Multi-loss Function to Improve the Text Detection and Segmentation

Problem Statement –

"Slidin' videos": Slide Transition Detection and Title Extraction in Lecture Videos

ITU AI/ML in 5G Challenge 2022

Team: AA_Vision

Dr. Anuj Abraham

Senior Researcher

TII, UAE.

Dr. Shitala Prasad

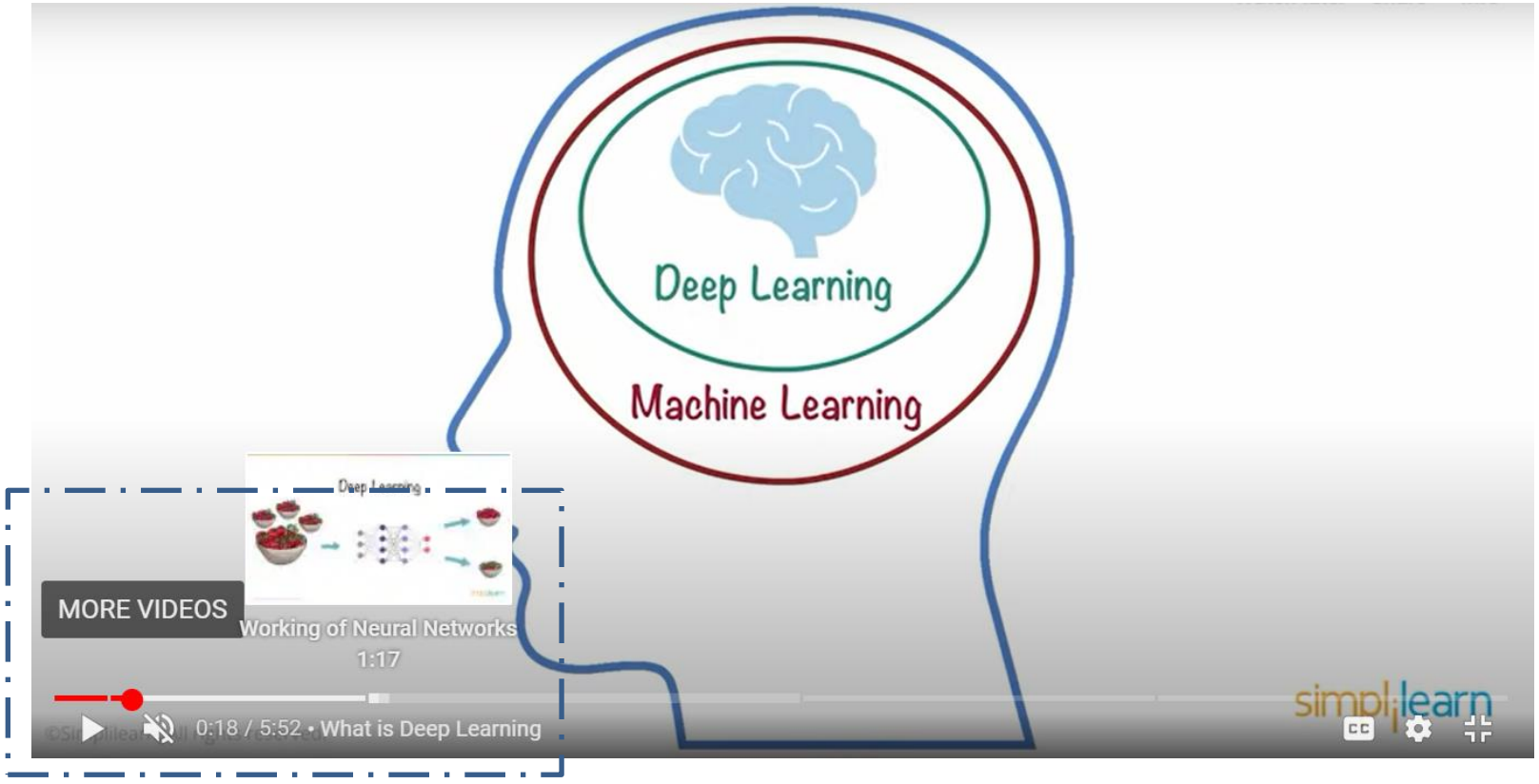
Scientist II

A*STAR, Singapore.

Outline

Introduction

YouTube's "Video Chapter" feature segments a video into sections marked by timestamps so that the user can easily navigate to the part of the video which is of most interest. This can be done by clicking or pressing the chapter marker, or by selecting the timestamp in the video description – (AI-5G Challenge)



