A Journey Through Video Research at AWS

Joseph Tighe

AWS Rekognition, Senior Scientist

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Content

- AWS AI Video Research
- Tracking
    - Pose Understanding (CVPR 2020)
    - Tracking (Under submission)
    - Self Supervised Tracking
- Action Understanding
    - Fine-grained (ICCV 2019)
    - Spatio-temporal modeling (Under submission)
    - Zero-shot Learning (CVPR 2020)
    - Unsupervised Training (Under submission)
- Fast Video Training and Inference (On-going)
- What Comes Next

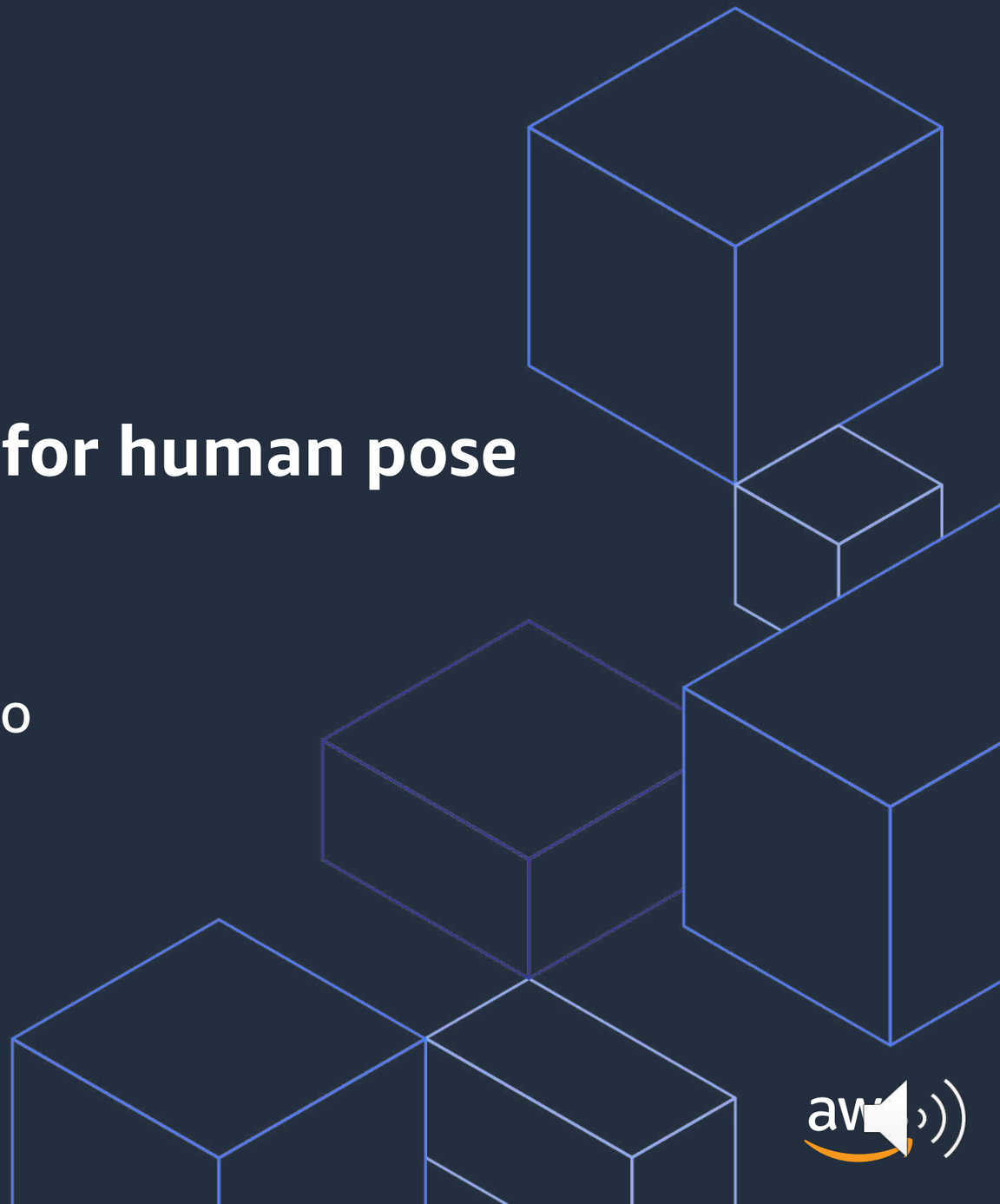
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Tracking to Understand Scene Dynamics

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

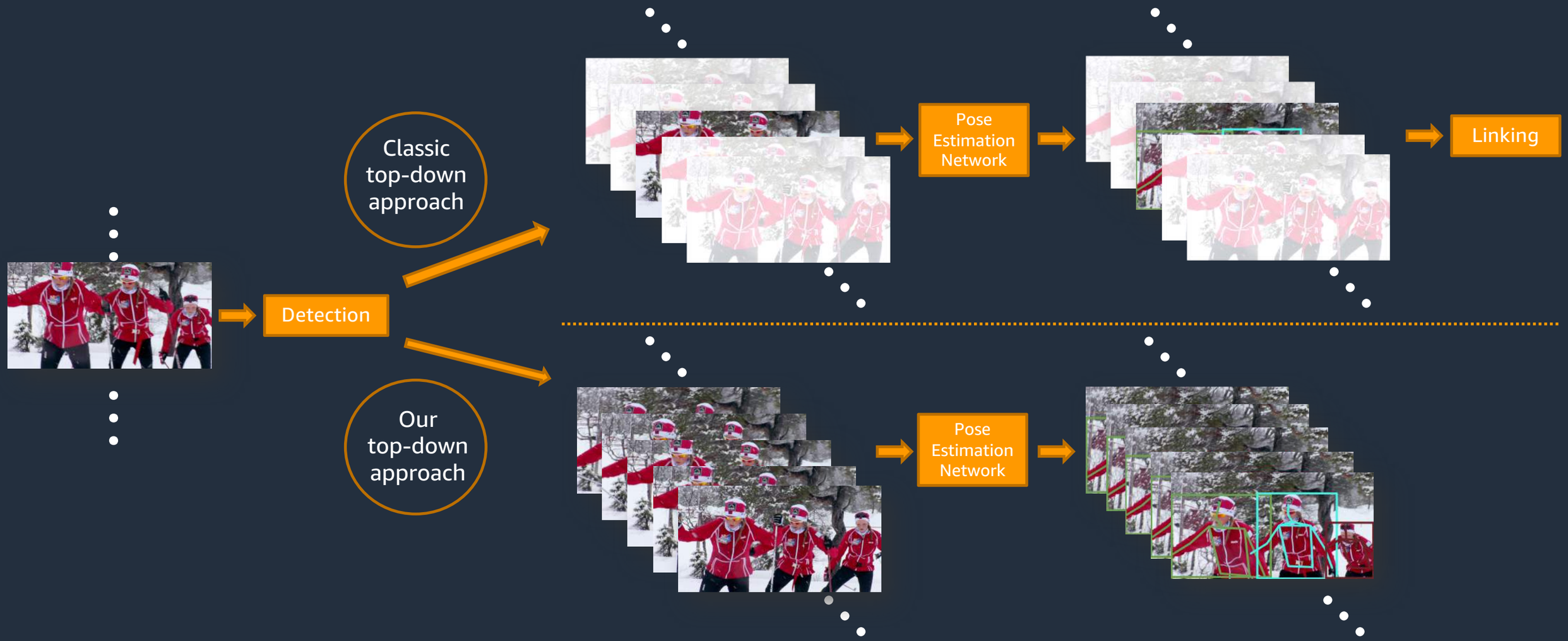
Combining detection and tracking for human pose estimation in videos (CVPR 2020 Oral)

Manchen Wang, Joseph Tighe, Davide Modolo



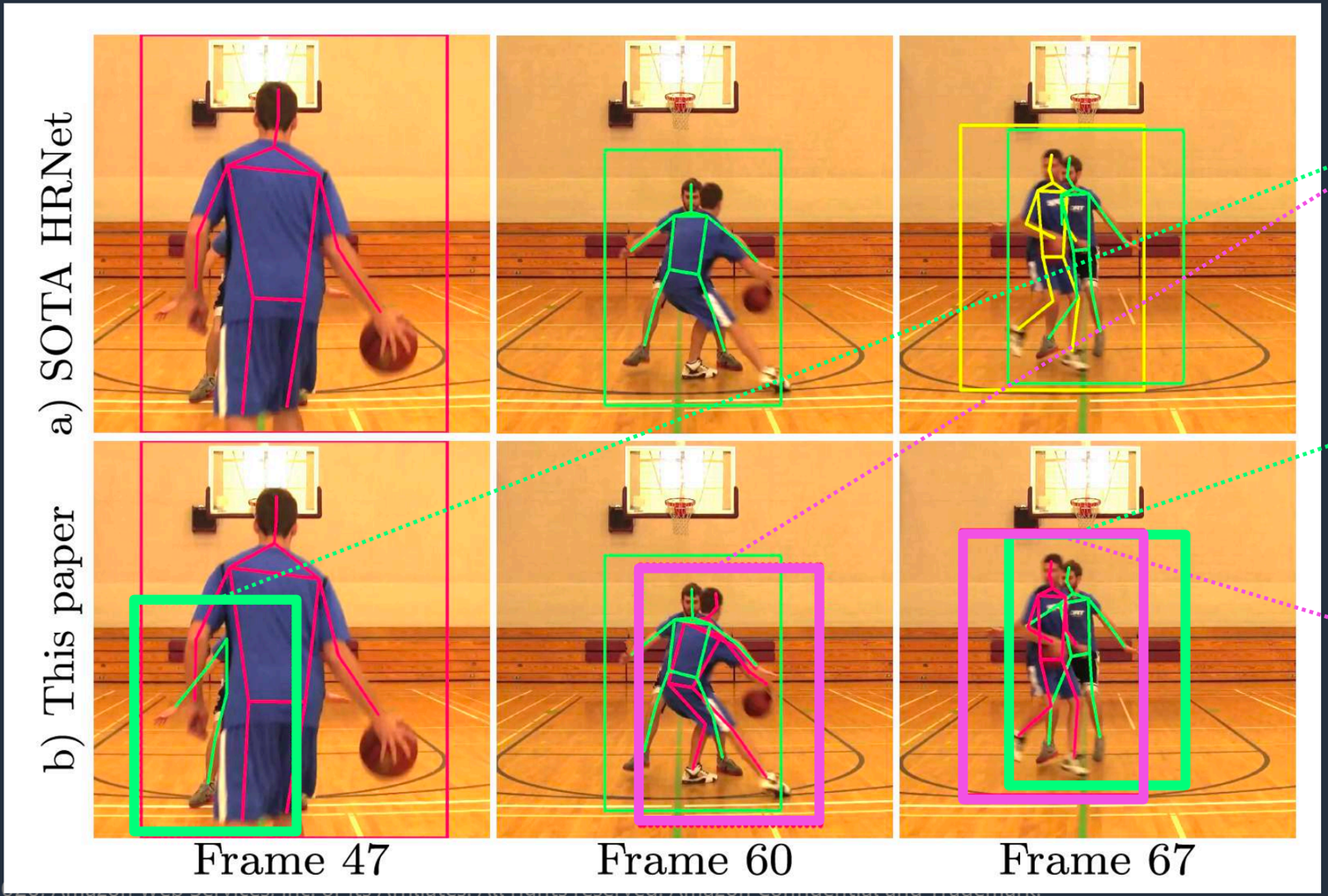
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Introduction



© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Differences and Benefits



1. Compensates for missed person detections

2. Better predictions on highly entangled people

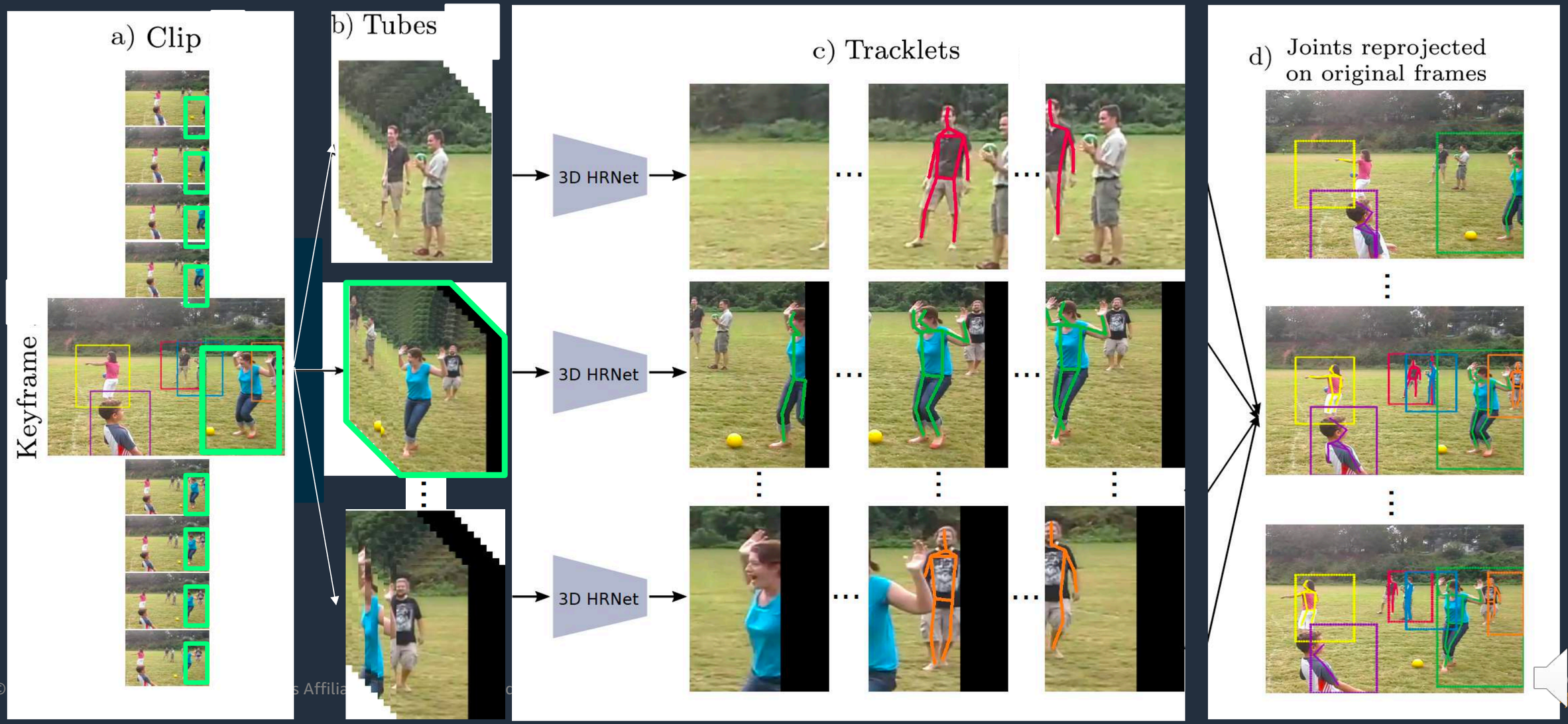
3. Produces more temporally consistent tracklets

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Our approach: 3 components

1. Clip tracking network

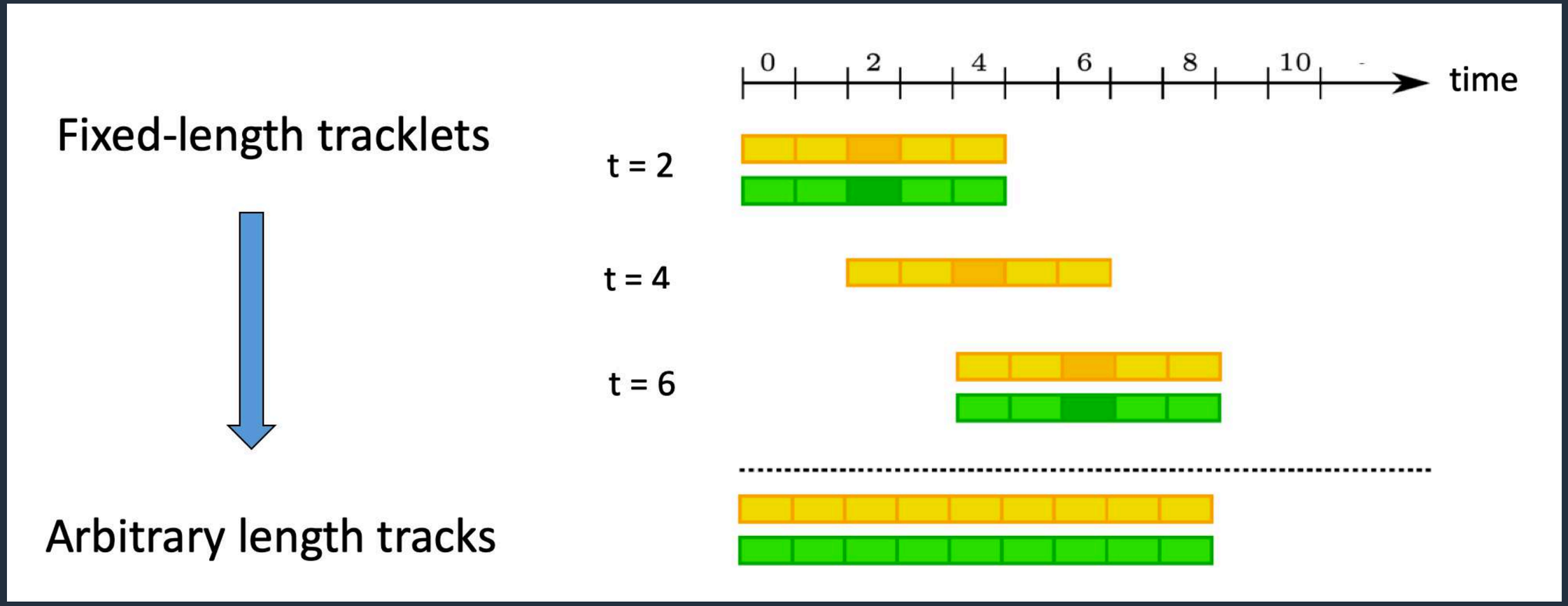
Performs both pose estimation and tracking simultaneously on a short video clip



©

Our approach: 3 components

1. Clip tracking network
2. **Video Tracking Pipeline**

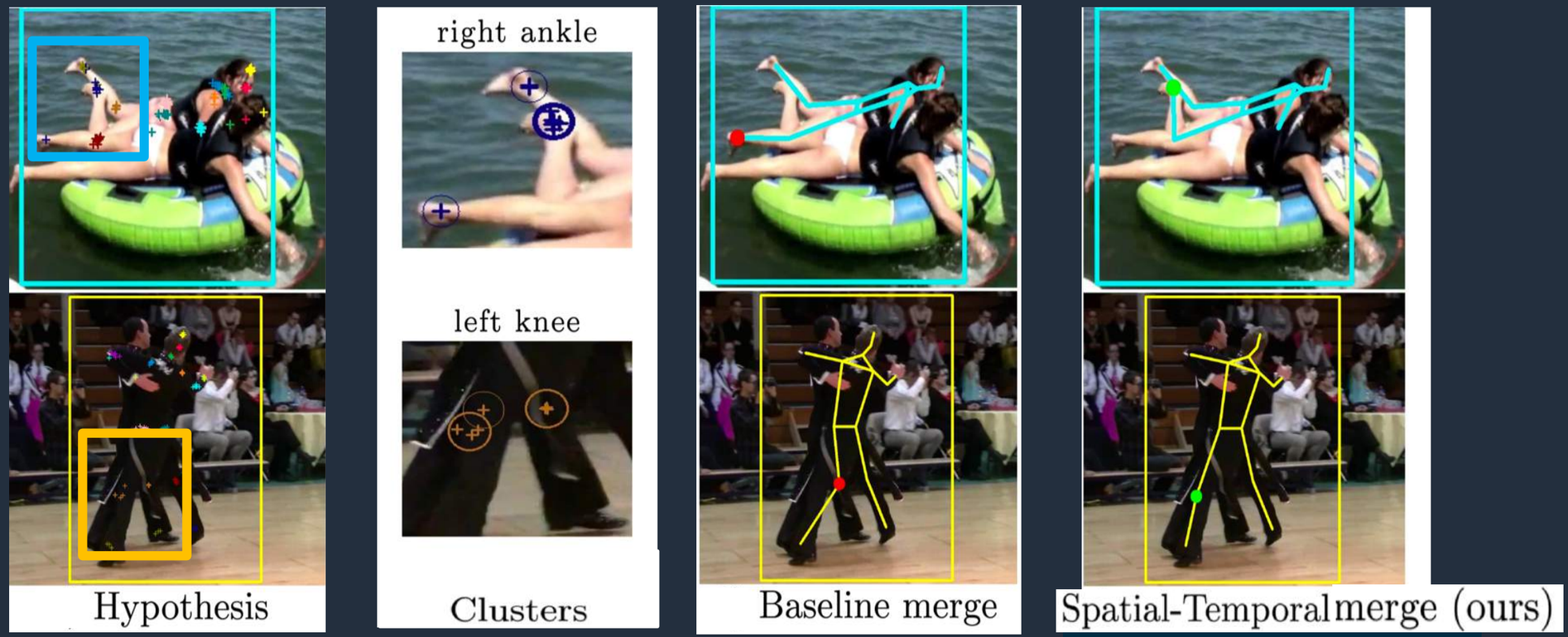


© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Our approach: 3 components

1. Clip tracking network
2. Video Tracking Pipeline
3. **Spatial-Temporal refinement**

   Merges multiple pose hypotheses --> picks the optimal one (spatial-temporal optimization)



© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Results on validation set

- SOTA on both human pose estimation (mAP) and tracking (MOTA)
- SOTA on both PoseTrack 2017 and 2018 validation
- SOTA against both top-down and bottom-up approaches

PoseTrack 2017 - validation

	Method	MOTA	mAP
Bottom-up	JointFlow [9]	59.8	69.3
	TML++ [13]	61.3	71.5
	STAF [23]	62.7	72.6
	STEmbedding [15]	**71.8**	77.0
Top-down	Detect&Track [11]	55.2	60.6
	PoseFlow [31]	58.3	66.5
	FastPose [33]	63.2	70.3
	FlowTrack [30]	65.4	76.7
	HRNet [26]	-	77.3
	Our approach	**71.6**	**83.8**

+6.2 MOTA
+6.5 mAP

PoseTrack 2018 - validation

	Method	MOTA	mAP
B-U	STAF [23]	60.9	70.4
	TML++ [13]	65.7	74.6
T-D	PT_CPN++ [32]	64.0	80.9
	Our approach	**68.7**	**81.5**

+3 MOTA

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Official challenge (Test set)

