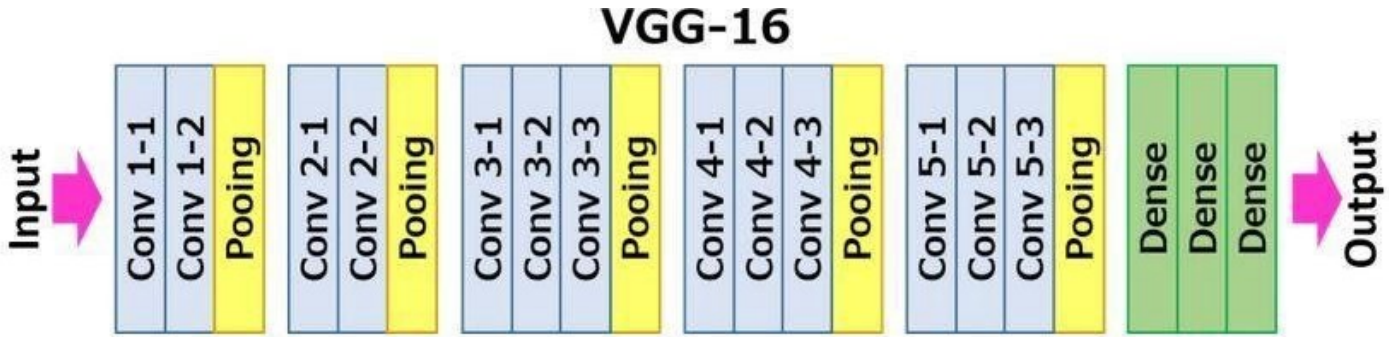
VGG-16

© Copyright National University of Singapore. All Rights Reserved.

Joint Modeling of Spatial and Temporal Info

- What do we look forward in Deep Learning?
  - End-to-end process – model spatial and temporal information jointly?
  - A balance between computation/memory cost and accuracy.
  - Essentially, C3D is only a 10/11 layer-CNN (only Conv and FC layers).





Method	Accuracy (%)
Imagenet + linear SVM	68.8
iDT w/ BoW + linear SVM	76.2
Deep networks [18]	65.4
Spatial stream network [36]	72.6
LRCN [6]	71.1
LSTM composite model [39]	75.8
C3D (1 net) + linear SVM	82.3

© Copyright National University of Singapore. All Rights Reserved.

Joint Modeling of Spatial and Temporal Info

- Computation cost of C3D?
  - Its computation cost (in GFLOPs) is 2× that of VGG-16.
  - Require initialize of all 3D kernels, which is also computationally costly.

- Any approaches to reduce cost?
  - Intuitive approach: easier initialization

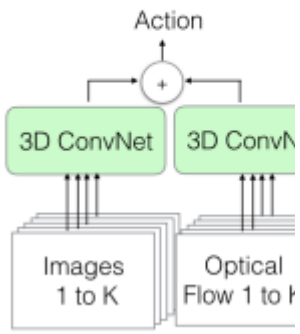
Layer	Size	MFLOPs
Input	3 x 16 x 112 x 112	
Conv1 (3x3x3)	64 x 16 x 112 x 112	1.04
Pool1 (1x2x2)	64 x 16 x 56 x 56	
Conv2 (3x3x3)	128 x 16 x 56 x 56	11.10
Pool2 (2x2x2)	128 x 8 x 28 x 28	
Conv3a (3x3x3)	256 x 8 x 28 x 28	5.55
Conv3b (3x3x3)	256 x 8 x 28 x 28	11.10
Pool3 (2x2x2)	256 x 4 x 14 x 14	
Conv4a (3x3x3)	512 x 4 x 14 x 14	2.77
Conv4b (3x3x3)	512 x 4 x 14 x 14	5.55
Pool4 (2x2x2)	512 x 2 x 7 x 7	
Conv5a (3x3x3)	512 x 2 x 7 x 7	0.69
Conv5b (3x3x3)	512 x 2 x 7 x 7	0.69
Pool5	512 x 1 x 3 x 3	
FC6	4096	0.51
FC7	4096	0.45
FC8	C	0.05

AlexNet: 0.7 GFLOP
VGG-16: 13.6 GFLOP
C3D: **39.5 GFLOP (2.9x VGG!)**

© Copyright National University of Singapore. All Rights Reserved.

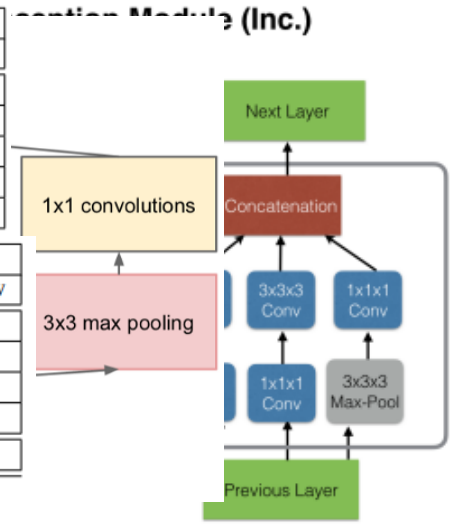
Joint Modeling of Spatial and Temporal Info

- I3D: inflating 2D CNNs to 3D;
  - Bootstrapping 3D filters from 2D filters, i.e., initialize through copying 2D parameters across temporal dimension.
  - Easy construct of 3D CNNs (Inception-I3D, subsequently ResNet-I3D)
  - Leverage on efficient structure from the original 2D CNNs (Inception is more efficient thanks to its dimension reduction design across channels)



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

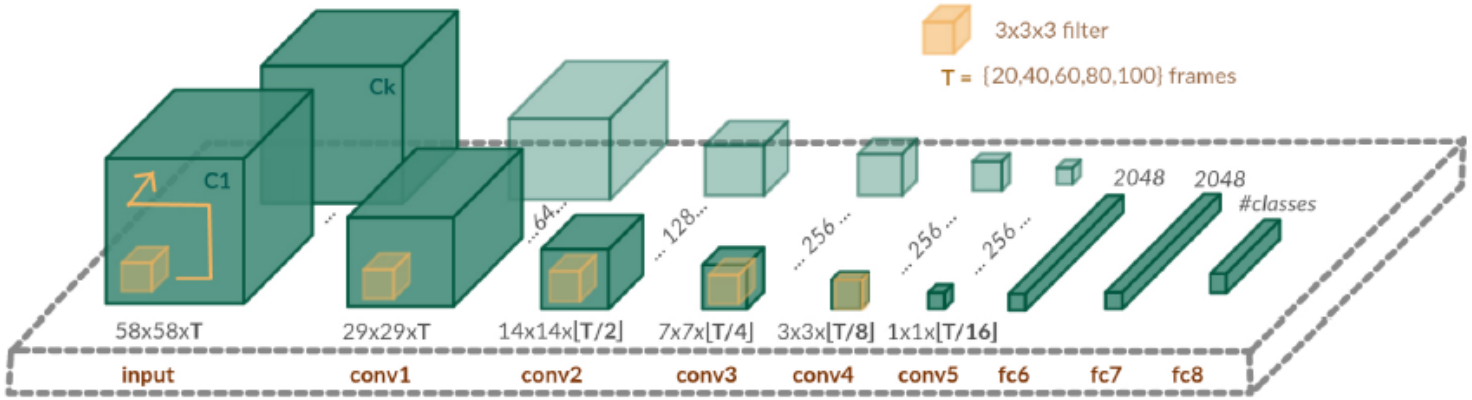
Architecture	UCF-101			HMDB-51			miniKinetics		
	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow
(a) LSTM	81.0	–	–	36.0	–	–	69.9	–	–
(b) 3D-ConvNet	51.6	–	–	24.3	–	–	60.0	–	–
(c) Two-Stream	83.6	85.6	91.2	43.2	56.3	58.3	70.1	58.4	72.9
(d) 3D-Fused	83.2	85.8	89.3	49.2	55.5	56.8	71.4	61.0	74.0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

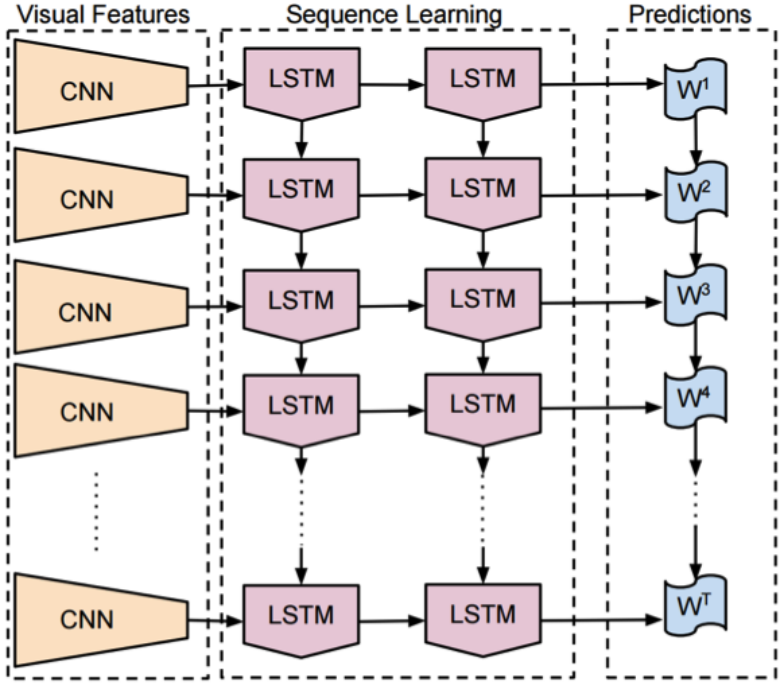


© Copyright National University of Singapore. All Rights Reserved.

Global/Long-Term Temporal Information

- Discussion on temporal information/motion information:
  - Temporal / motion information can be monitored **globally** or **locally**.
- RNN/LSTM and 3D CNNs extract temporal information “locally”.
  - Global temporal information obtained by stacking RNN / CNN blocks.