Introduction: Problem Definition

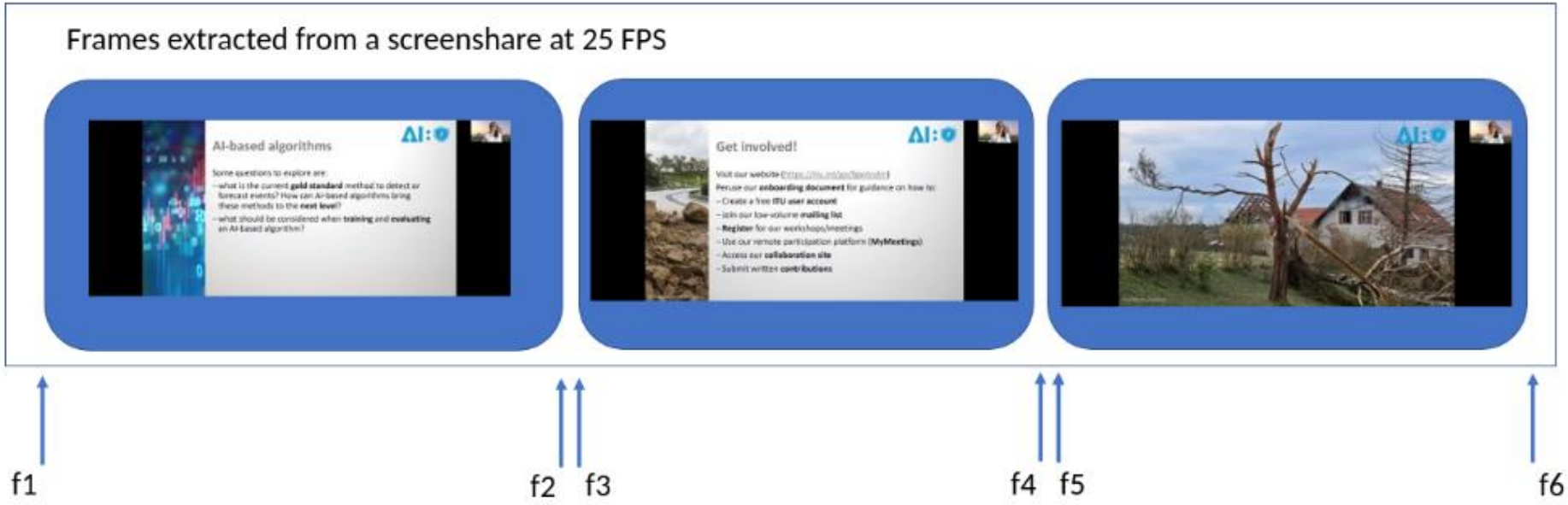
AI model which annotates slide transitions by:

- Identifying starting and ending frames of each slide shown in the video
- Extracting (apparent) titles of each slide
- All videos were recorded at 25FPS

Dataset:

- ❑ Video files covering the presentation from when speaker started screen share right to the moment when it was turned off. Video files vary in duration (from several minutes to several hours) and resolution (from **1600** x **1200** to **3840** x **2160**).
- ❑ A ground truth data set with **2500+** slide transitions showing the starting and ending frame of each slide including (apparent) titles.

Introduction: Problem Detail



frame_start, frame_end, is_slide, title
f1, f2, 1, "AI-based algorithms"
f3, f4, 1, "Get involved!"
f5, f6, 1, "NO_TITLE"

where f1, f2, f3, f4, f5, f6 – are frame numbers,
f2 is adjacent to f3, f4 is adjacent to f5

“groundtruth.csv”

A1 starting_frame1

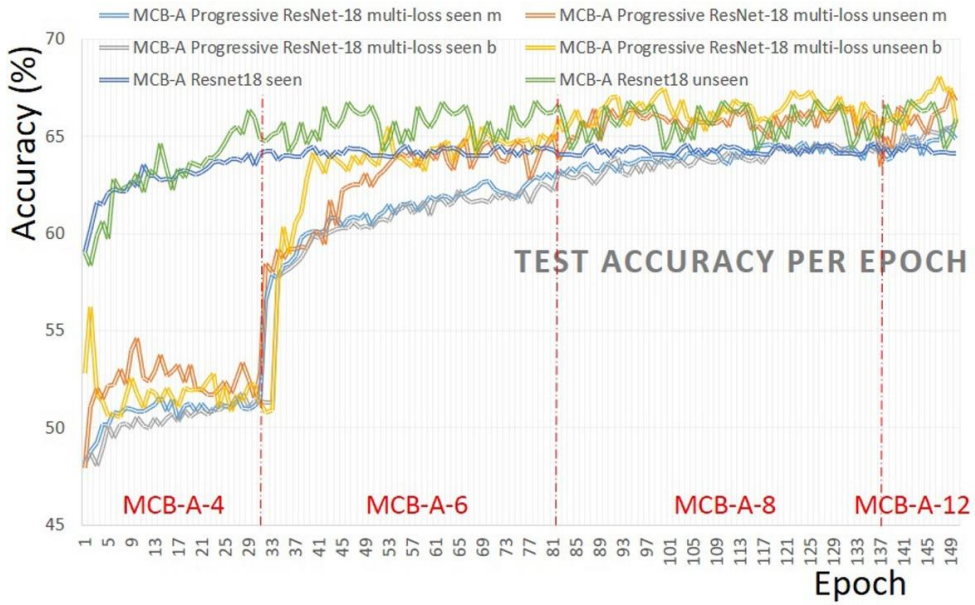
	A	B	C	D	E	F	G	H	I	J	K	L
1	starting_frame1	starting_frame2	ending_frame1	ending_frame2	title1	title2	title3	title4	bonus_title1	bonus_title2	is_hybrid	
2	25	25	97	97	CAN TECHNOLOGY SCALE TO FEED THE WORLD ?						0	
3	98	102	949	977	CAN TECHNOLOGY SCALE TO FEED THE WORLD ?						0	
4	950	978	1332	1345	HPE-STUDENTS-FARMERS						0	
5	1333	1346	1765	1782	HPE-STUDENTS-FARMERS						0	
6	1766	1783	2721	2730	HPE-STUDENTS-FARMERS						0	
7	2722	2731	4268	4279	RESULTS						0	