- SOTA on both PoseTrack2017 and 2018 benchmarks

Challenge 3: Multi-Person Pose Tracking

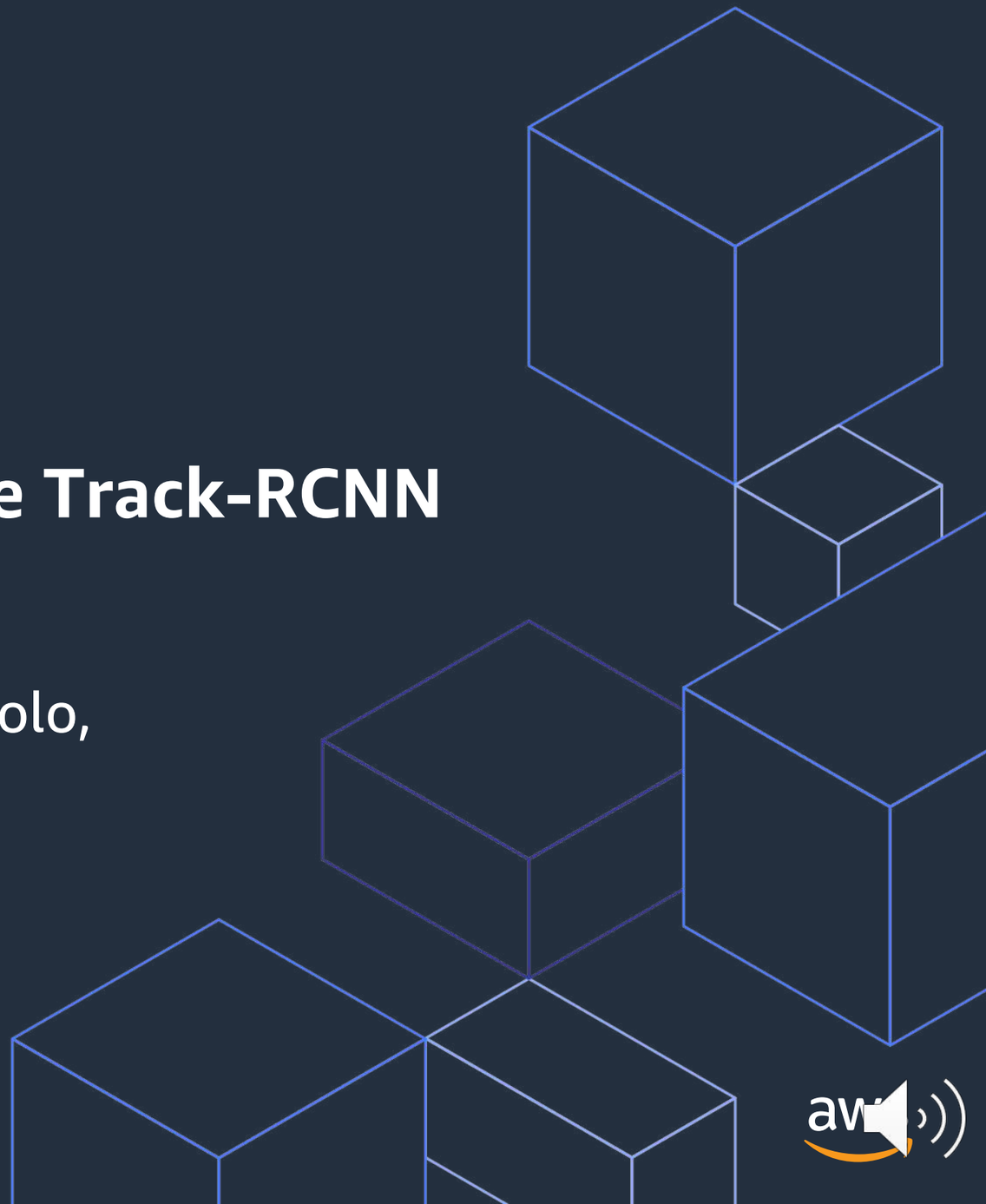
No.	Entry	Additional Training Data	total AP	total MOTA
1	DetTrack	+ COCO	74.14	64.09
2	KeyTrack	+ COCO	74.04	61.15
3	PGPT	+ COCO	72.57	60.17
4	CorrTrack	+ COCO	74.21	60.01
5	POINet	+ COCO	72.49	58.41

+3 MOTA

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Multi-object Tracking with Siamese Track-RCNN (Under submission)

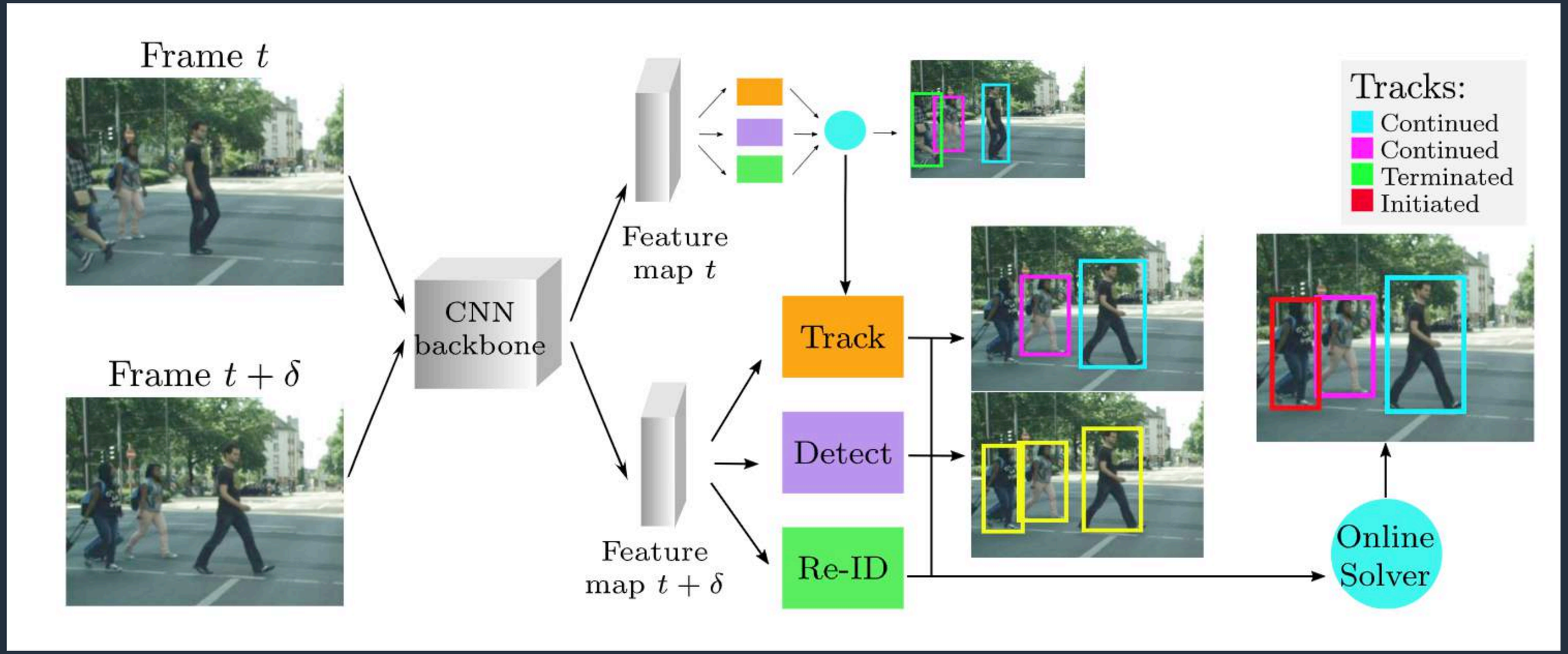
Bing Shuai, Andrew Berneshawi, Davide Modolo, Joseph Tighe



© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Siamese Track-RCNN

Three functional branches with joint training



© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

MOT Challenge 2017

Method	Year	MOTA ↑	IDF1 ↑	MT ↑	ML ↓	FP ↓	FN ↓	IDsw ↓
Siamese Track-RCNN	2020	**59.6**	**60.1**	**23.9%**	**33.9%**	15532	**210519**	2068
DeepMOT-Tracktor [54]	2019	53.7	53.8	19.4%	36.6%	**11731**	247447	1947
Tracktor++ [5]	2019	53.5	52.3	19.5%	36.6%	12201	248047	2072
DeepMOT-SiamRPN [54]	2019	52.1	47.7	16.7%	41.7%	12132	255743	2271
eHAF [44]	2018	51.8	54.7	23.4%	37.9%	33212	248047	**1834**
FWT [27]	2017	51.3	47.6	21.4%	35.2%	24101	247921	2648
jCC [30]	2018	51.2	54.5	20.9%	37.0%	25937	247822	1802
STRN [53]	2019	50.9	56.5	20.1%	37.0%	27532	246924	2593
MOTDT17 [7]	2018	50.9	52.7	17.5%	35.7%	24069	250768	2474
MHT_DAM [31]	2015	50.7	47.2	20.8%	36.9%	22875	252889	2314

Table 1. *Results on MOT17 test set using the provided public detections.*

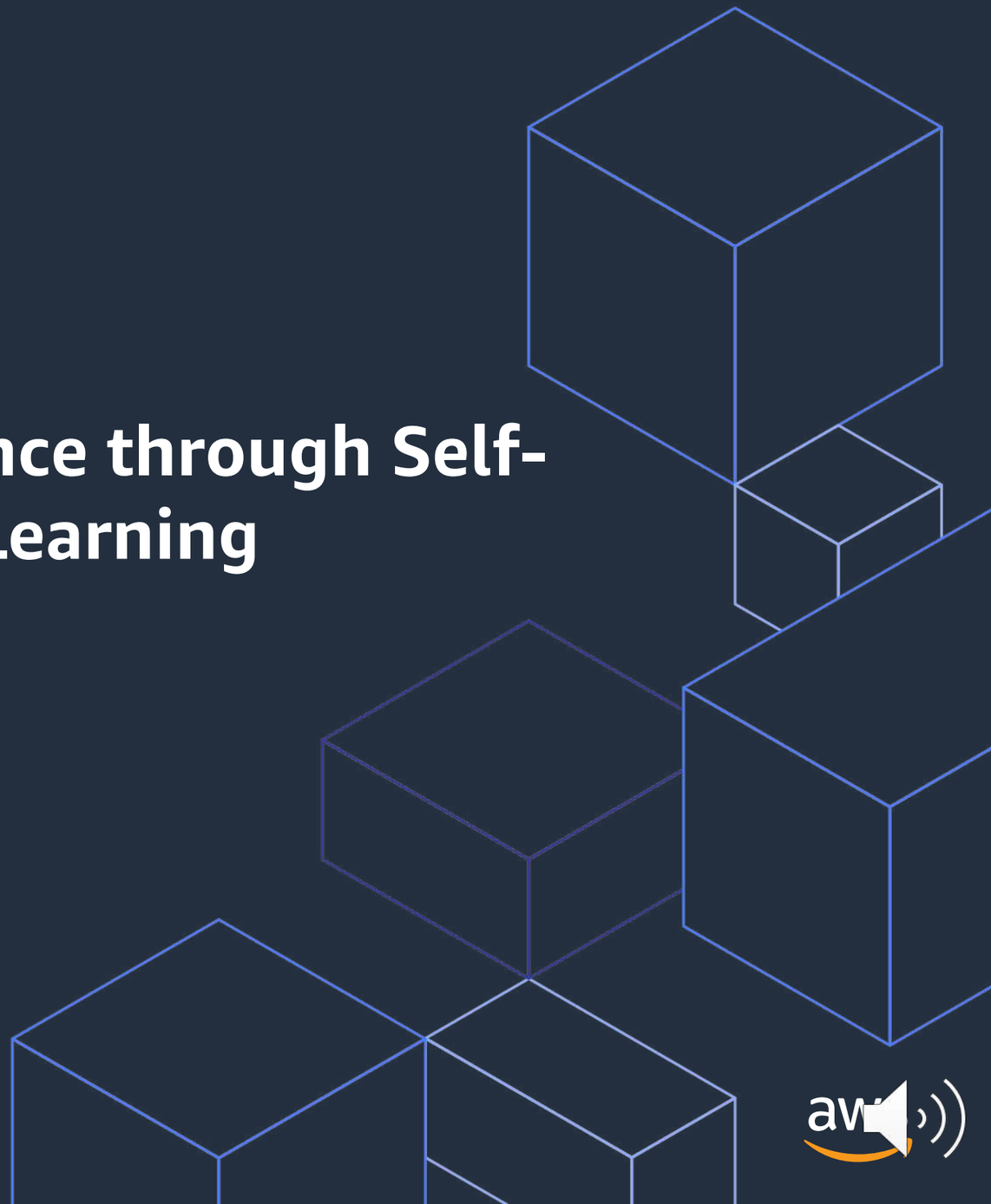
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Tracking on JTA

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Improving Temporal Correspondence through Self-supervised Visual Representation Learning (Under submission)