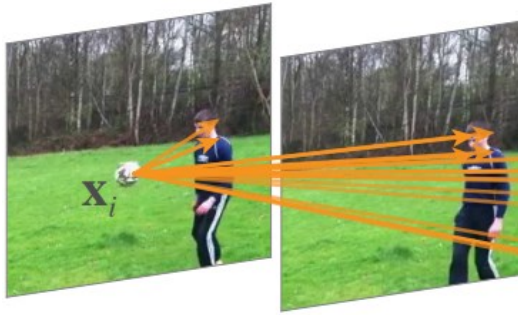


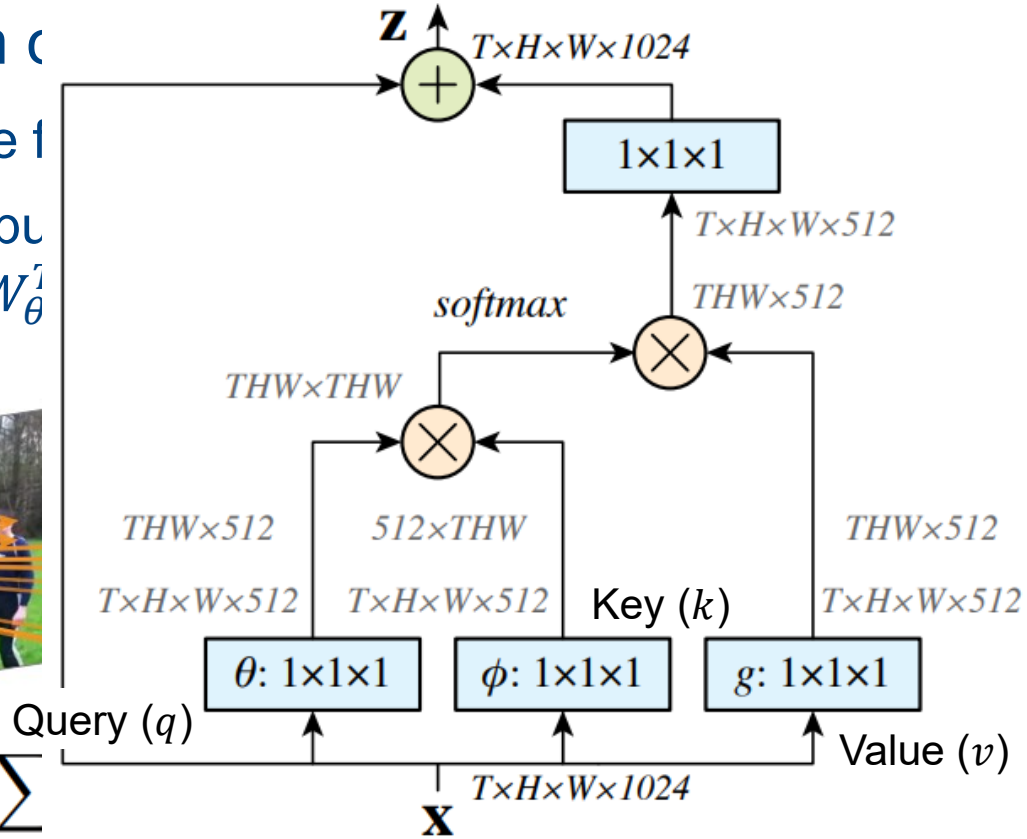


© Copyright National University of Singapore. All Rights Reserved.

Global/Long-Term Temporal Information

- The Non-Local (NL) module: capturing long-range dependencies with a non-local operation that computes the response at a position as a weighted sum c input feature maps.
  - If the pairwise f $\frac{1}{\mathbf{x})} f(\mathbf{x}_i, \mathbf{x}_j)$ is equal to the *softmax* compu an be re-written as: $\mathbf{y} =$ $softmax(\mathbf{x}^T W_\theta$ **mer**"-style self-attention!





$$\mathbf{y}_i = \frac{1}{\mathcal{C}(\mathbf{x})} \sum_{\forall j}$$

ssian

$$f(\mathbf{x}_i, \mathbf{x}_j) = e^{\mathbf{x}_i^T \mathbf{x}_j}.$$

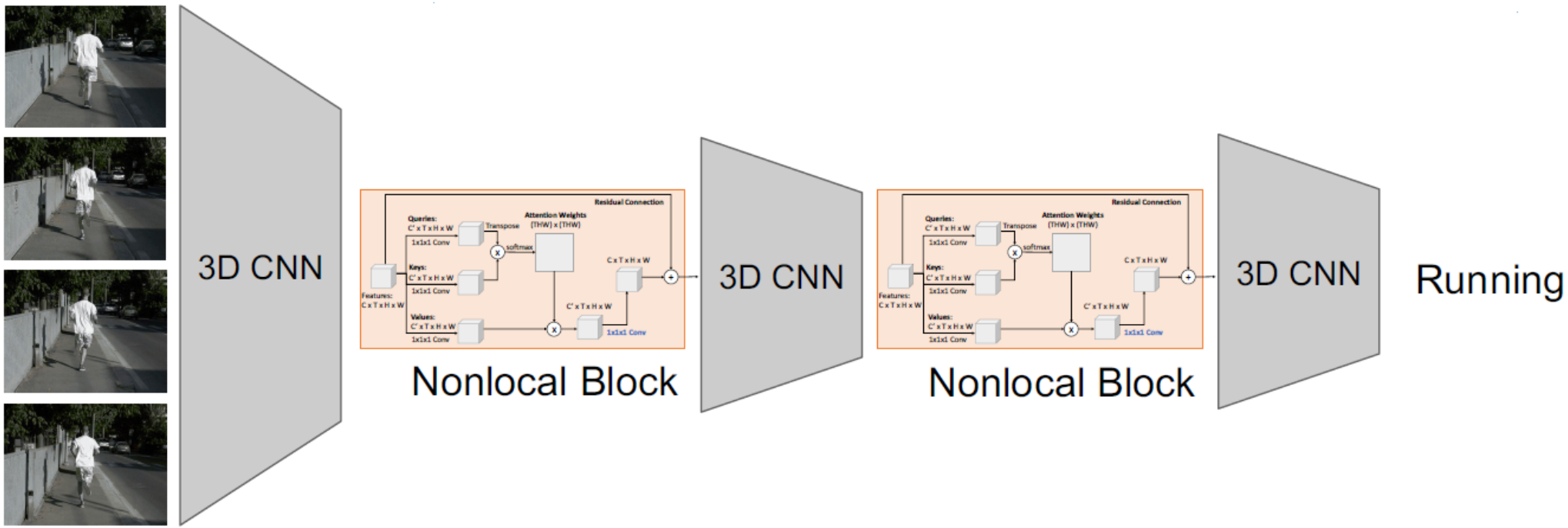
$$f(\mathbf{x}_i, \mathbf{x}_j) = e^{\theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j)}.$$

$$f(\mathbf{x}_i, \mathbf{x}_j) = \theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j).$$

$$f(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{ReLU}(\mathbf{w}_f^T [\theta(\mathbf{x}_i),\ \phi(\mathbf{x}_j)]).$$

© Copyright National University of Singapore. All Rights Reserved.

Global/Long-Term Temporal Information



model	backbone	modality	top-1 val	top-5 val	top-1 test	top-5 test	avg test†
I3D in [7]	Inception	RGB	72.1	90.3	71.1	89.3	80.2
2-Stream I3D in [7]	Inception	RGB + flow	75.7	92.0	74.2	91.3	82.8
RGB baseline in [3]	Inception-ResNet-v2	RGB	73.0	90.9	-	-	-
3-stream late fusion [3]	Inception-ResNet-v2	RGB + flow + audio	74.9	91.6	-	-	-
3-stream LSTM [3]	Inception-ResNet-v2	RGB + flow + audio	77.1	93.2	-	-	-
3-stream SATT [3]	Inception-ResNet-v2	RGB + flow + audio	77.7	93.2	-	-	-
NL I3D [ours]	ResNet-50	RGB	76.5	92.6	-	-	-
	ResNet-101	RGB	**77.7**	**93.3**	-	-	**83.8**

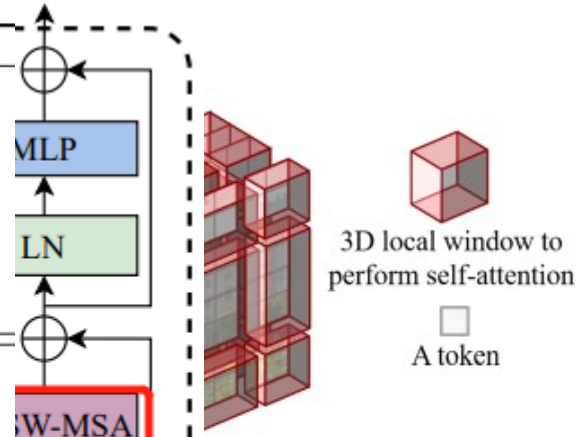
© Copyright National University of Singapore. All Rights Reserved.

Global/Long-Term Temporal Information

- Since obtaining global/long-term temporal information is useful towards HAR, can we apply it w/o 3D-CNN (pure global temporal feature)?
  - Video Transformer (Video Swin Transformer as example)

Embed token

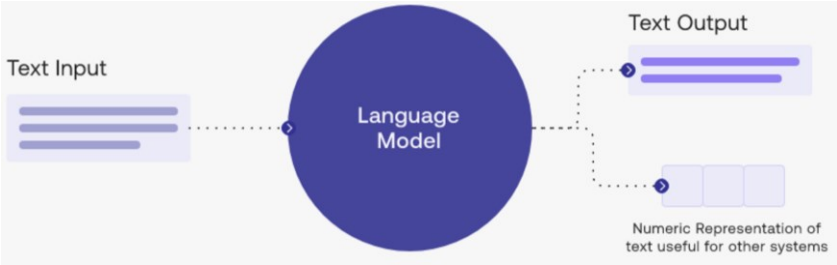
Method	Pretrain	Top-1	Top-5	Views	FLOPs	Param
R(2+1)D [37]	-	72.0	90.0	10 × 1	75	61.8
I3D [6]	ImageNet-1K	72.1	90.3	-	108	25.0
NL I3D-101 [40]	ImageNet-1K	77.7	93.3	10 × 3	359	61.8
ip-CSN-152 [36]	-	77.8	92.8	10 × 3	109	32.8
CorrNet-101 [39]	-	79.2	-	10 × 3	224	-
SlowFast R101+NL [13]	-	79.8	93.9	10 × 3	234	59.9
X3D-XXL [12]	-	80.4	94.6	10 × 3	144	20.3
Swin-T	ImageNet-1K	78.8	93.6	4 × 3	88	28.2
Swin-S	ImageNet-1K	80.6	94.5	4 × 3	166	49.8
Swin-B	ImageNet-1K	80.6	94.6	4 × 3	282	88.1
Swin-B	ImageNet-21K	82.7	95.5	4 × 3	282	88.1
Swin-L	ImageNet-21K	83.1	95.9	4 × 3	604	197.0
Swin-L (384↑)	ImageNet-21K	84.6	96.5	4 × 3	2107	200.0
Swin-L (384↑)	ImageNet-21K	**84.9**	**96.7**	10 × 5	2107	200.0

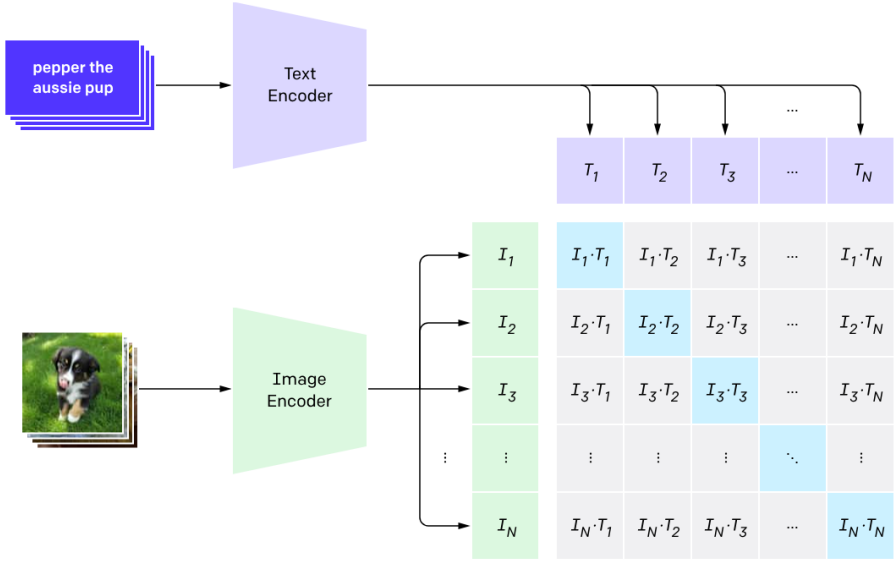


© Copyright National University of Singapore. All Rights Reserved.