Introduction: Motivation

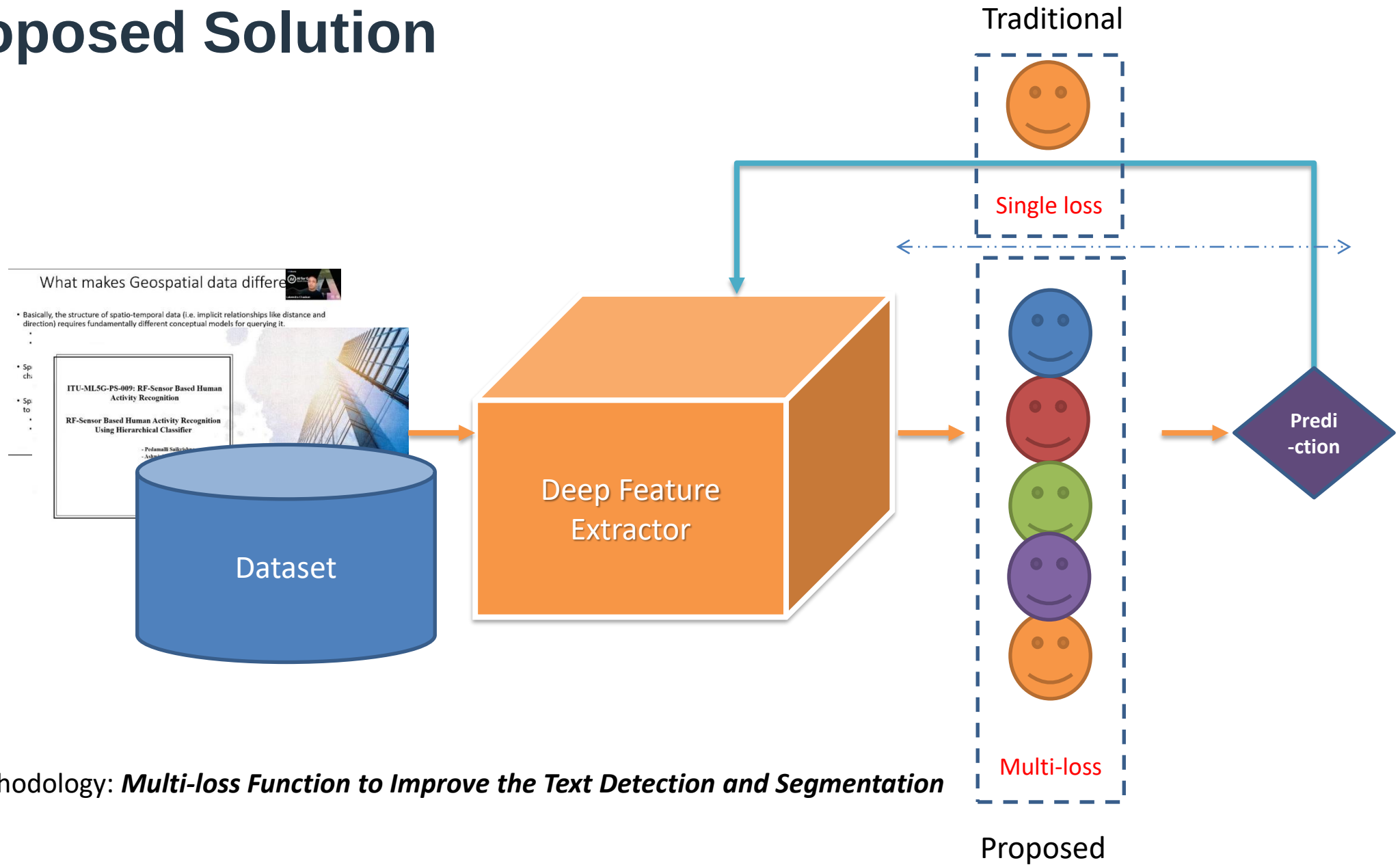
YouTube introduced a multi-loss view invariant stochastic prototype embedding to minimize and improve the recognition accuracy of novel objects at different viewpoints by using a progressive multi-view learning approach.

- Prasad, S., Li, Y., Lin, D., Dong, S. and Nwe, M.T.L., 2021. A Progressive Multi-View Learning Approach for Multi-Loss Optimization in 3D Object Recognition. *IEEE Signal Processing Letters*, 29, pp.707-711.

- Multi-loss in Detection and Segmentation



Proposed Solution

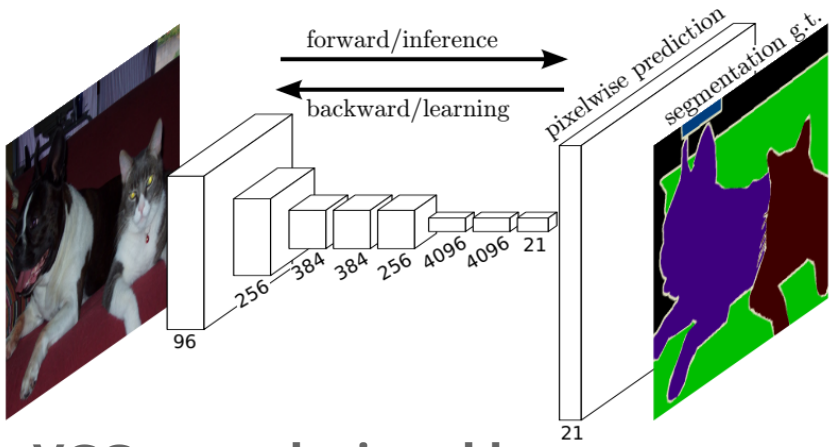


Methodology: ***Multi-loss Function to Improve the Text Detection and Segmentation***