Daniel McKee, Bing Shuai, Davide Modolo,
Joseph Tighe, Svetlana Lazebnik



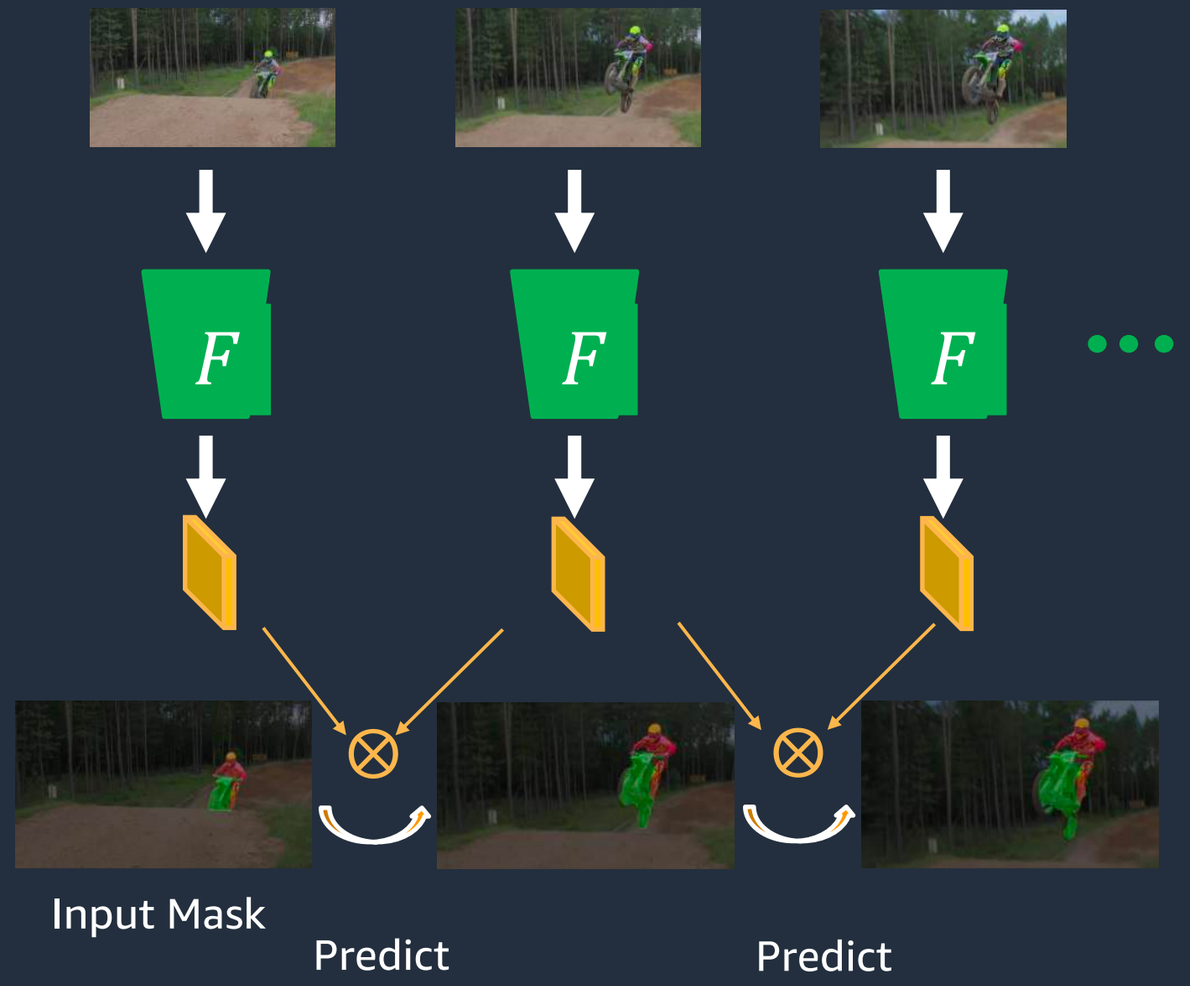
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Task

Given labeled starting frame, propagate dense labels through video using encoder F

Labels are propagated based on computed affinities between feature locations

We train F in a fully self-supervised manner



© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Previous Temporal Correspondence Methods

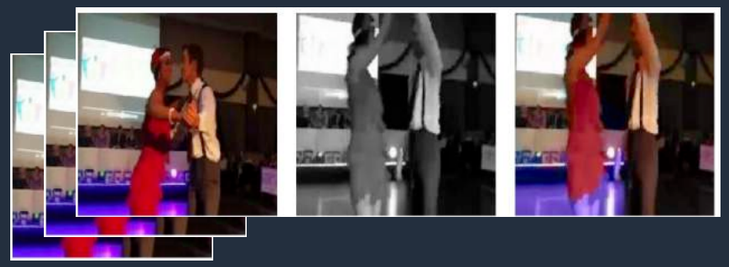
Self-supervised signals:

Colorization

Temporal cycle consistency

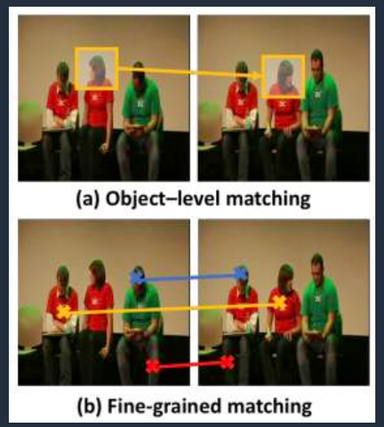
All these methods rely only inter-frame signals for training

Colorizing video frames



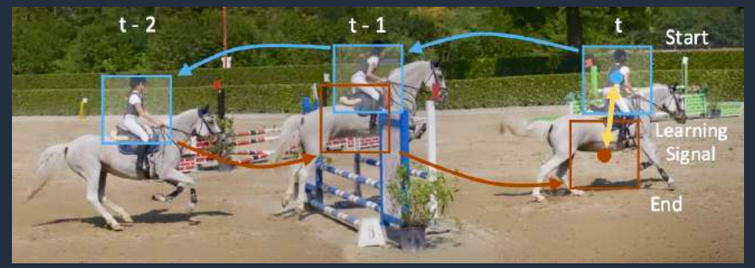
Vondrick et al., 2018

Localization+colorization



Li et al., 2019

Temporal cycle consistency



Wang et al., 2019

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Image-based Representation Learning

Motivation: ImageNet Classifiers sets a strong baseline without training on video, what about self-supervised visual representation learning?

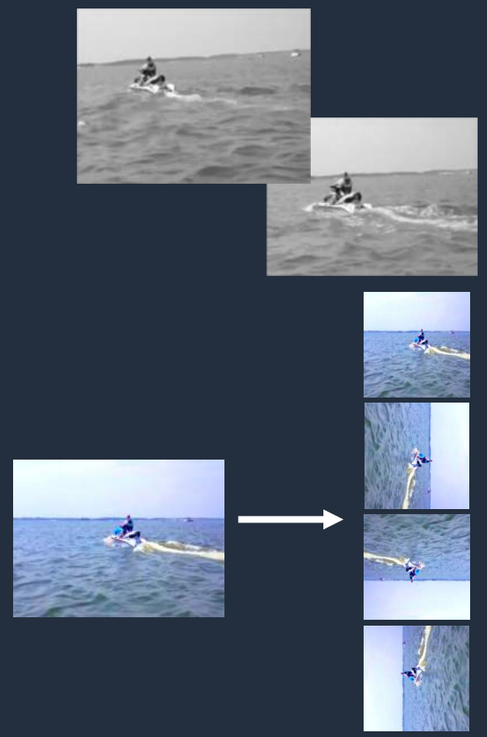
Self-supervised methods show similarly strong performance:

Method	Supervision	$\mathcal{J}$	$\mathcal{F}$
RotNet	✗	44.8	50.7
DeepCluster VGG16	✗	48.2	53.2
ImageNet Classifier	✓	51.3	56.6

Our goal: incorporate temporal-based objectives with image-based self-supervised representation learning

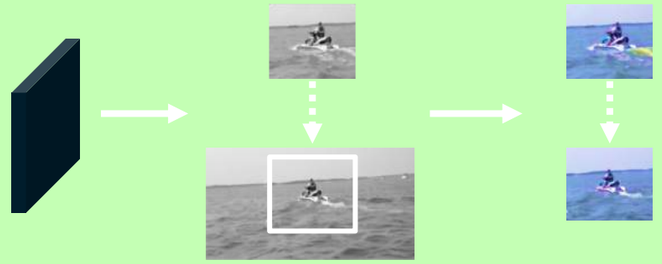
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

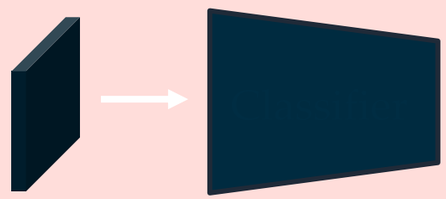
Our Method



Encoder

Patch localization & colorization





0°
90°
180°
270°

Visual representation learning

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Results

DAVIS: video object segmentation

Method	Supervision	$\mathcal{J}$	$\mathcal{F}$
RotNet	✗	44.8	50.7
TimeCycle [7]	✗	46.4	50.0
JointTask [4]	✗	57.7	61.3
Ours	✗	**61.3**	**66.1**
ImageNet Classifier	✓	51.3	56.6
OSVOS [1]	✓✓	56.6	63.9
DMM-Net [8]	✓✓	68.1	73.3

+3.6 J
+4.8 F

✓or ✓✓denote image-level or pixel supervision resp.

VIP: human part segmentation

Method	Supervision	mIoU	AP^r_{vol}
RotNet	✗	26.9	11.6
TimeCycle [7]	✗	28.9	15.6
JointTask [4]	✗	34.1	17.7
Ours	✗	**38.3**	**22.2**
ImageNet Classifier	✓	31.9	15.9
ATEN [9]	✓✓	37.9	24.1

+4.2 mIoU
+4.5 AP

JHMDB: human keypoints

Method	Supervision	PCK@.1	PCK@.2
RotNet	✗	56.0	76.1
TimeCycle [7]	✗	57.3	78.1
JointTask [4]	✗	58.6	79.8
Ours	✗	**59.8**	**81.3**
ImageNet Classifier	✓	57.3	78.5

+1.2 PCK@.1
+1.5 PCK@.2

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Qualitative Results

Input JointTask Ours GT

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Qualitative Results

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Semantic Video Understanding

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

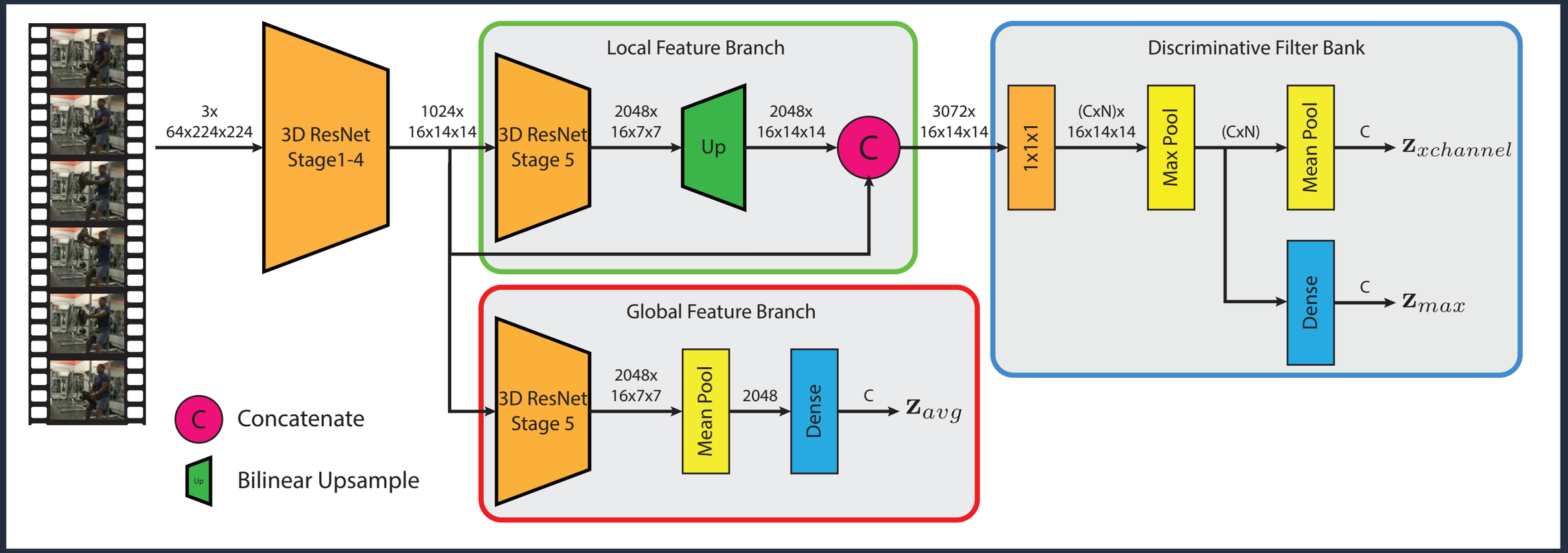
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Activity Classification using Spatial-Temporal Finegrained Discriminative Filter Banks