From Numerical to Semantic Classes (LLVM)

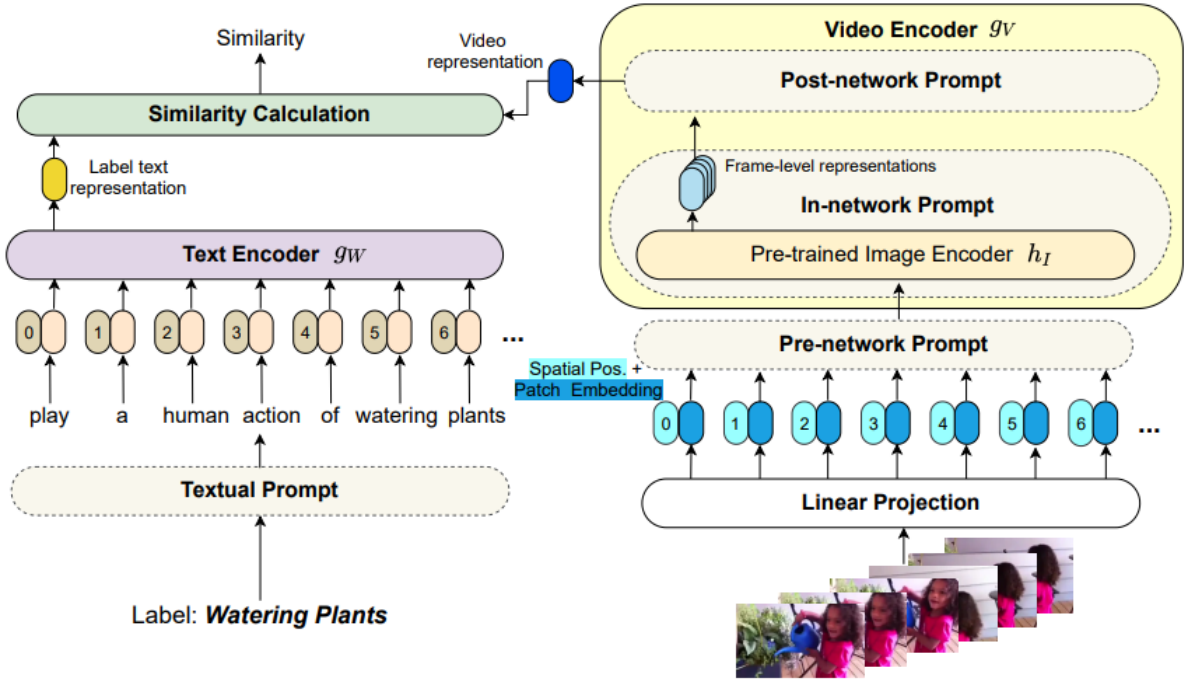
- The success of language models: obtaining semantic information from words (text features).
  - One-hot classes vs. a descriptive sentence? (E.g., class “Walking” vs description “A video of a man walking”)
  - How can we leverage semantic texts rather than numerical classes?
  - Visual Prompt + Textual Prompt: aligning visual information with text information.

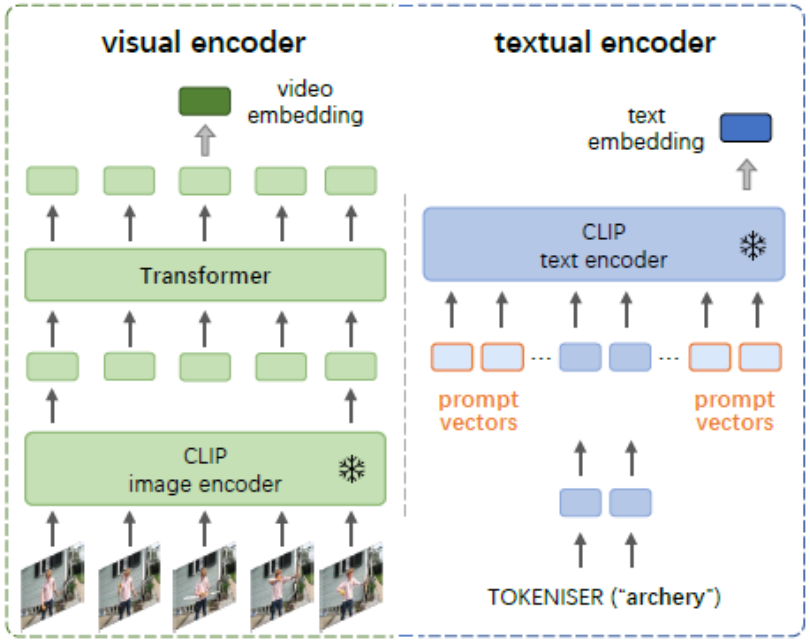




From Numerical to Semantic Classes (LLVM)

- Intuitively: applying the same strategy towards videos
- Open question: what is the challenge of adapting this strategy to videos?
  - The preciseness of the descriptive text.
  - The computation cost of training.





© Copyright National University of Singapore. All Rights Reserved.

3D Analysis:

The Basic, the Tasks and Development

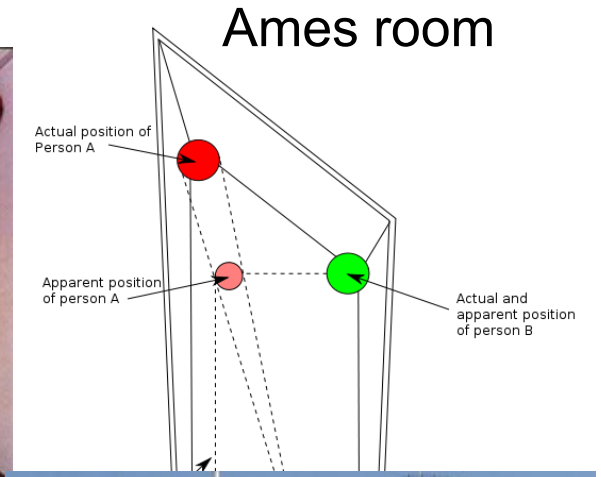
- ✓ Depth Estimation: Starting from Stereo
- ✓ Point Cloud and Voxels
- ✓ DL in 3D Analysis

© Copyright National University of Singapore. All Rights Reserved.

What Images Cannot Perceive: Depth

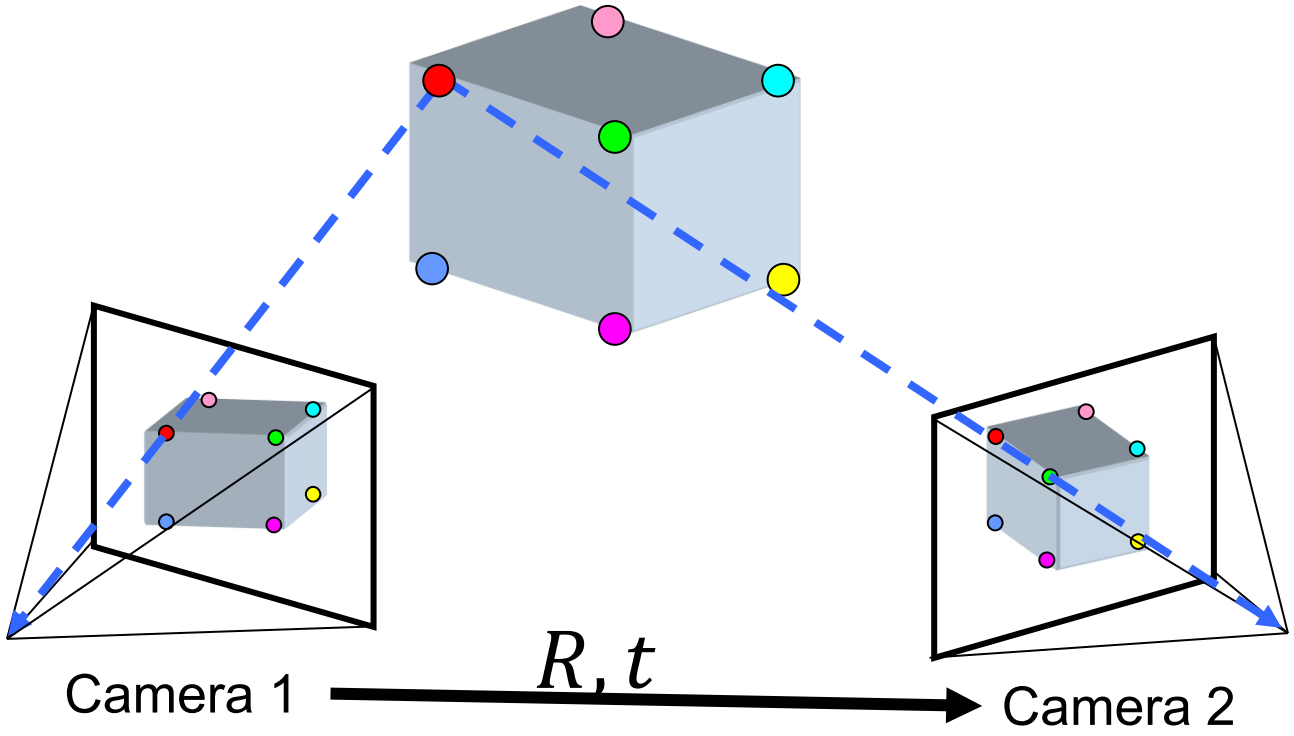
- Single-view Ambiguity: What we can / cannot understand from images?
  - Only know how a 3D point in world space corresponds to each pixel
  - Can only project 3D points to 2D points, process cannot reverse.
  - Nowhere near to find 3D point from 2D point

Ames room



© Copyright National University of Singapore. All Rights Reserved.