Proposed Solution

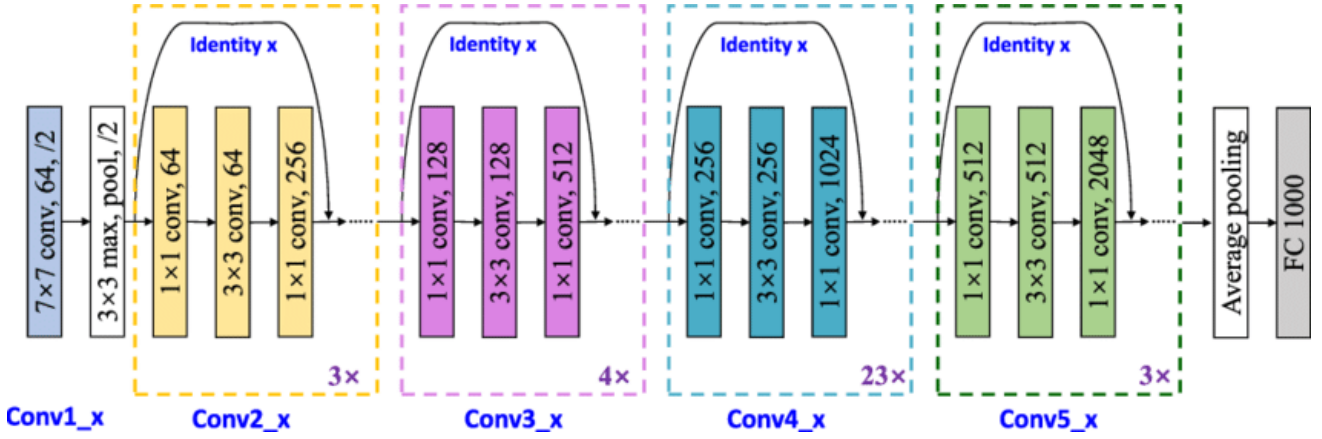
Advantages :

- ❑ Model used **ResNet101** (createDeepLabv3)
- ❑ No architectural changes
- ❑ Negligible model computation cost (equivalent to the original)
- ❑ Multi-loss training strategy converges the network much faster
- ❑ Gives a significant boost in the performance by ~5%



VGG convolutional layers

ResNet convolutional layers



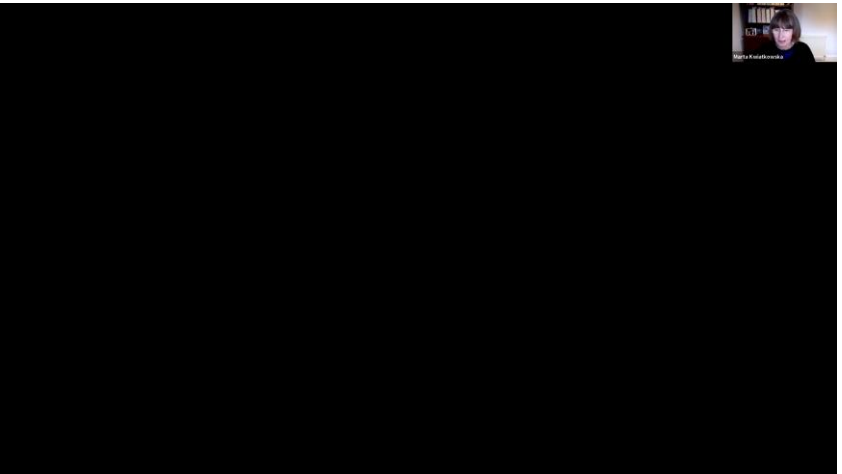
https://github.com/msminhas93/DeepLabv3FineTuning/blob/bcdc3dfc79a5b75bc30c52b32315661c0a4da17e/model.py#L6

Proposed Solution: Dataset Preparation

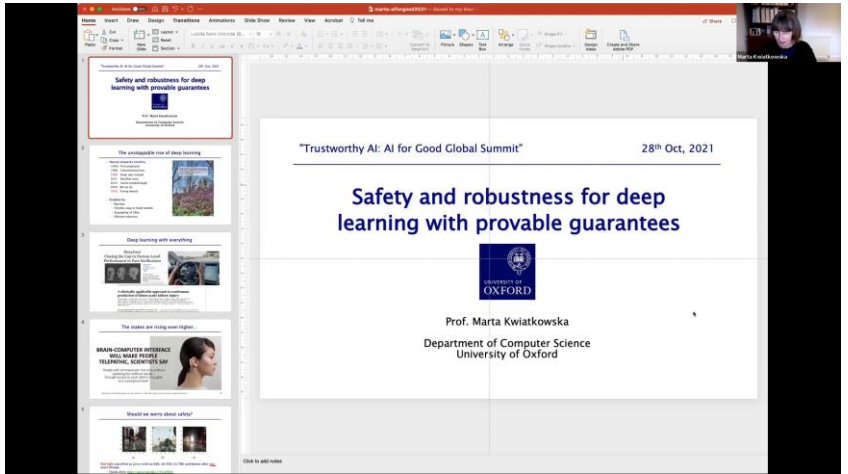
Since this challenge has video as the input source, we needed to extract frames from these video files for feeding deep model. Therefore, for dataset preparation we used following steps:

- ❑ The videos frames per second (fps) is 25
- ❑ Use video files to extract images, 25 frames per second video
- ❑ That is, if the video is of X length, the total number of frames will be X(minute)*60(second)*25(fps)
- ❑ Once the images are extracted, they are categorized into training and validation sets
- ❑ There are three types of images: no title, same title and new title slides