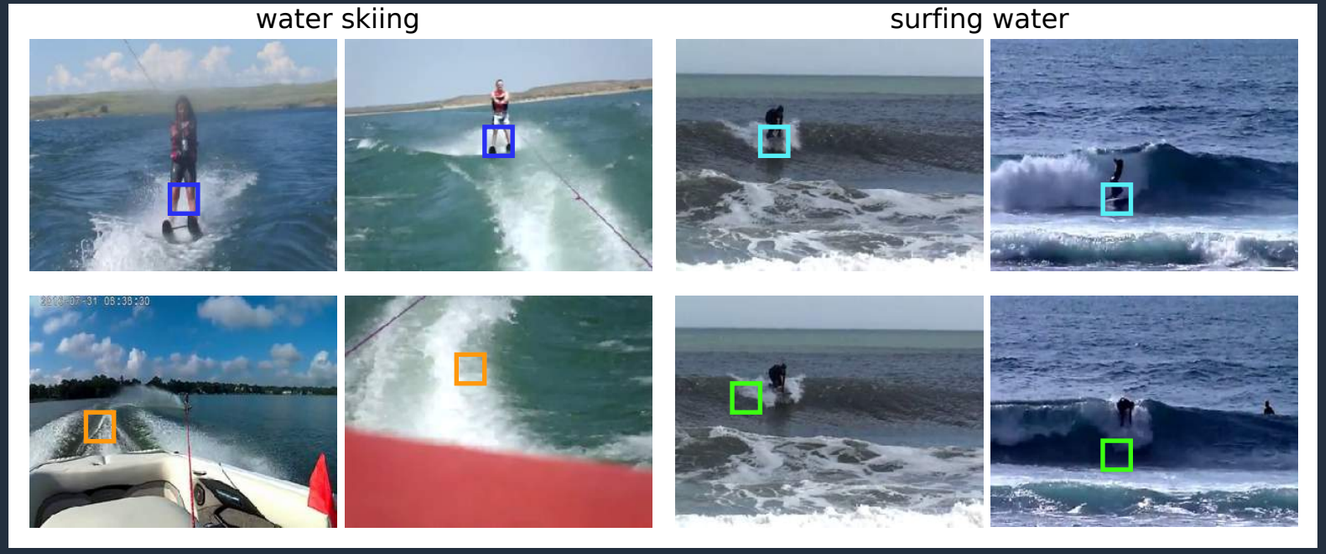
Brais Martinez Alonso, Davide Modolo, Yuanjun Xiong, Joseph Tighe

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

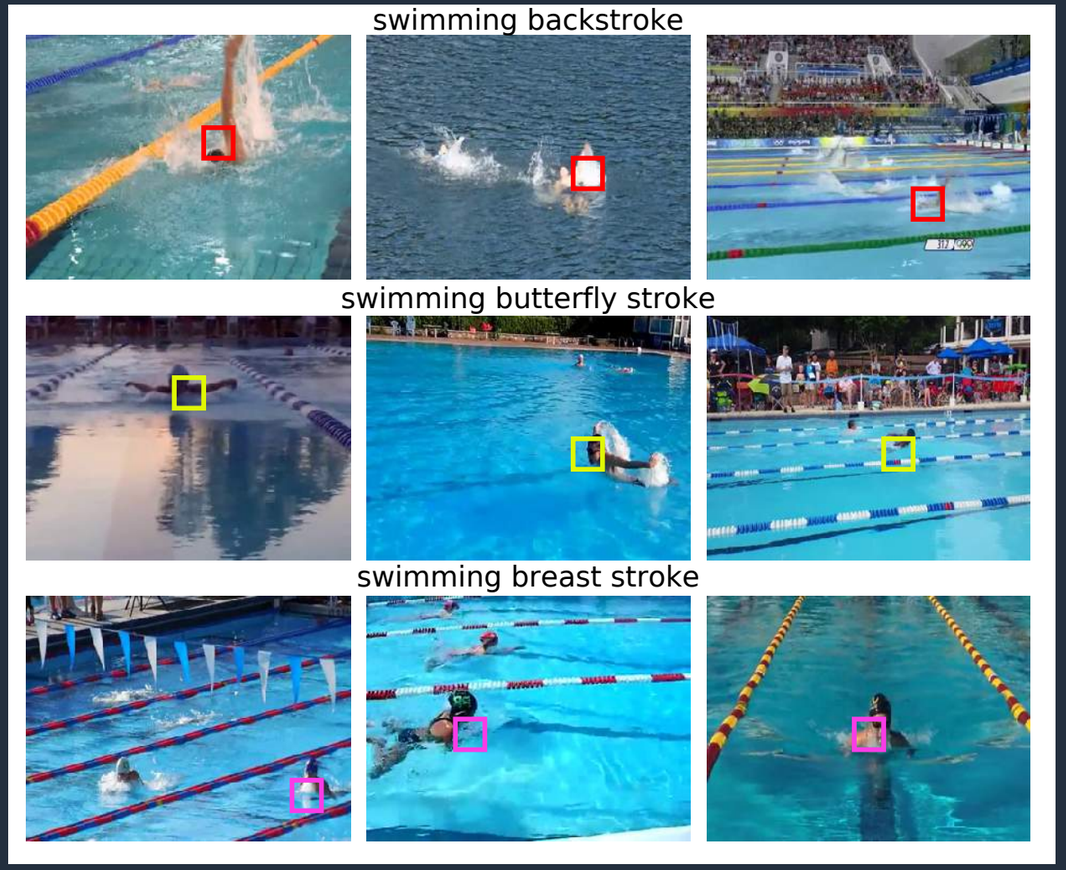
Activity Classification using Spatial-Temporal Finegrained Discriminative Filter Banks



© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.



- **Top**: filters that have specialized on the person's leg and the object being ridden (ski or surf)
- ***Bottom***: filters that have specialized on the texture of the water (wake or wave)



- Filters activate on the frame and location of the swimmer appearing in a canonical pose
- Robust to viewpoints and scenes' changes

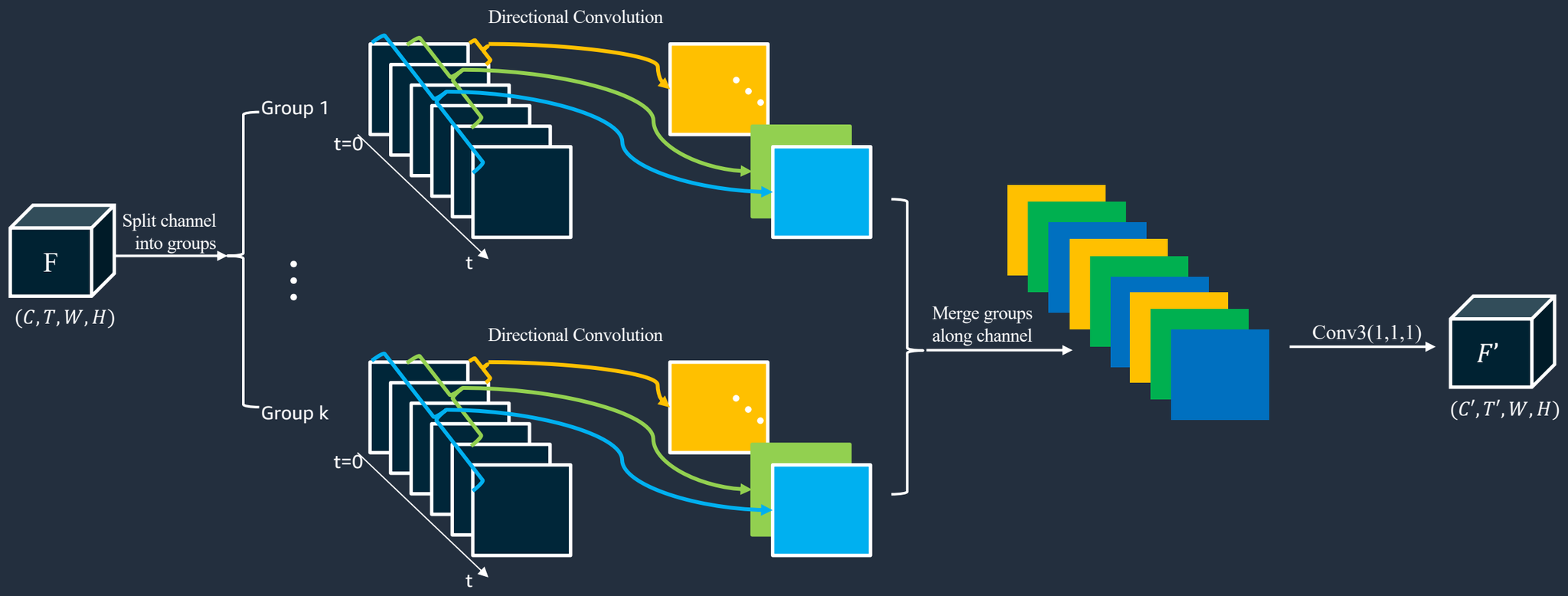
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Directional Temporal Modeling For Action Recognition (Under submission)

Xinyu Li, Bing Shuai, Joseph Tighe

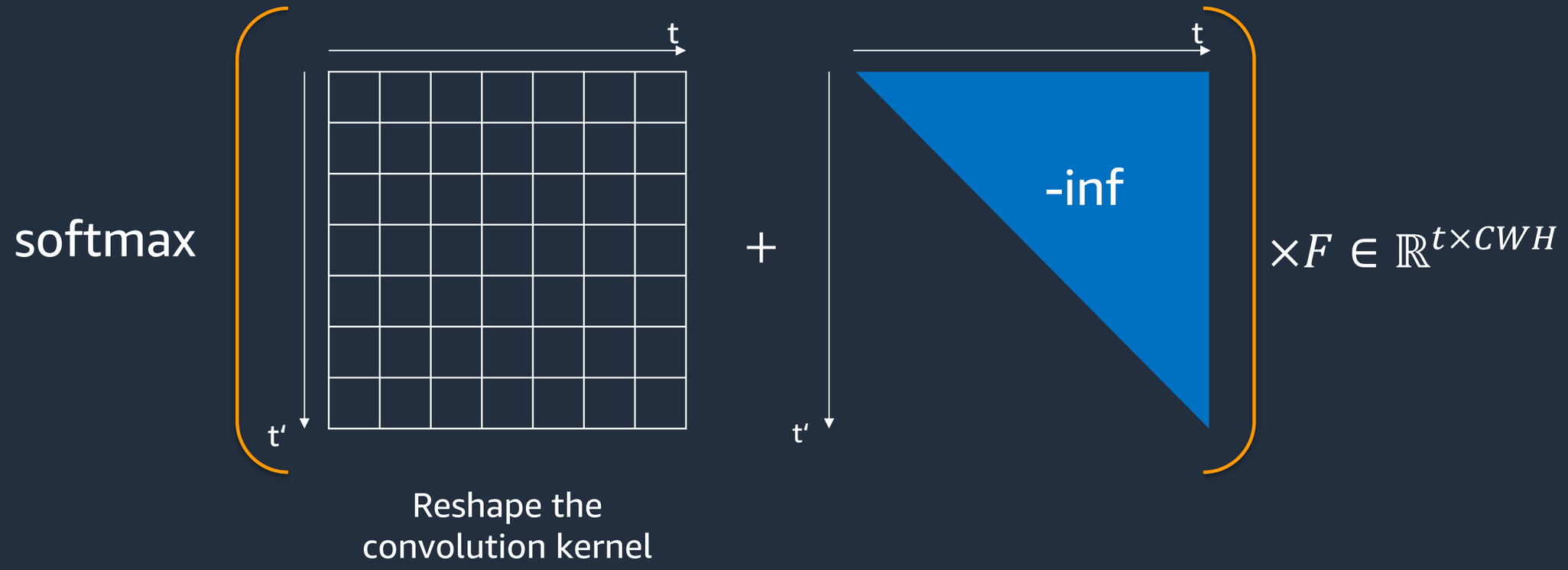
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Channel Independent Directional Convolution (CIDC)