Resolving Single-view Ambiguity: Two-view Geometry

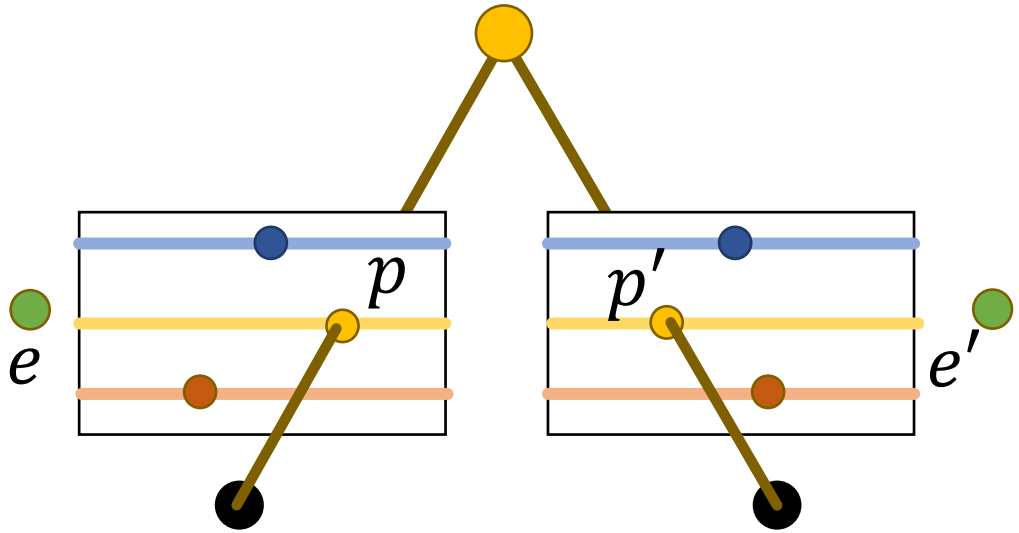


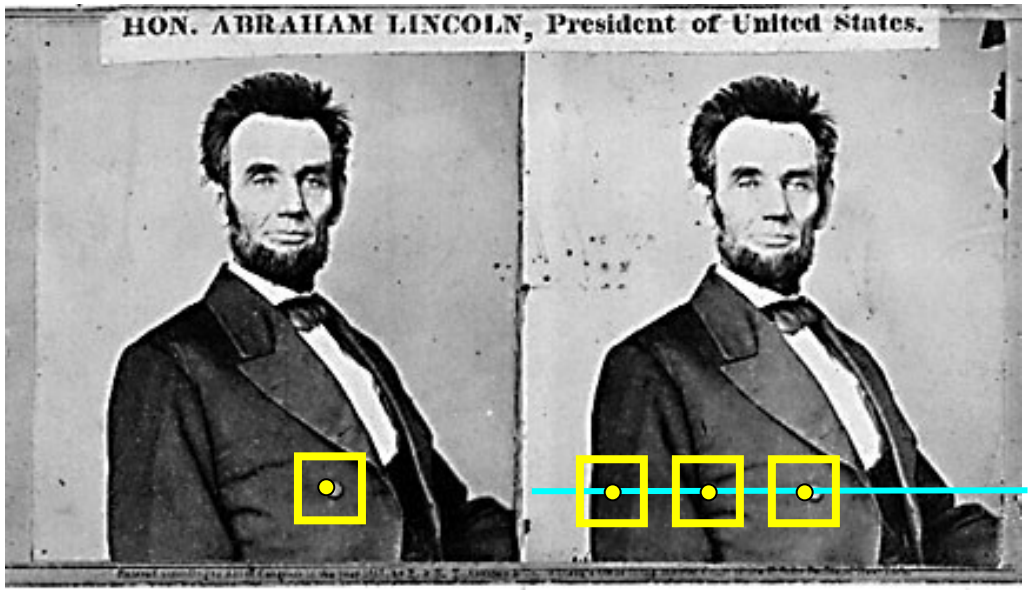
- Two-view geometry:
  - Stereo/Epipolar geomery:
  - Given 2 cameras with the extrinsic matrix (i.e., Rotation and Translation matrix) of Camera 2 with respect to Camera 1 known, find where a point could be in 3D space.
  - Stereo geometry is analyzed under a special scenario where the two cameras are parallel to each other.
  - The key: finding correspondence.

© Copyright National University of Singapore. All Rights Reserved.

Stereo Geometry: Two Parallel Image Planes

- Similar to how humans perceive the environment through stereo vision.
- Image planes of cameras are parallel to each other, while the camera centers are set at same height, the focal lengths are the same.
- Correspondence: at the horizontal parallel line.





© Copyright National University of Singapore. All Rights Reserved.

Stereo Geometry: Depth from Disparity

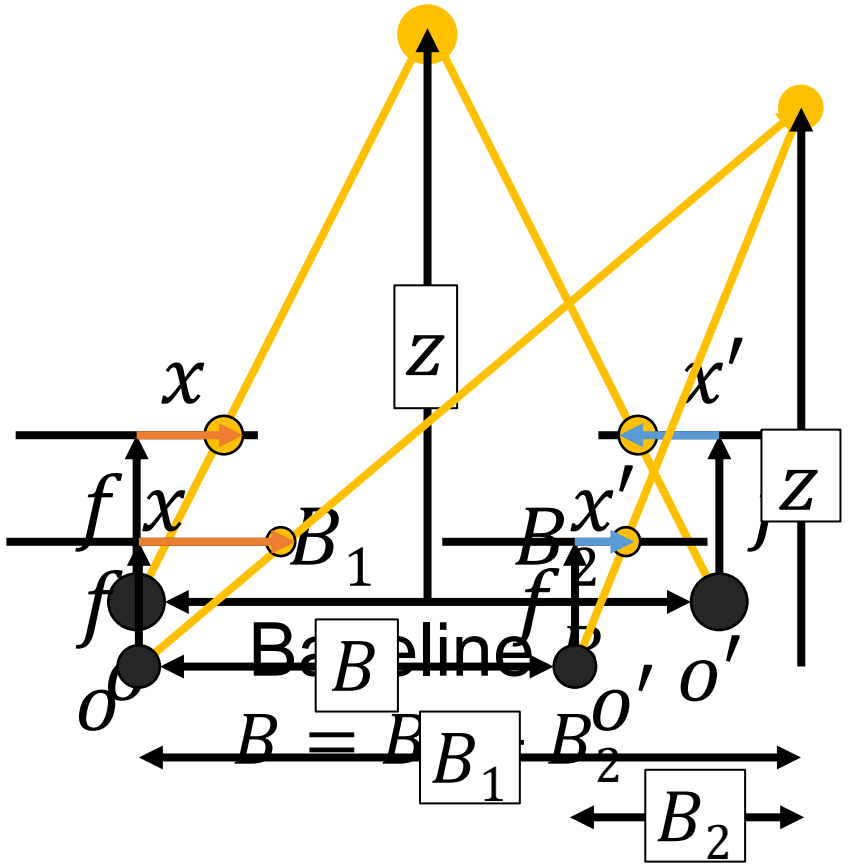
- From similar triangles, the following two equation holds:

$$\frac{x}{f} = \frac{B_1}{z}; \frac{-x'}{f} = \frac{B_2}{z}; \frac{x - x'}{f} = \frac{B_1 + B_2}{z} = \frac{B}{z}$$

- Denote the disparity as the distance btw two corresponding points in the left and right image of a stereo pair, that is $d = x - x'$, obtain the depth from disparity by:

$$d = x - x' = \frac{fB}{z}$$

$$z = \frac{fB}{d}$$



- How to obtain depth from image? Obtain disparity map!

© Copyright National University of Singapore. All Rights Reserved.

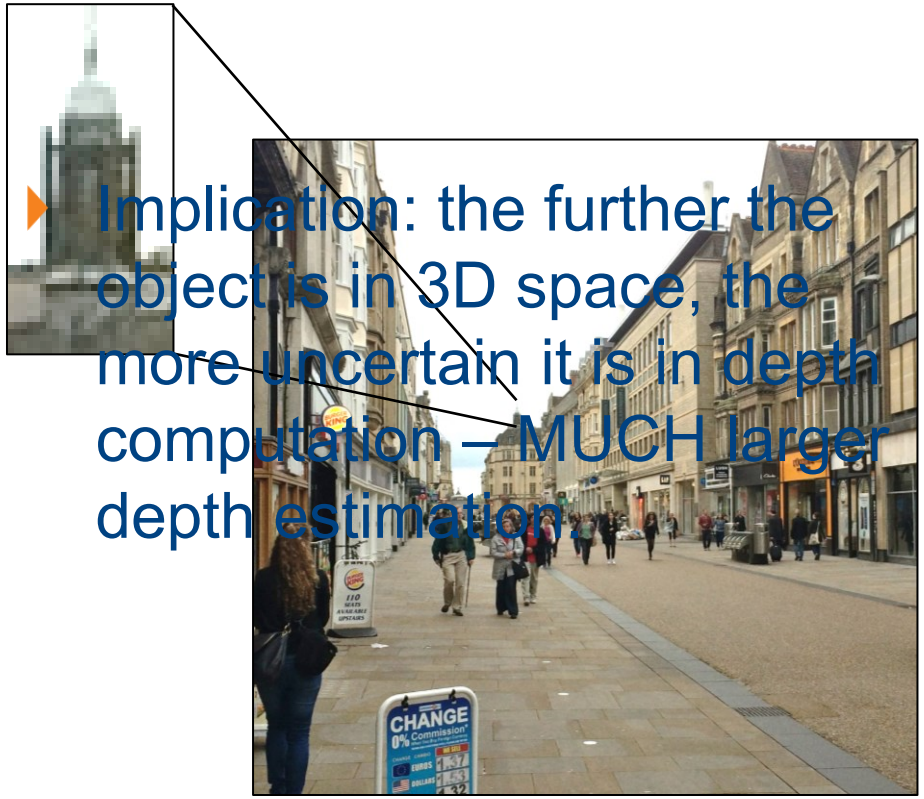
Depth Error from Disparity Error

- Depth estimation – disparity measurements may not be correct due to blurry images.
  - Assume that we may not know the exact disparity d, but the estimated depth Z, focal length f, baseline length B and disparity error Δd are known.
  - To estimate ΔZ, we start from the depth estimation:

$$Z = \frac{fB}{d}; \Delta Z = \frac{fB}{d} - \frac{fB}{d + \Delta d} = fB\frac{\Delta d}{d(d + \Delta d)} = \frac{Z^2\,\Delta d}{fB + Z\Delta d}$$

  - To simplify, we use first order Taylor series estimation: $g(x)|_{x \to a} = g(a) + g'(x)|_{x=a}(x - a)$
- Depth error estimated from the disparity error:

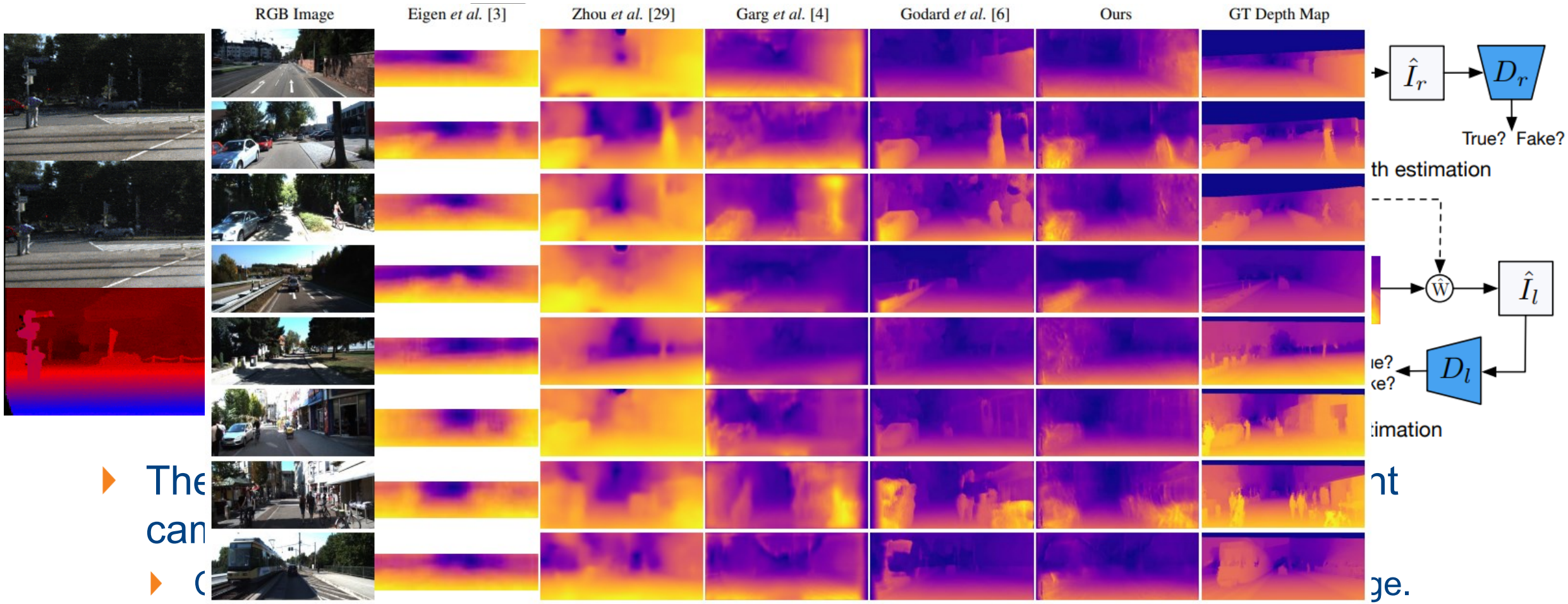
$$\Delta Z = \frac{Z^2}{fB}\Delta d$$



- Implication: the further the object is in 3D space, the more uncertain it is in depth computation – MUCH larger depth estimation.

© Copyright National University of Singapore. All Rights Reserved.

Learning-based Depth Est.: From Disparity Maps

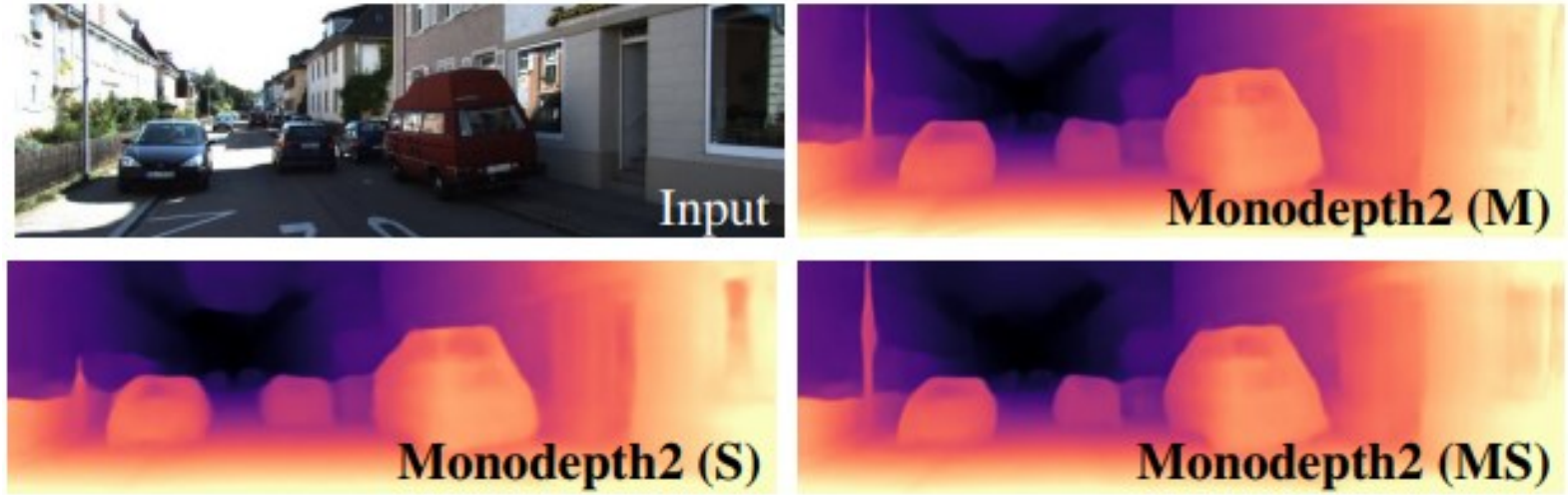


- The nt can
  - C ge.

© Copyright National University of Singapore. All Rights Reserved.

Binocular or Monocular?

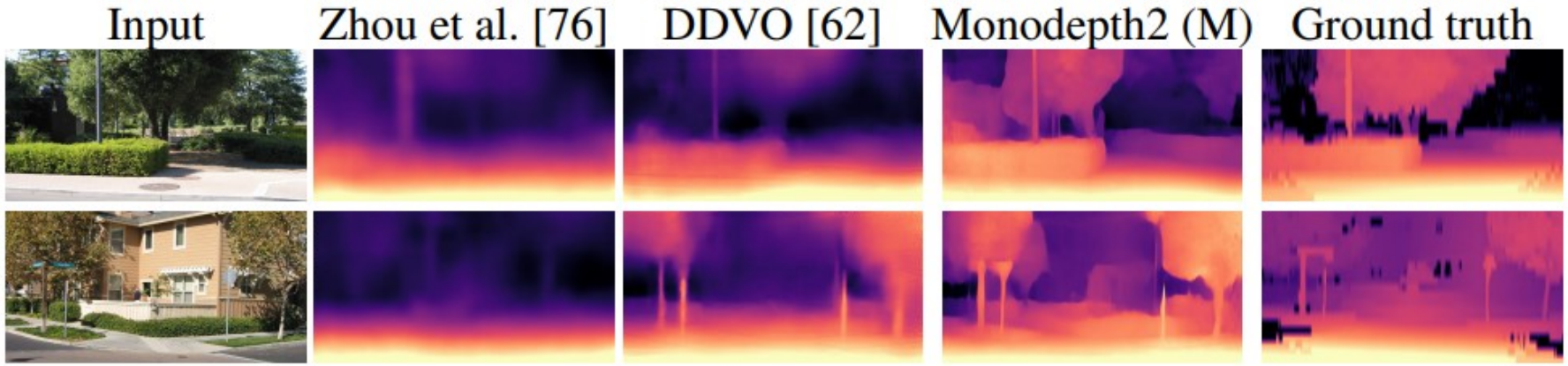
- Humans normally perceive depth information with two eyes: Binocular
- Can humans perceive depth information with one eye : Monocular?
- Actually, in most cases it is possible, and you can’t observe much difference to that of using two eyes!
- How? Human observe “videos”, not “still images”.



© Copyright National University of Singapore. All Rights Reserved.

Monocular Depth Estimation

- Bino … rt
  mon … em!
- Diffe
  bino
- Not a … ded!



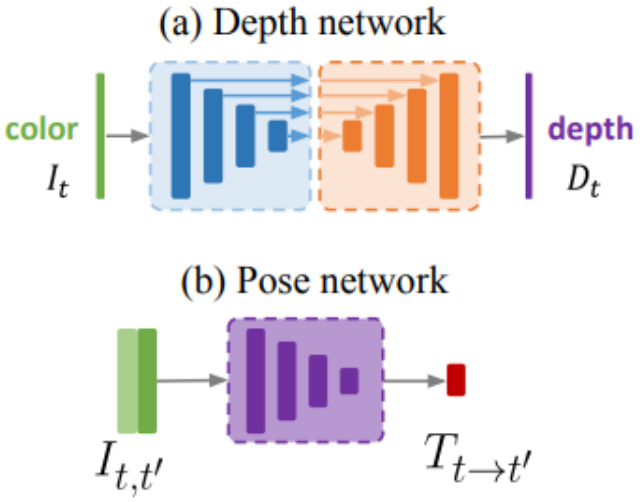
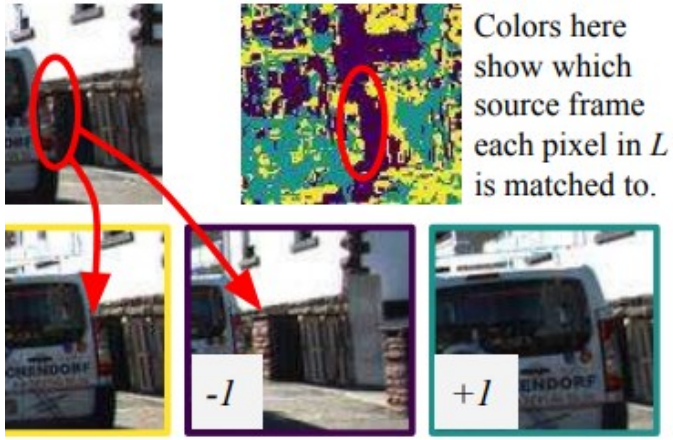


Figure 8. **Failure cases. Top:** Our self-supervised loss fails to learn good depths for distorted, reflective and color-saturated regions. **Bottom:** We can fail to accurately delineate objects where boundaries are ambiguous (left) or shapes are intricate (right).



Colors here show which source frame each pixel in L is matched to.

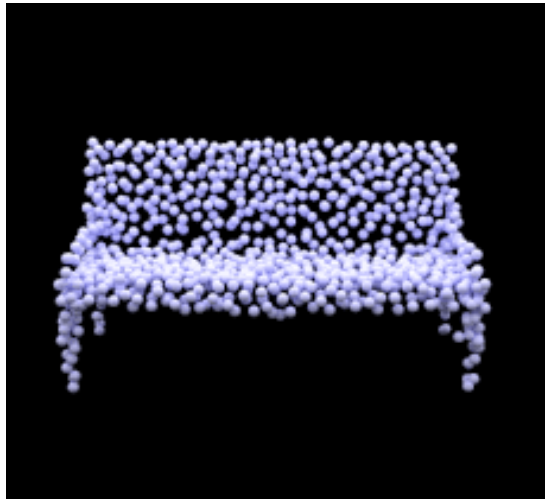
© Copyright National University of Singapore. All Rights Reserved.