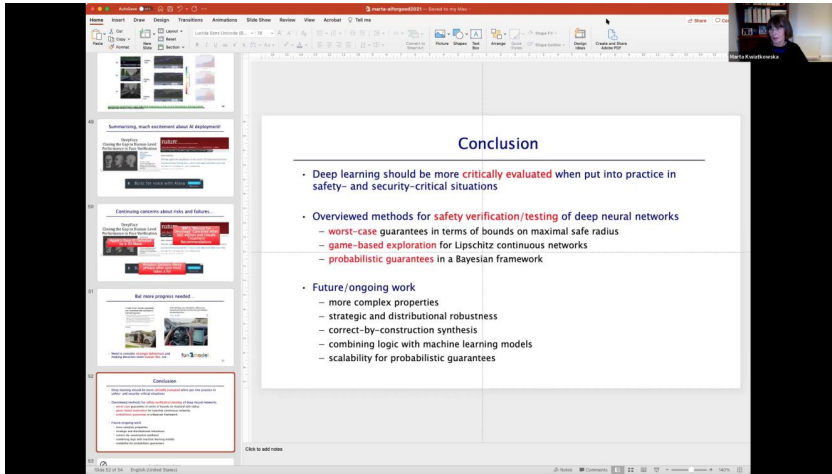
No Title Slide



Title Slide



Title Slide



Proposed Solution: Dataset Preparation

Creating training and validation sets

- ❑ From the ground truth CSV file, we extracted the number of frames with the same titles by using starting and ending frame numbers
- ❑ Split the computed frame number by 80%-20%
- ❑ Using the frame number, split the dataset

groundtruth.csv

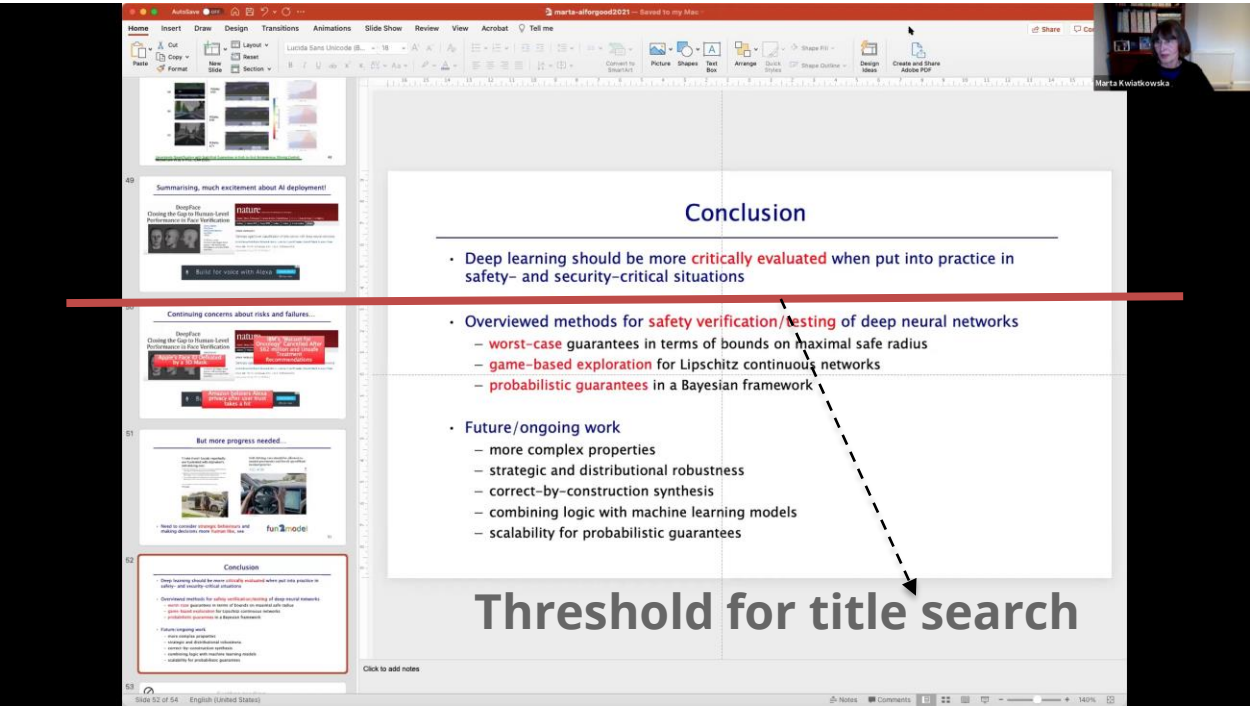
A	B	C	D	E	F	
starting_frame1	starting_frame2	ending_frame1	ending_frame2	title1	title2	title3
35	35	219	219	Safety and robustness for deep learning with pro		
220	220	954	954	Safety and robustness for deep learning with pro		
955	955	3217	3217	The unstoppable rise of deep learning		
3218	3218	3258	3258	Should we worry about safety?		
3259	3259	3273	3273	The stakes are rising even higher...		
3274	3274	3288	3288	Deep learning with everything		
3289	3289	3314	3314	The unstoppable rise of deep learning		
3315	3315	3354	3354	Deep learning with everything		