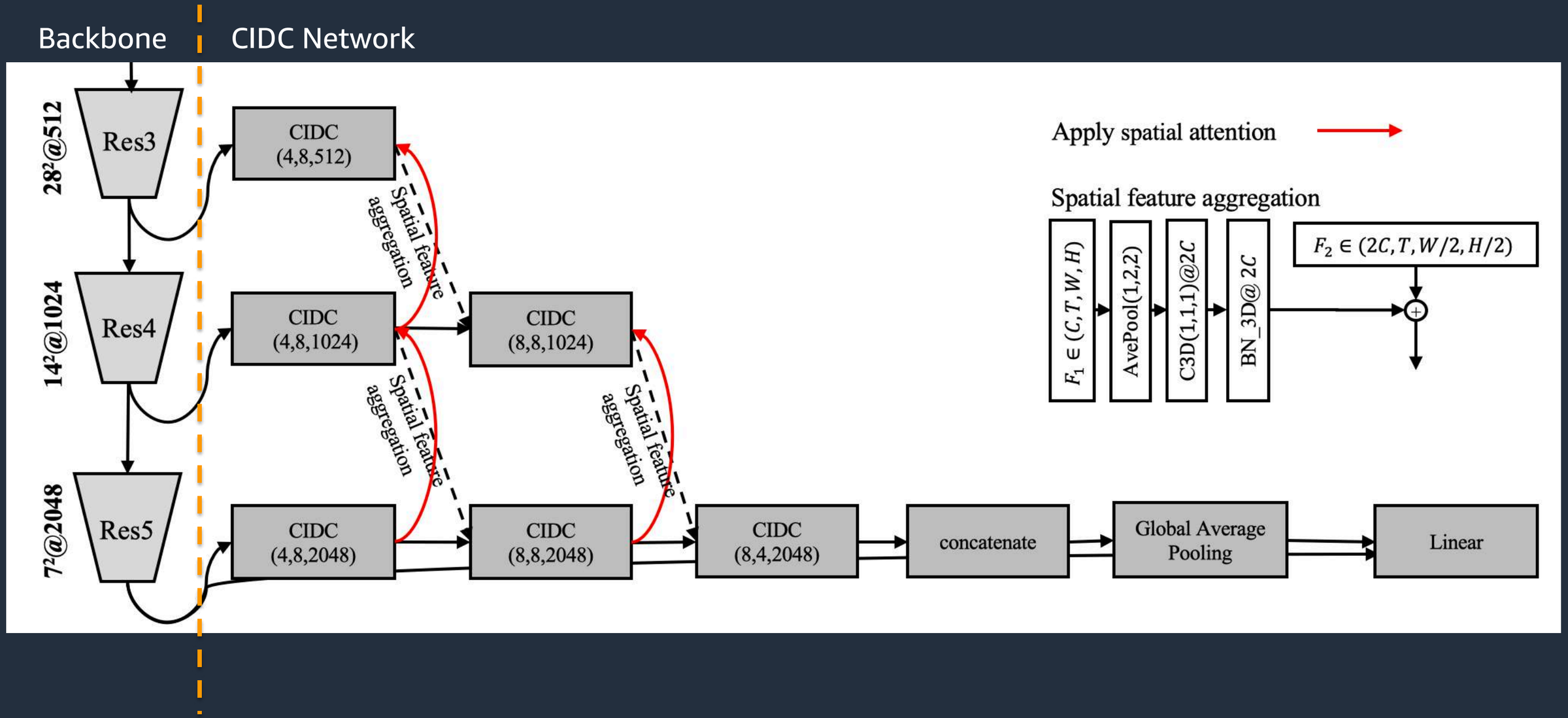
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

CIDC Implementation (Directional Mask)



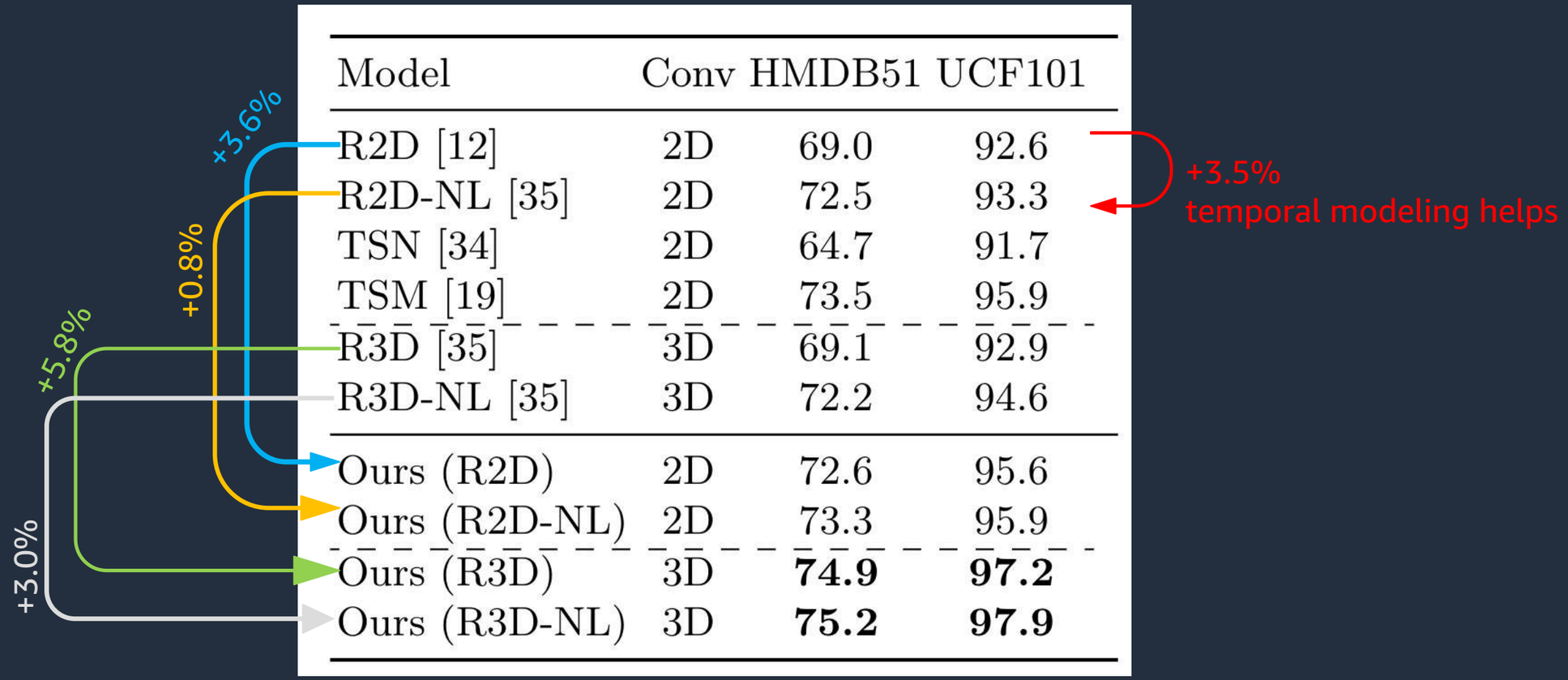
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

CIDC Network



© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

CIDC Results – Small dataset



Model	Conv	HMDB51	UCF101
R2D [12]	2D	69.0	92.6
R2D-NL [35]	2D	72.5	93.3
TSN [34]	2D	64.7	91.7
TSM [19]	2D	73.5	95.9
R3D [35]	3D	69.1	92.9
R3D-NL [35]	3D	72.2	94.6
Ours (R2D)	2D	72.6	95.6
Ours (R2D-NL)	2D	73.3	95.9
Ours (R3D)	3D	**74.9**	**97.2**
Ours (R3D-NL)	3D	**75.2**	**97.9**

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

CIDC Results – Large Dataset

Table 2: Result comparison on Kinetics-400 dataset. For fair comparison, we only compare with methods that use ResNet-50.

Model	Conv	Top1	Top5
TSN [34]	2D	70.6	89.2
R2D [35]	2D	70.2	88.7
R2D-NL [35]	2D	72.4	89.8
R3D[35]	3D	73.1	90.1
R3D-NL [35]	3D	73.6	90.9
TSM [19]	2D	74.1	91.2
bLVNet [5]	3D	73.5	91.2
Slowfast (4 × 16) [6]	3D	74.1	91.1
Ours (R2D)	2D	72.2	90.1
Ours (R2D-NL)	2D	72.8	90.5
Ours (R3D)	3D	73.6	90.9
Ours (R3D-NL)	3D	74.1	91.4

+2.0%
+0.5%
+0.4%
+0.50%
+2.2%

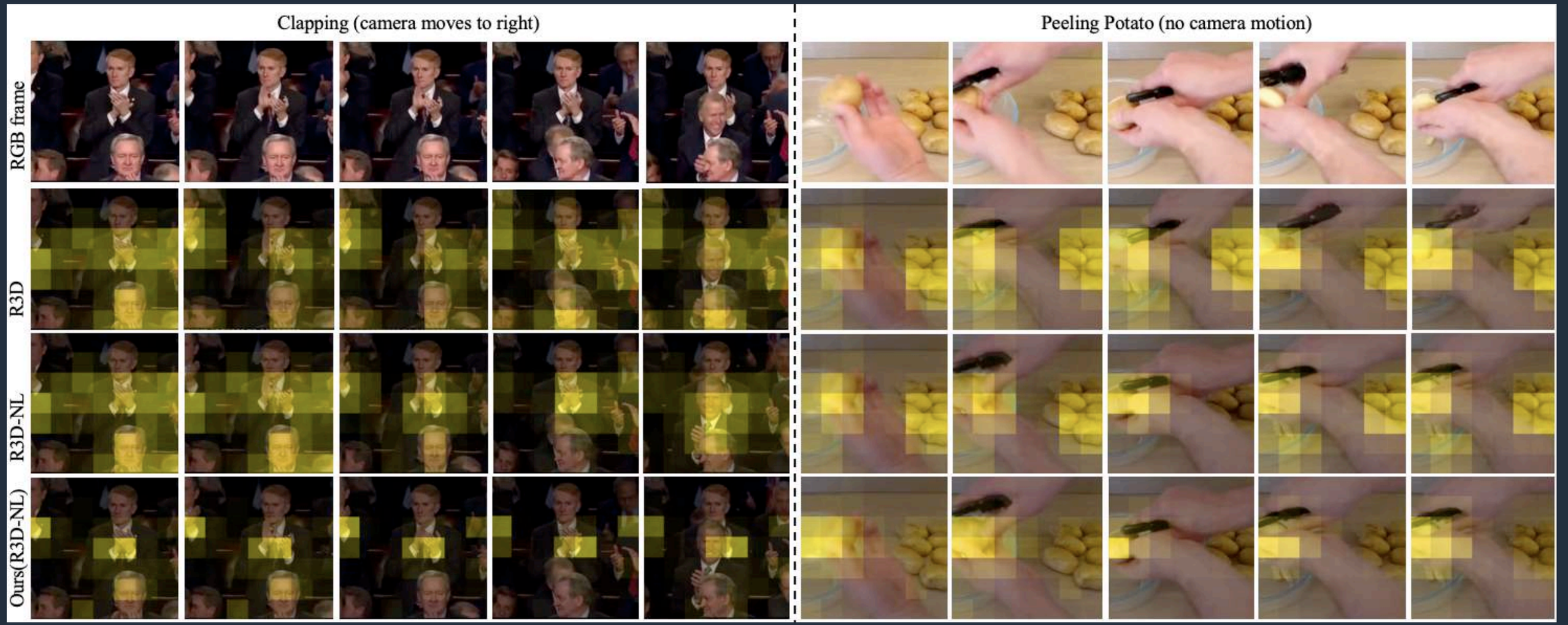
Table 3: Result comparison on Something-Something V2 dataset. We only compare with methods that use ResNet-50 as backbone and take as input the RGB video.