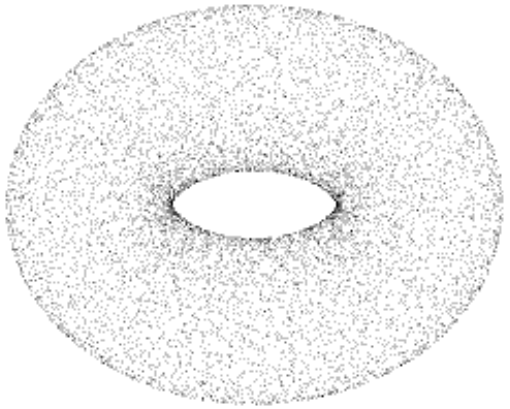
Point Clouds

- A more direct way to understand 3D points.
- A Point Cloud is an unordered set of 3D points in a frame of reference (Cartesian coordinate system) on the surface of objects.

$$\boldsymbol{P} = \{(\boldsymbol{x_i}, \boldsymbol{y_i}, \boldsymbol{z_i}) | \boldsymbol{i} \in \boldsymbol{N}\}$$

- More info can be associated with the cartesian coordinates: e.g., RGB color information, normal to surface, info about meshes (vertices & edges).
- Dense / Sparse, it gives the info of the observed object / environment in 3D space.

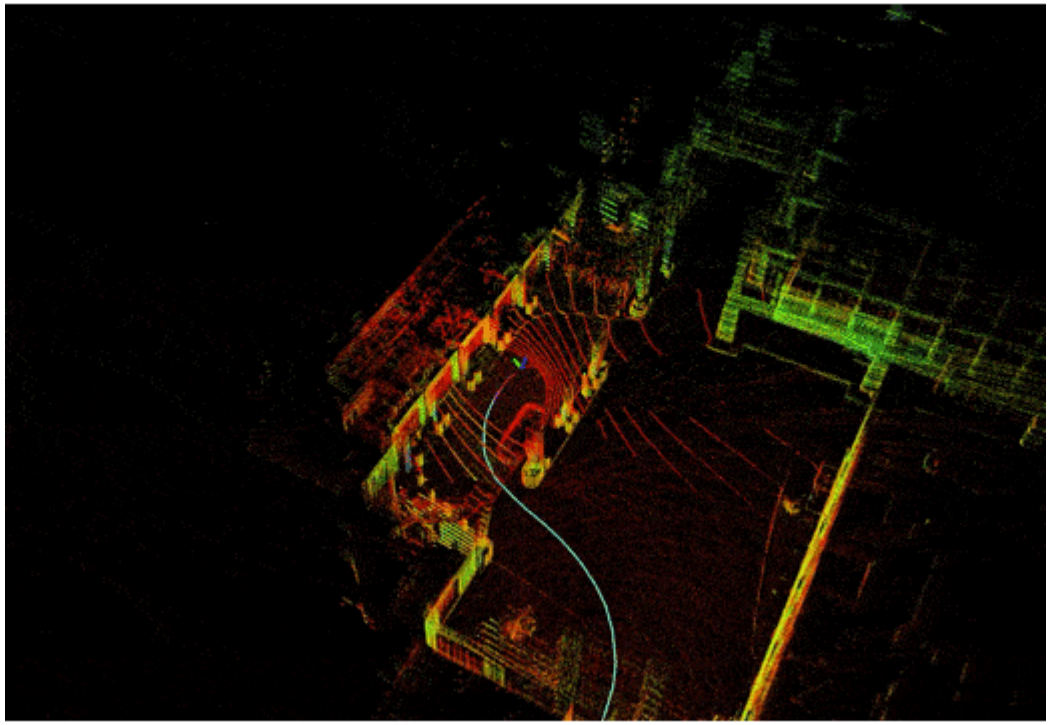




© Copyright National University of Singapore. All Rights Reserved.

Applications of Point Cloud

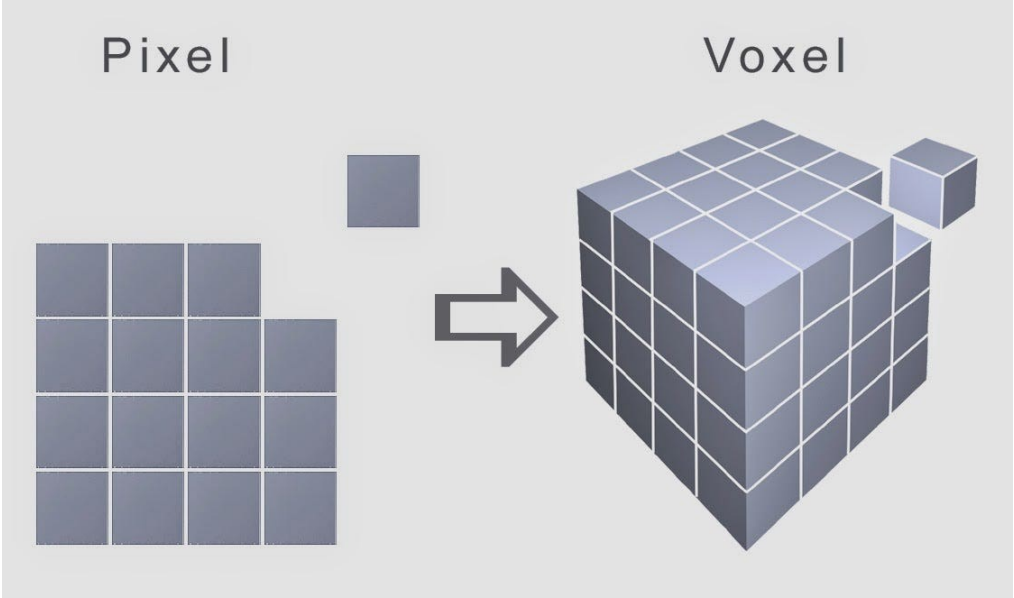
- Applications: Reconstruction and SLAM (Simultaneous Localization And Mapping)



© Copyright National University of Singapore. All Rights Reserved.

Point Clouds and Voxels

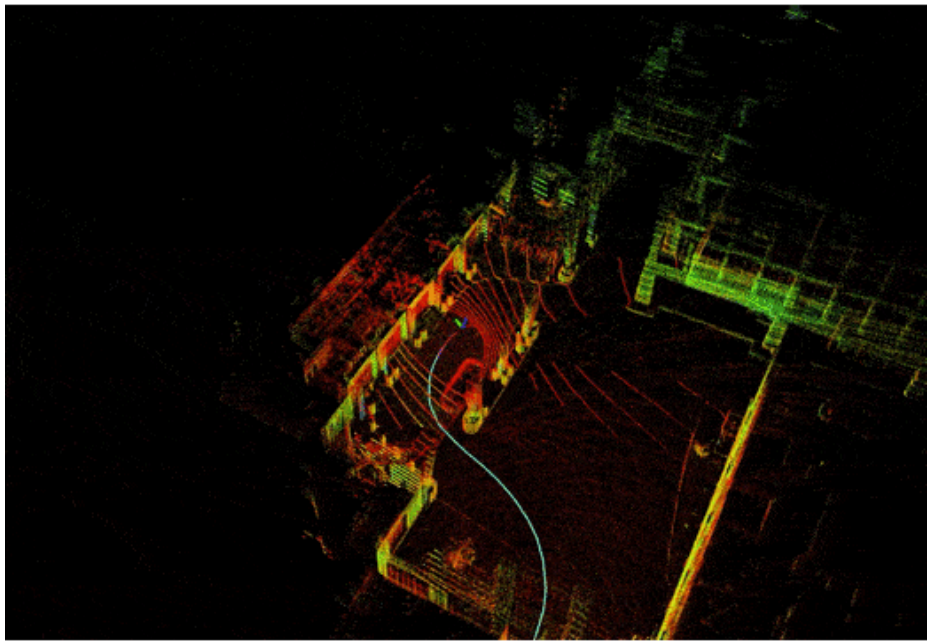
- Point clouds themselves are 3D points, impossible to leverage pixels for description.
- Voxels: a 3D geometric cube, (can be viewed as) the building block of the 3D space.
- Extending pixels in 2D to voxels in 3D.



© Copyright National University of Singapore. All Rights Reserved.

Point Cloud: the Pros and Cons

- Pros:
  - Point cloud models can accurately represent relatively complex objects with a finite number of elements (points).
  - Point cloud models can be the quickest to create -- assuming you have access to 3D scanning technology or CAD conversion software.
- Cons:
  - Point-cloud models lack the precision and cannot create mathematically perfect curves.
  - Point-cloud data does not include information on surfaces, and therefore cannot be used natively for rendering or manufacturing.



© Copyright National University of Singapore. All Rights Reserved.

DL for Point Cloud: PointNet to Pointformer

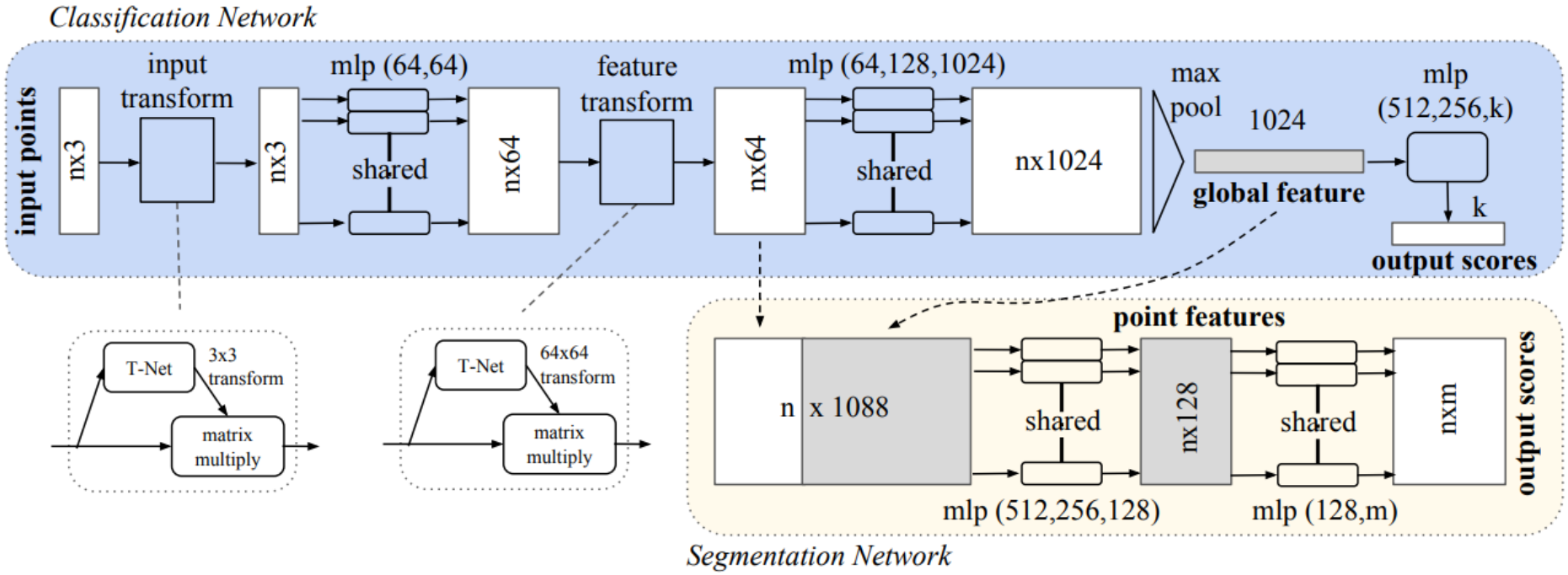


Figure 2. **PointNet Architecture.** The classification network takes n points as input, applies input and feature transformations, and then aggregates point features by max pooling. The output is classification scores for k classes. The segmentation network is an extension to the classification net. It concatenates global and local features and outputs per point scores. "mlp" stands for multi-layer perceptron, numbers in bracket are layer sizes. Batchnorm is used for all layers with ReLU. Dropout layers are used for the last mlp in classification net.