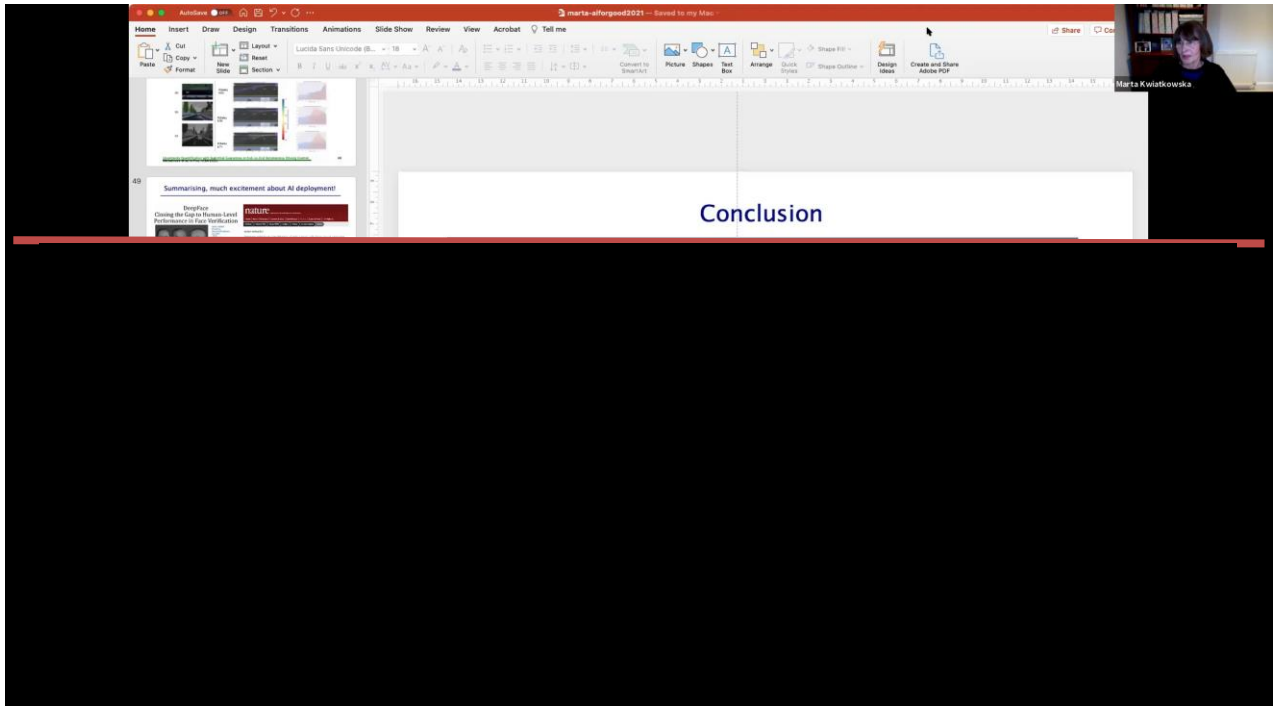
```
train:  147  and validation:  37 out of  184  total frames
train:  587  and validation:  147 out of  734  total frames
train:  1810  and validation:  452 out of  2262  total frames
train:  32  and validation:  8 out of  40  total frames
train:  11  and validation:  3 out of  14  total frames
train:  11  and validation:  3 out of  14  total frames
train:  20  and validation:  5 out of  25  total frames
train:  31  and validation:  8 out of  39  total frames
train:  46  and validation:  12 out of  58  total frames
train:  84  and validation:  21 out of  105  total frames
train:  132  and validation:  33 out of  165  total frames
train:  26  and validation:  6 out of  32  total frames
train:  156  and validation:  39 out of  195  total frames
train:  45  and validation:  11 out of  56  total frames
train:  22  and validation:  6 out of  28  total frames
train:  254  and validation:  64 out of  318  total frames
train:  215  and validation:  54 out of  269  total frames
train:  105  and validation:  26 out of  131  total frames
train:  46  and validation:  12 out of  58  total frames
train:  459  and validation:  115 out of  574  total frames
train:  516  and validation:  129 out of  645  total frames
```

Proposed Solution: Dataset Preparation

Creating training and validation sets

- ❑ From the ground truth CSV file, we extracted the number of frames with the same titles by using starting and ending frame numbers
- ❑ Split the computed frame number by 80%-20%
- ❑ Using the frame number, split the dataset





Results: Comparison

Metrics	Baseline	Proposed Multi-loss (4)
Epochs	10	10
Train Loss	0.00279	0.00192
Train F1	0.853	0.911
Train AUROC	0.993	0.990
Test Loss	0.0241	0.0200
Test F1	0.764	**0.812**
Test AUROC	0.925	**0.949**