Model	Conv	Top1	Top5
TSN [19, 34]	2D	30.0	60.5
MultiScale TRN [40]	2D	48.8	77.6
2-stream TRN [40]	2D	55.5	83.1
Fine-grain [21]	3D	53.4	81.1
TSM [19]	2D	63.4	88.5
bLVNet [5]	3D	61.7	88.1
R2D	2D	35.5	65.4
R3D	3D	48.6	78.2
Ours (R2D)	2D	40.2	68.6
Ours (R3D)	3D	56.3	83.7

+4.7%
+7.6%

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

CIDC Visualization



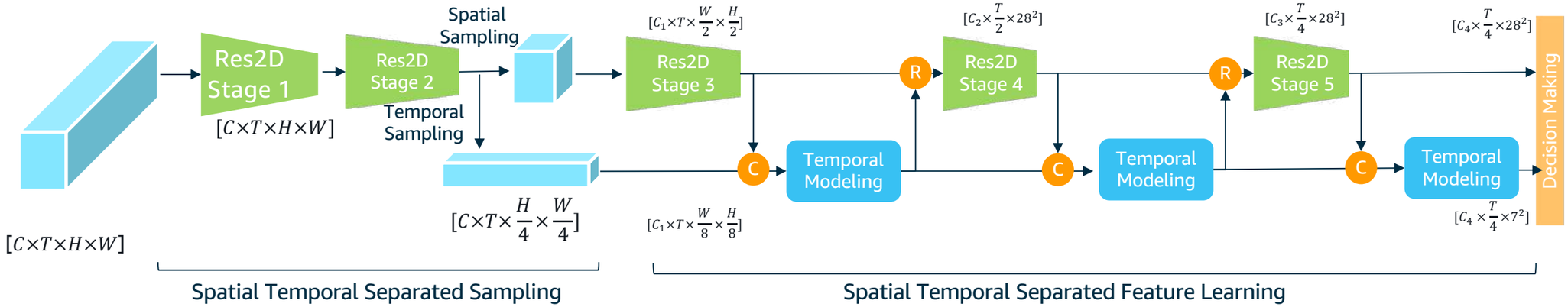
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Spatial Temporal Separated Network (Under submission)

Xinyu Li*, Chunhui Liu*, Hao Chen, Yi Zhu,
Joseph Tighe
* equally contributed

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

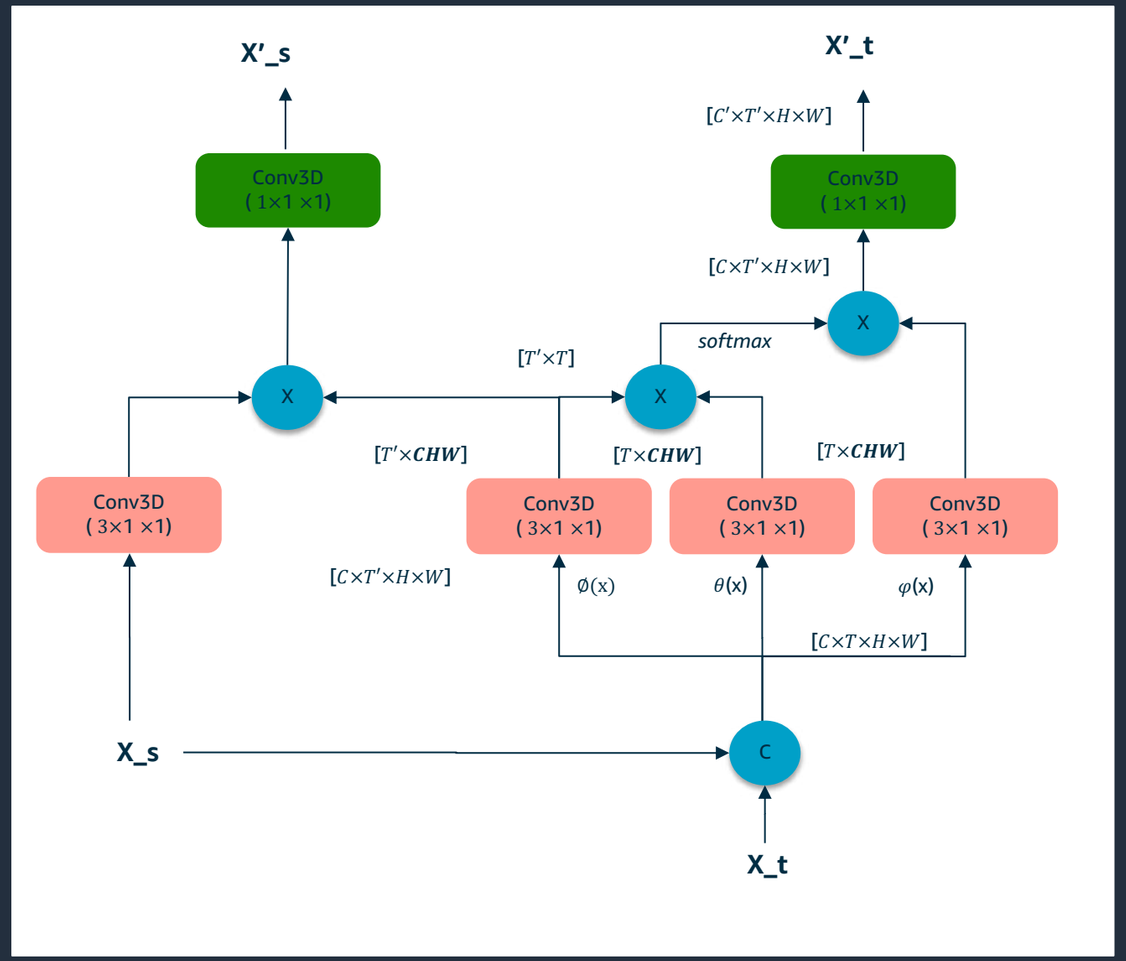
Spatial Temporal Separated Network



© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Spatial Temporal Separated Network: Temporal Modeling

- Spatial temporal information exchange by concat along channel
- Temporal to spatial feature aggregation by apply the same attention matrix



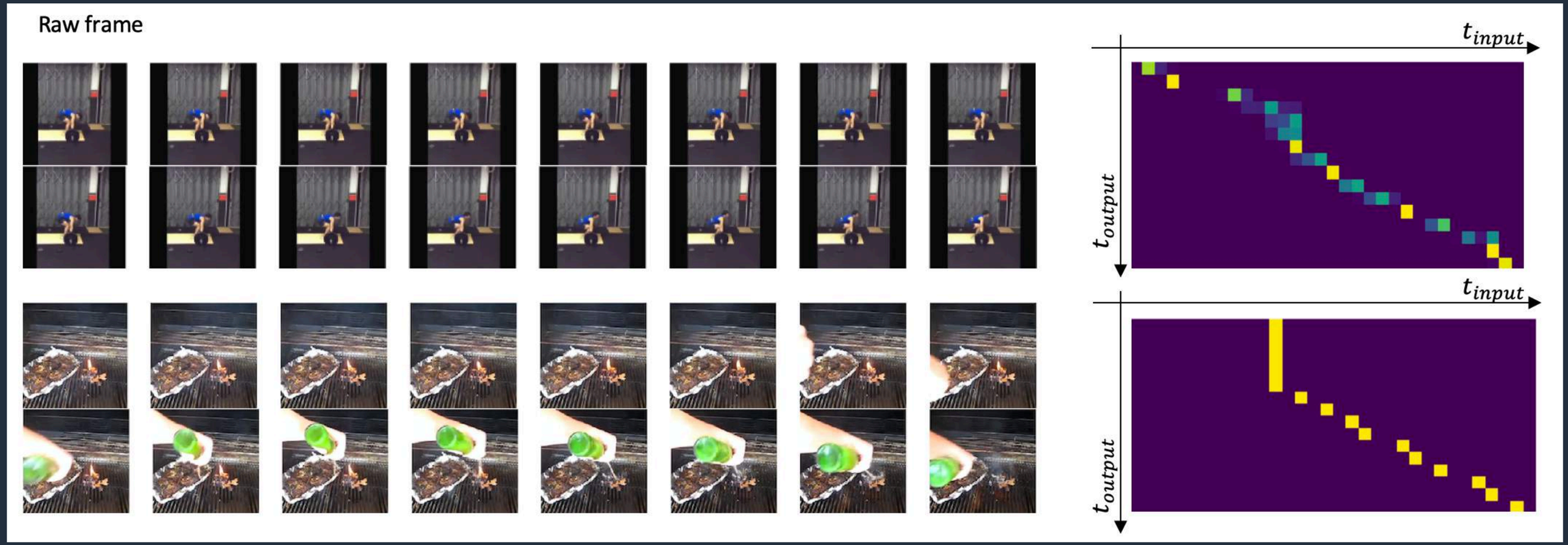
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Spatial Temporal Separated Network

Version	Spatial Branch	GFLOPs	Top1
R2D-50		64	69.9
R2D-50 NL			72.9
I3D-50		153	74.0
I3D-50 NL		282	75.2
I3D-101		167	75.1
I3D-101-NL		359	76.0
IR-CSN-101		73	76.2
Slowfast-50 4x16		36	75.3
Slowfast-50 8x8		66	76.6
Slowfast-101 8x8		106	77.2
STS	R2D-50	49	74.1
STS	R2D-101	64	75.3
STS	I3D-50	34	75.5
STS	I3D-101	102	76.6
STS	I3D-152	192	78.2
STS	X3D-M*	5.7	73.1