© Copyright National University of Singapore. All Rights Reserved.

DL for Point Cloud: PointNet to Pointformer

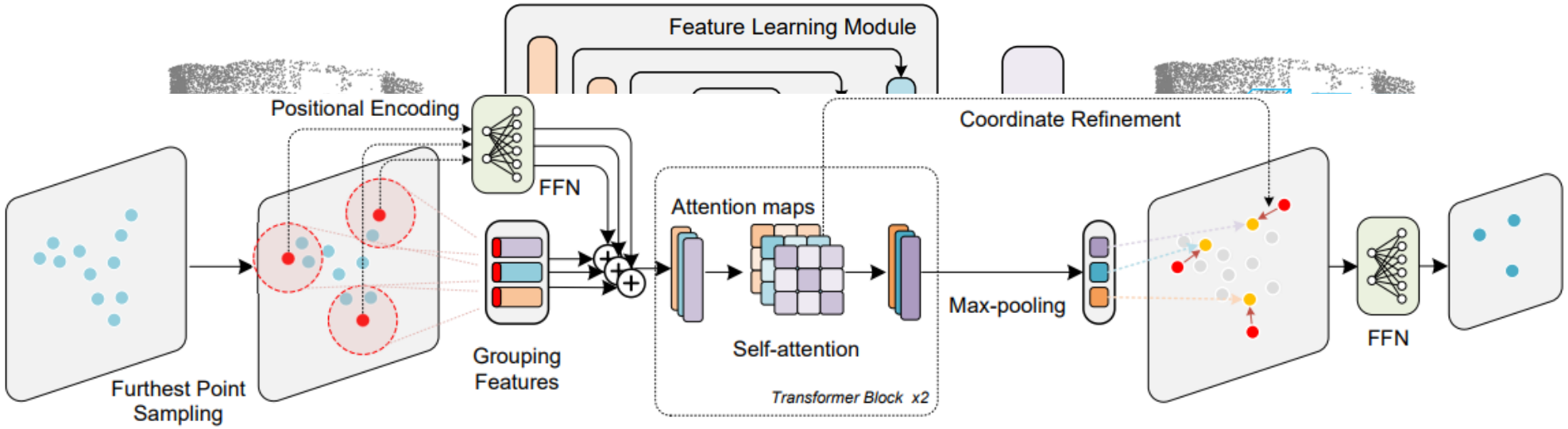
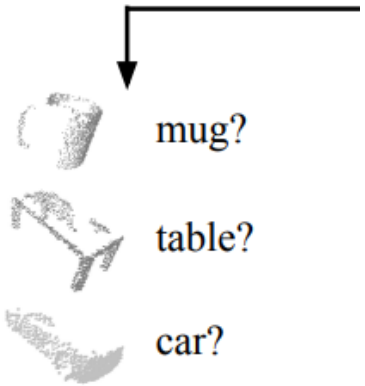


Figure 3. Illustration of the Local Transformer. Input points are first down-sampled by FPS and generate local regions by ball query. Transformer block takes point features and coordinates as input and generate aggregated features for the local region. To further adjust the centroid points, attention maps from the last Transformer layer are adopted for coordinate refinement. As a result, points are pushed closer to the object centers instead of surfaces.

Figure 2. The **Pointformer** backbone for 3D object detection in point clouds. A basic feature learning block consists of three parts: a Local Transformer to model interactions in the local region; a Local-Global Transformer to integrate local features with global information; a Global Transformer to capture context-aware representations at the scene level.

© Copyright National University of Singapore. All Rights Reserved.

DL for Point Cloud: The Tasks



Classification

Figure 1. **Applicatio**
architecture that cons
voxelization or rende
both global and local
and effective approac

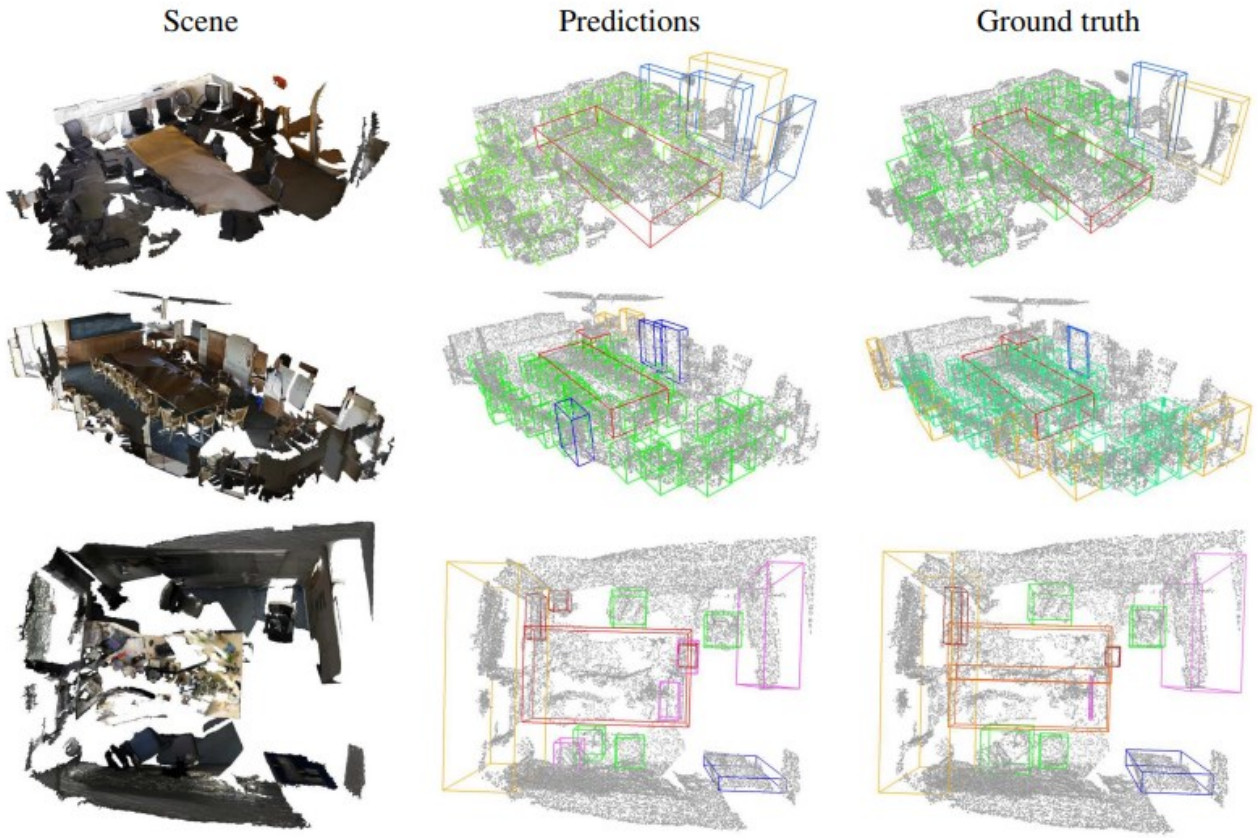


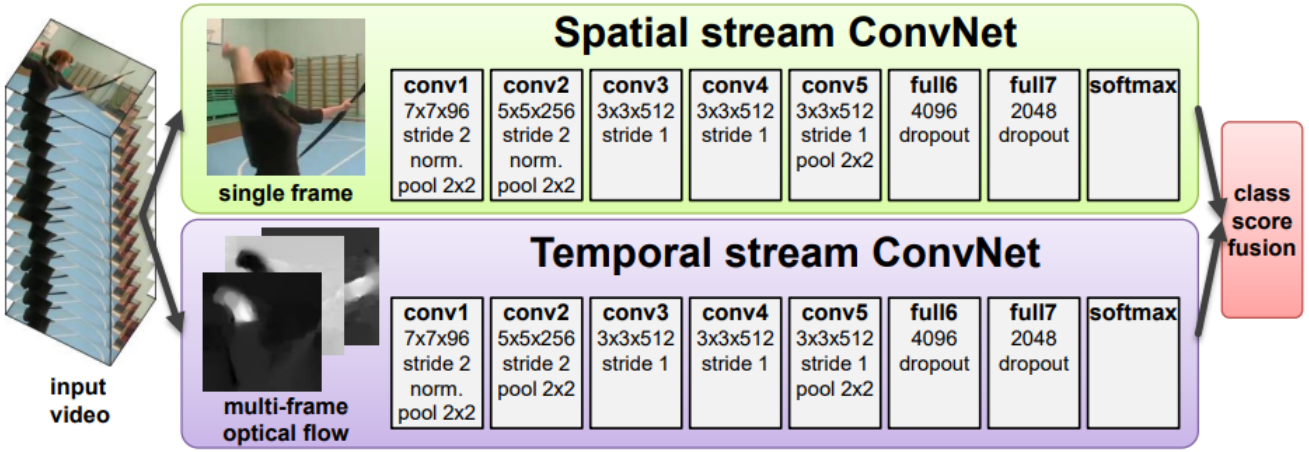
Figure 7. **Qualitative results of 3D object detection on ScanNetV2.** From left to right: Original scene image, our model's prediction, and annotated ground truth boxes. Different object categories are presented with different colors.