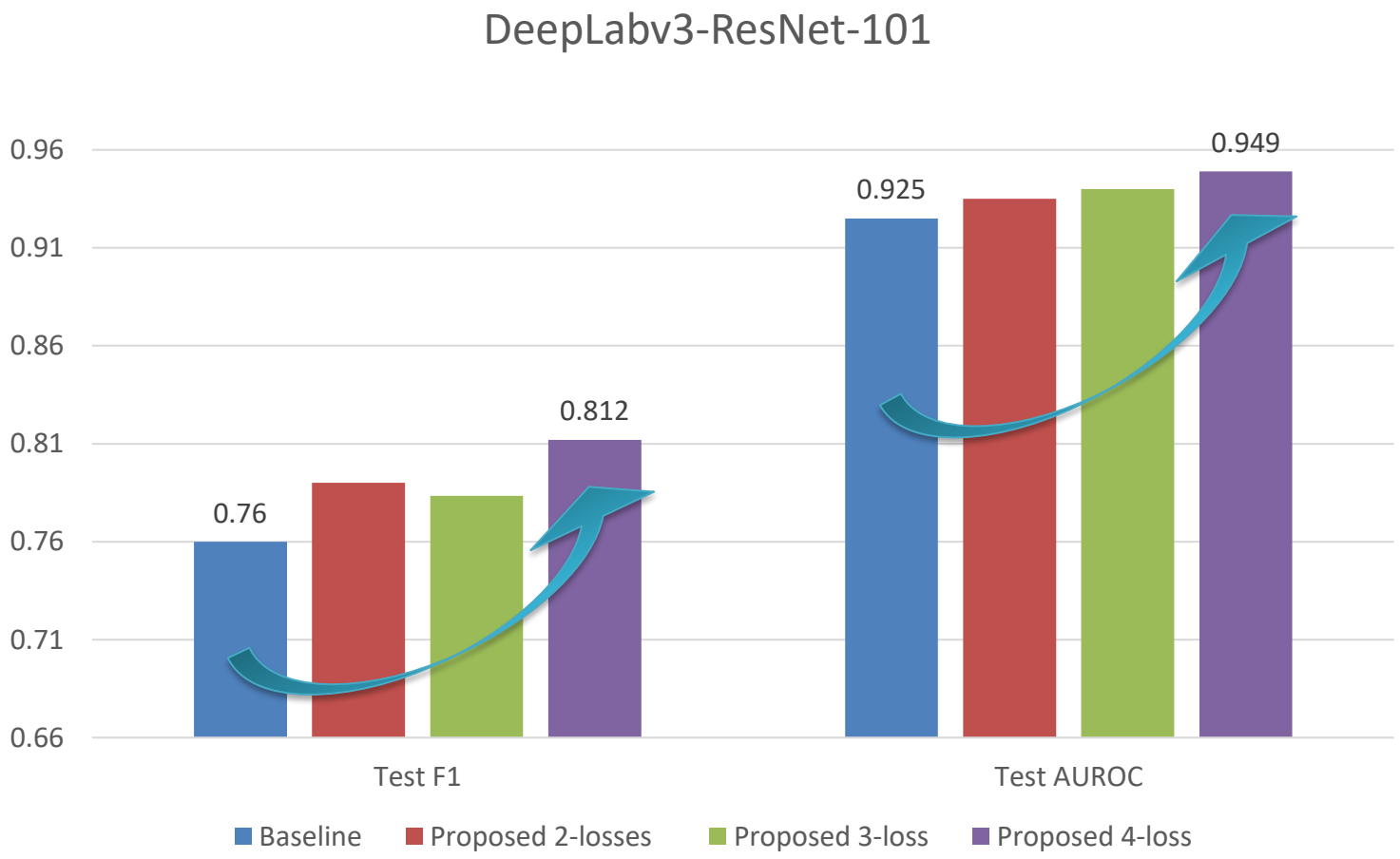
Average of three runs

- We used MAE, MSE, MSLE and variations of Huber loss functions for our experiments

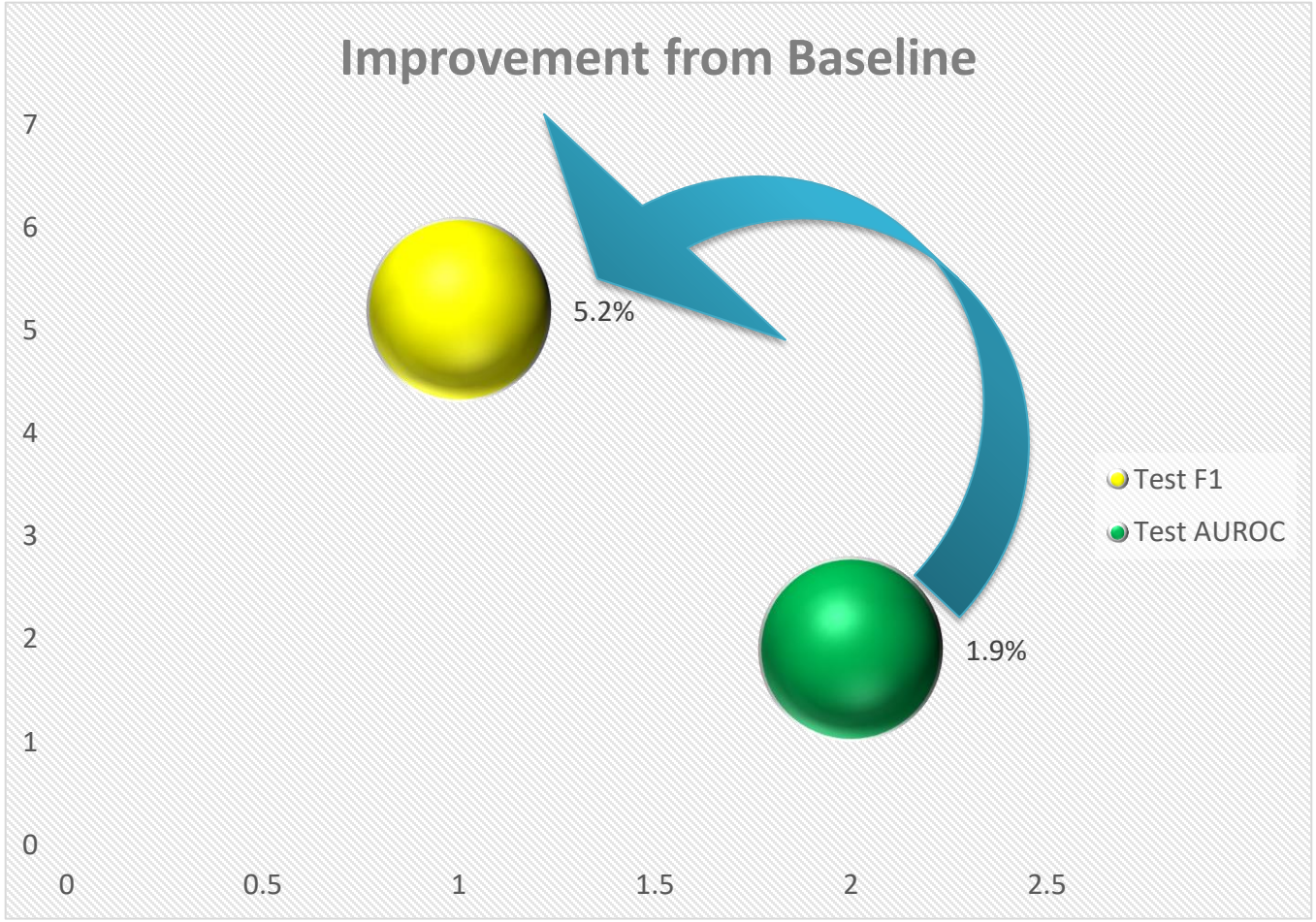
Evaluations: $F_1 = 2 * \frac{precision * recall}{precision + recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}(\mathrm{FP} + \mathrm{FN})}$