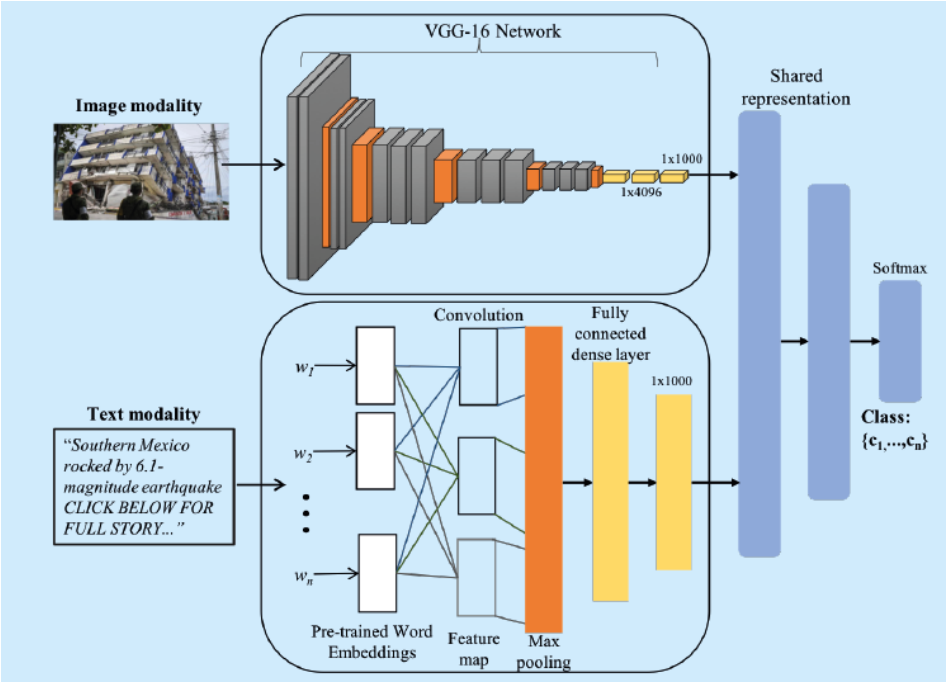
ıtic segmentation. Top
n row is output semantic
ed in the same camera

© Copyright National University of Singapore. All Rights Reserved.

Multi-Modal Learning and 3D Vision

- Information from a single modality can have various constraints (e.g., the pros and cons of point cloud / RGB images).
- One intuitive but also effective way of learning: multi-modal learning!
  - Action Recognition: Optical flow + RGB; 3D: Point cloud + RGB; Text + Image

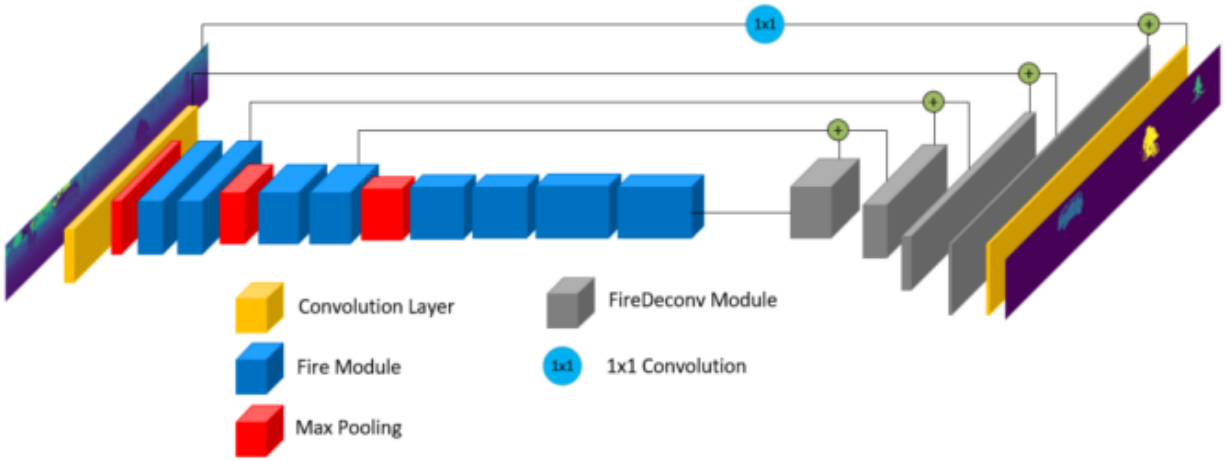




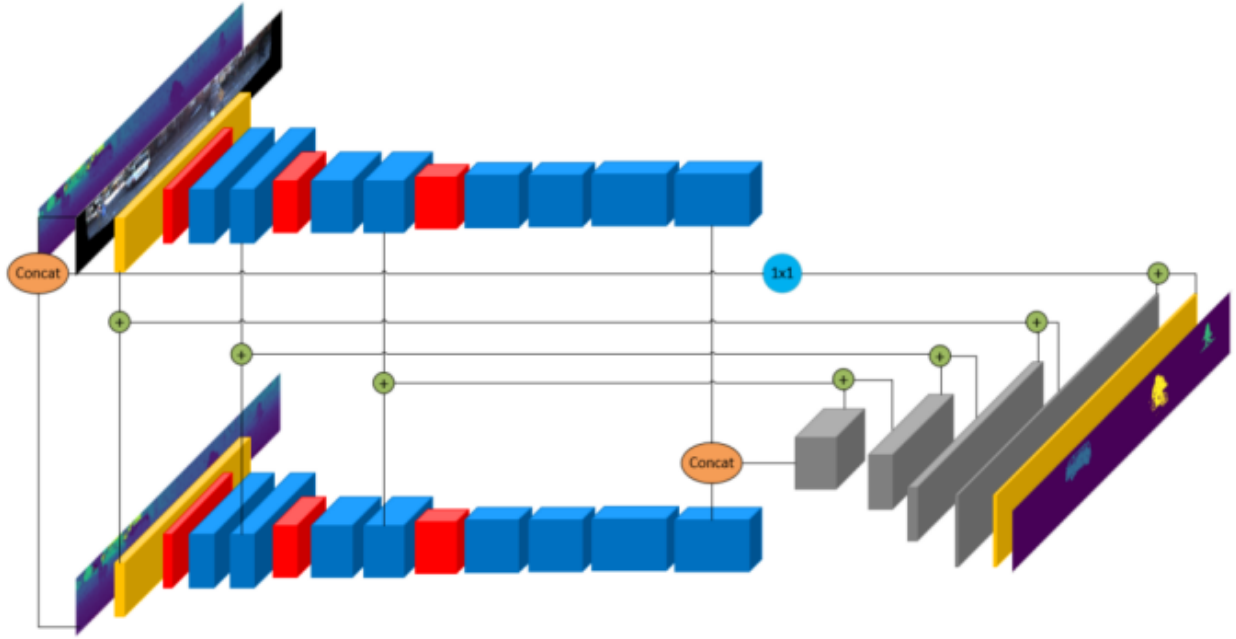
© Copyright National University of Singapore. All Rights Reserved.

Earlier Fusion: Network-Based

- Similar to slow fusion for videos.



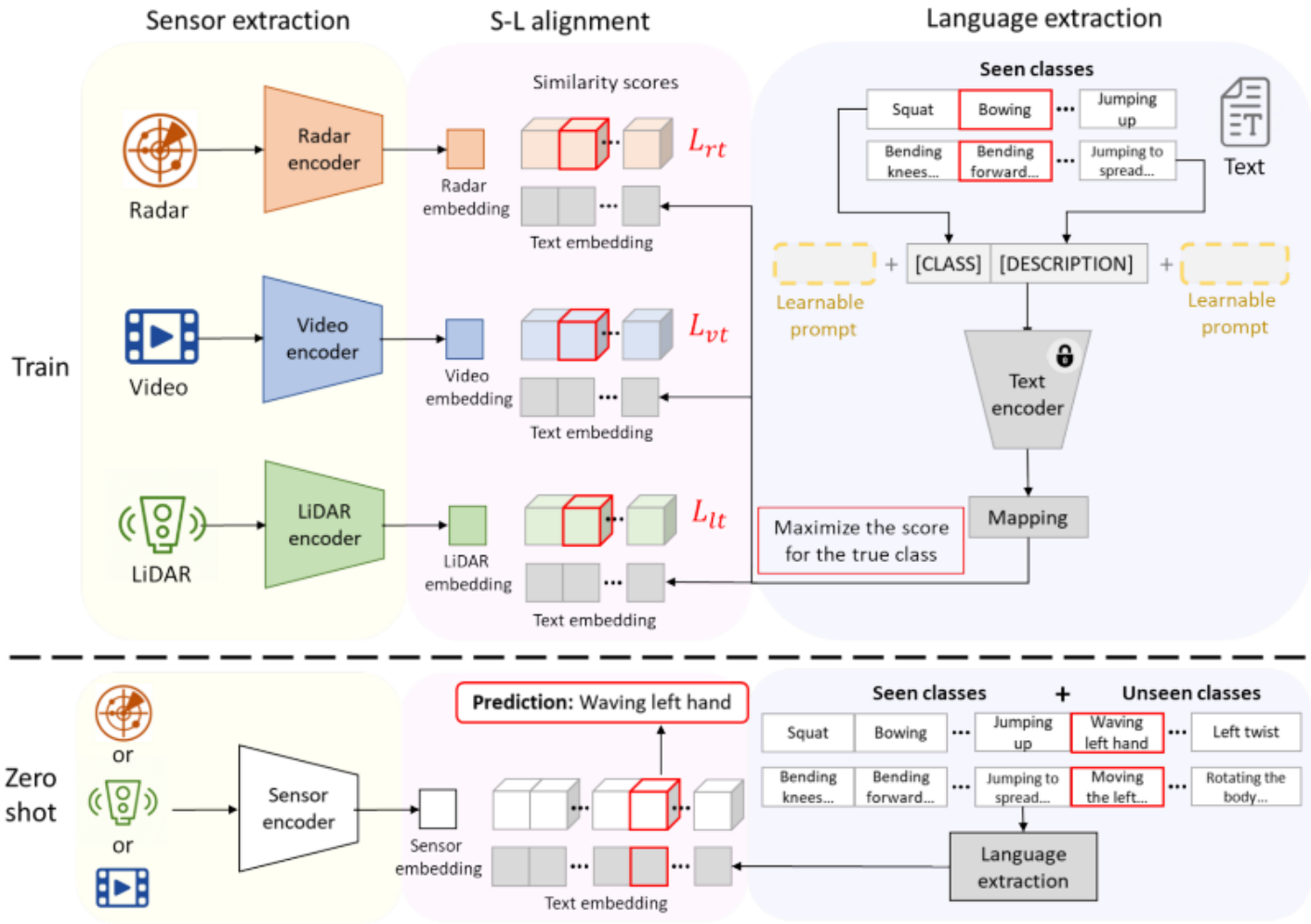
(a) LiDAR baseline architecture based on SqueezeSeg [22].



(b) Proposed RGB+LiDAR mid-fusion architecture

© Copyright National University of Singapore. All Rights Reserved.

Recent Trends: LLXM (LMM) with Visual as Anchor



© Copyright National University of Singapore. All Rights Reserved.