https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

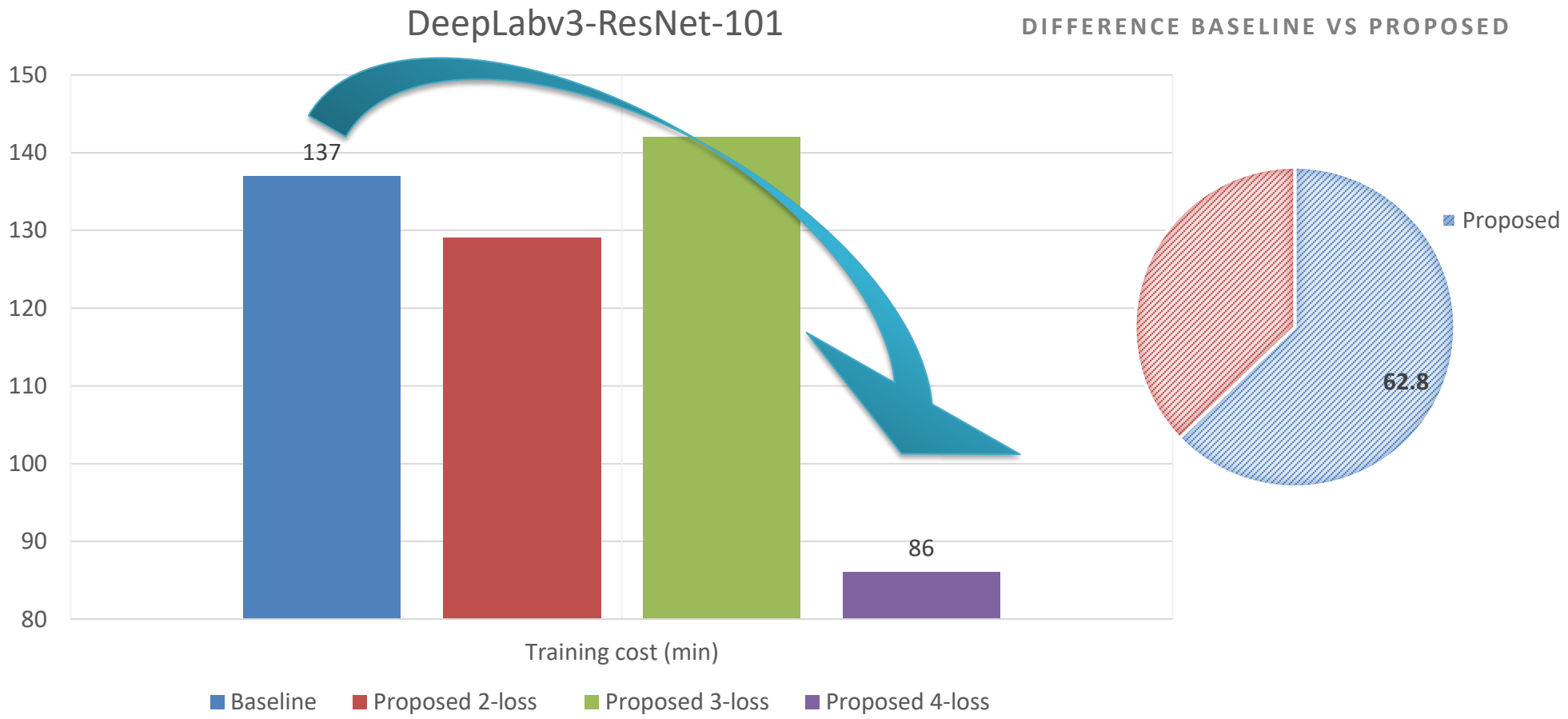
Results: Ablations Study



Results: Ablations Study



Results: Ablations Study (Computations)



Conclusion

In deep digital world,

- Not very important to improve the network architectural to improve the model performance
- Training strategy is important to optimize the learning
- Multi-loss training strategy converges the network
- Significant boost in the performance by simply involving several loss functions for same task
- Gradient calculation is optimized for detection and segmentation

Technology Innovation Institute

Thank you for attention