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Temporal Affinity Map

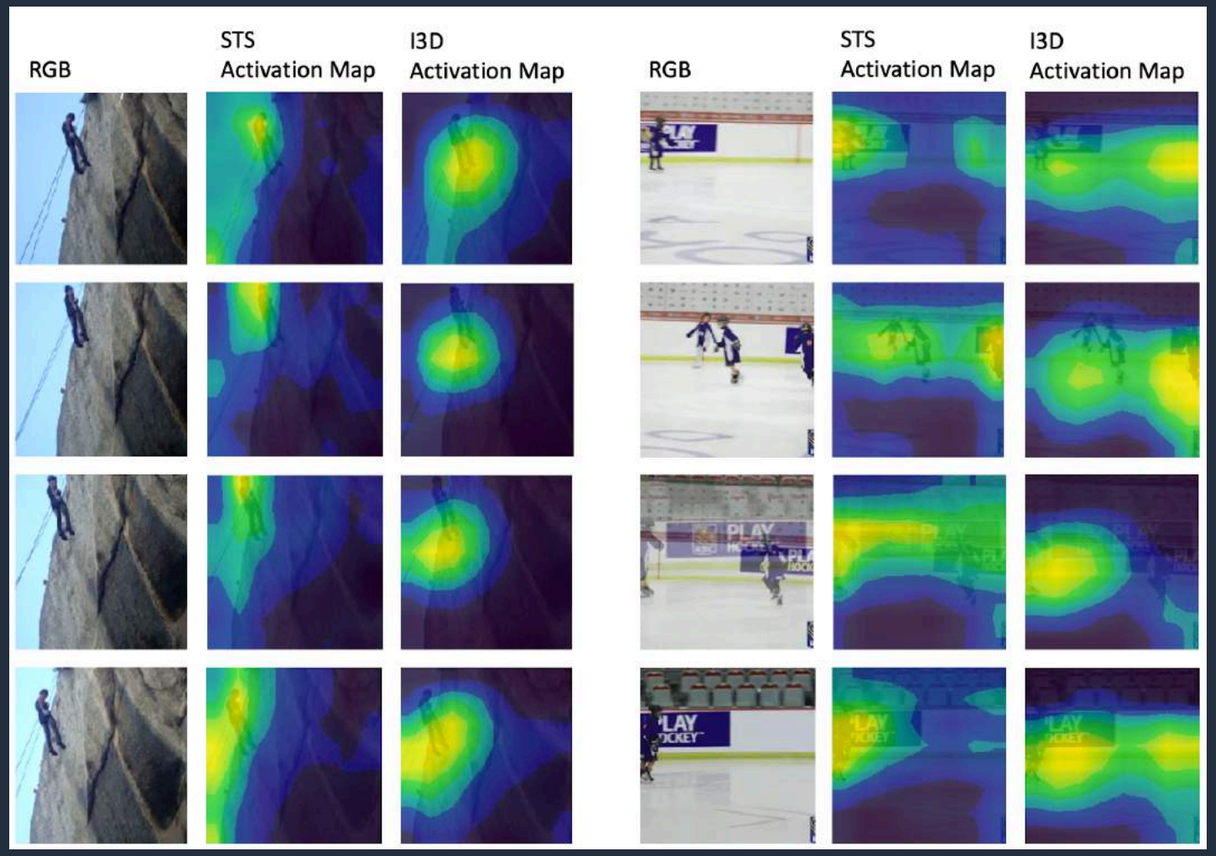


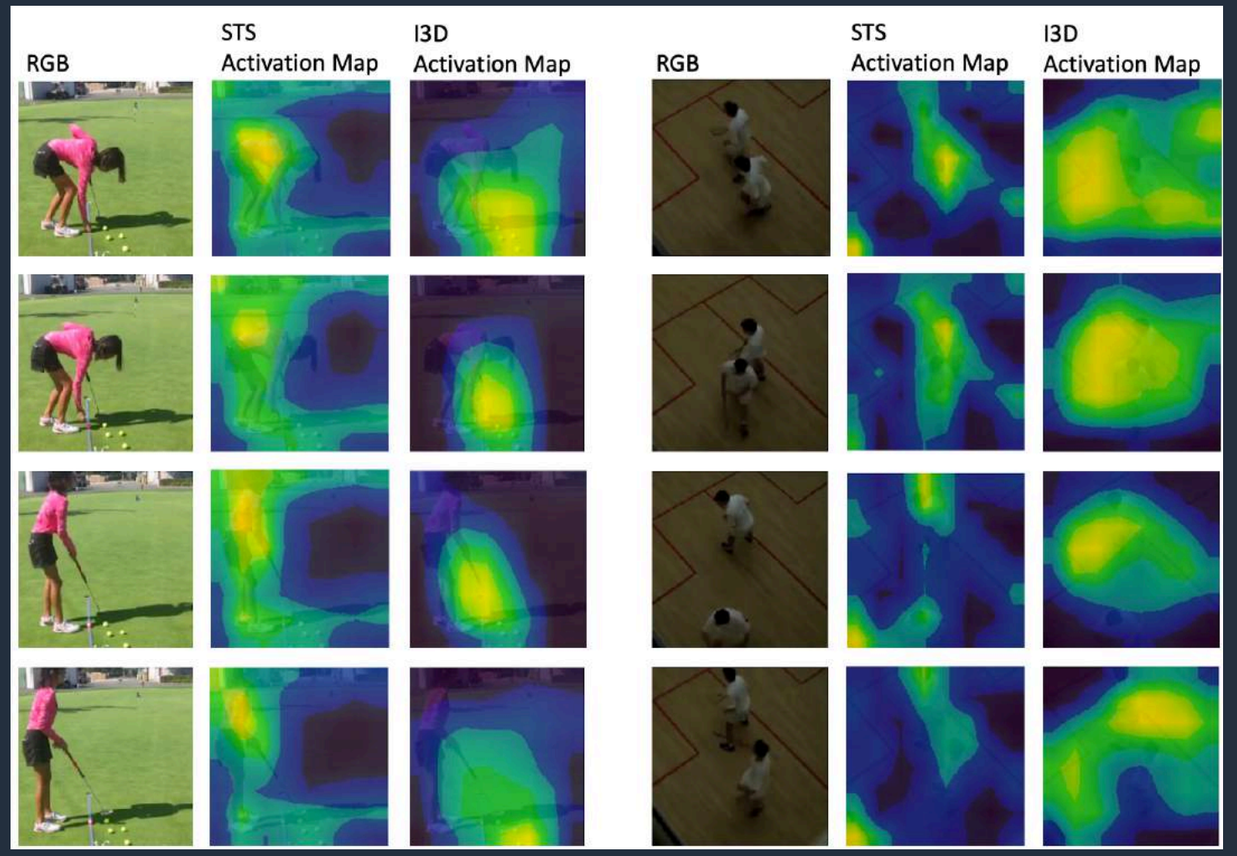
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Visualization – More examples

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Visualization – Compare with I3D





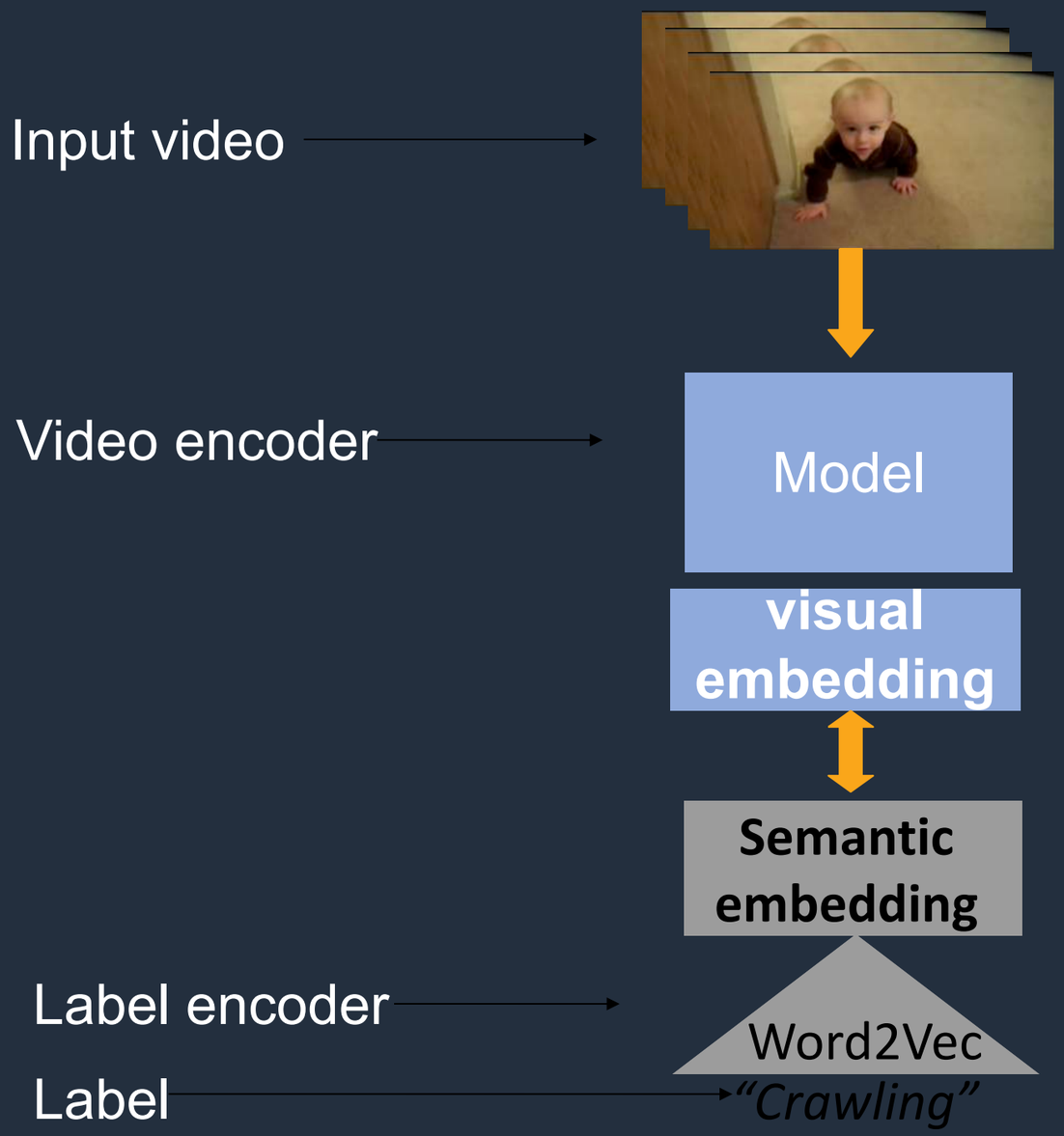
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Rethinking Zero-shot Video Classification: End-to-end Training for Realistic Applications

Biagio Brattoli, Joseph Tighe, Fedor Zhdanov,
Pietro Perona ， Krzysztof Chalupka

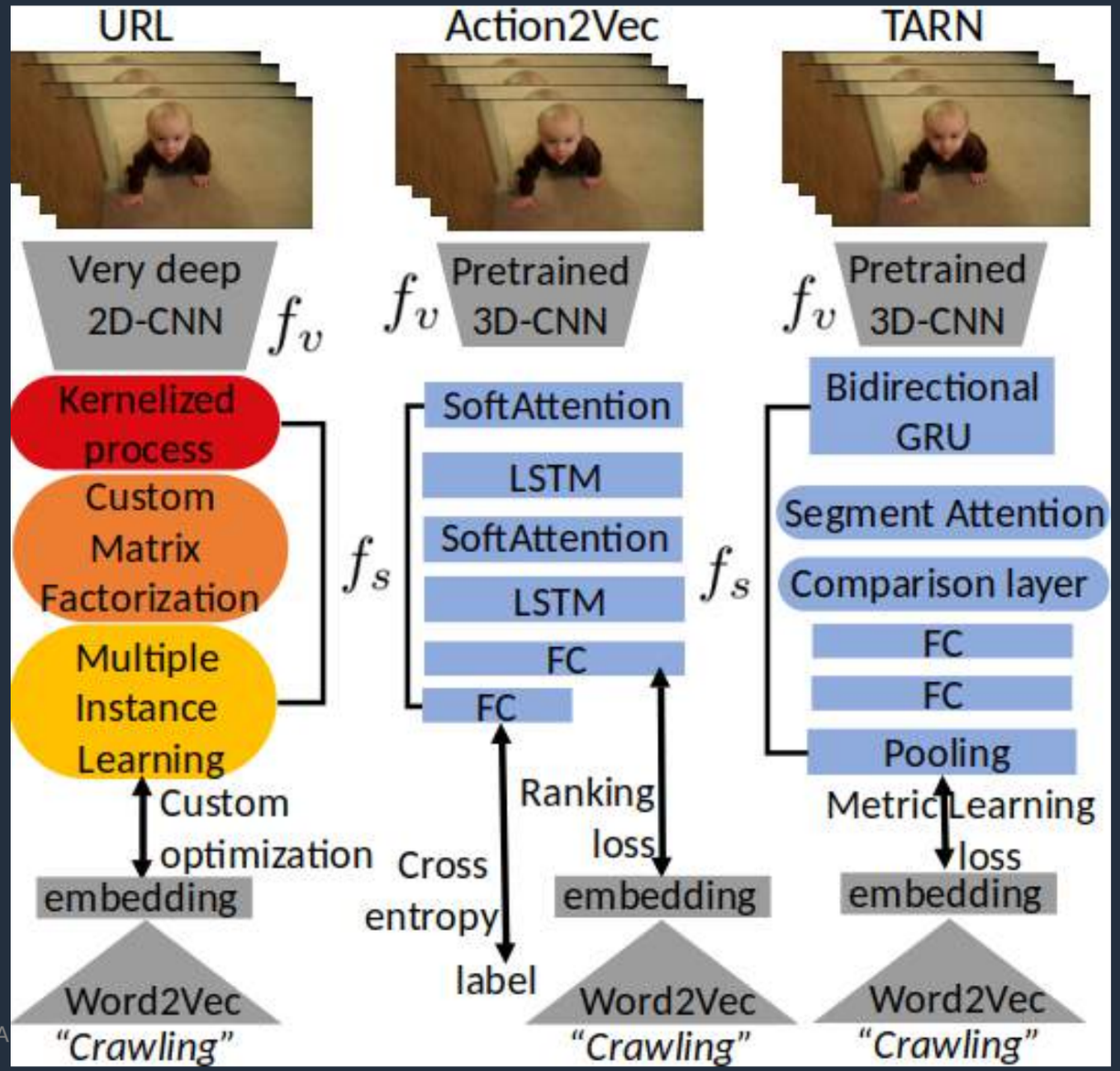
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Task



© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Related Work



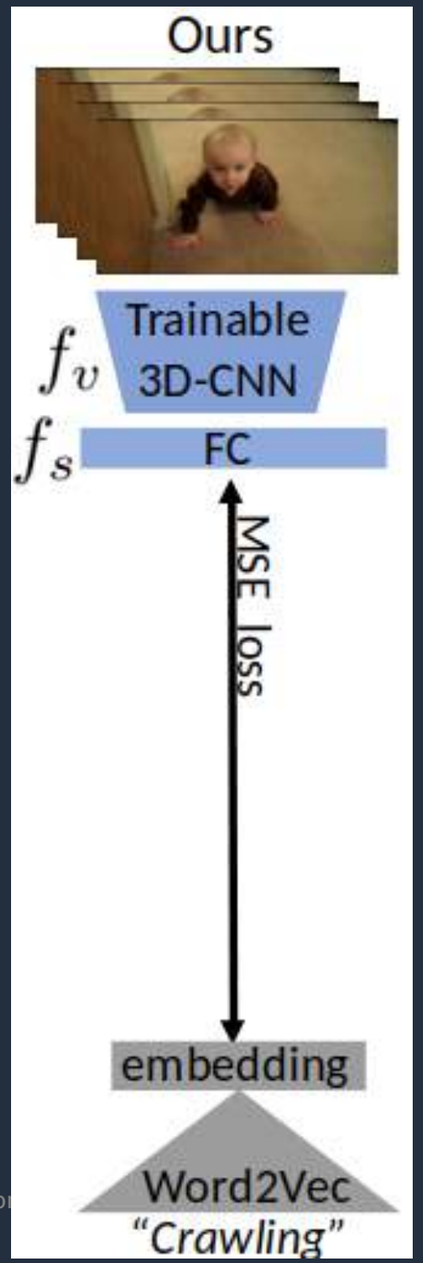
- Complex models
- Hard to reproduce
- Poor performance
- Zero-shot Learning paradigm not enforced in the pre-trained network

© 2020, A

Our Model

Out-of-the-box
3D CNN

Out-of-the-box
MSE loss



Protocol:
Trained once on Kinetics and tested on multiple datasets (UCF101, HMDB, Activity)

- Simple
- Easy to reproduce
- Good performance
- ZSL paradigm enforced

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Co

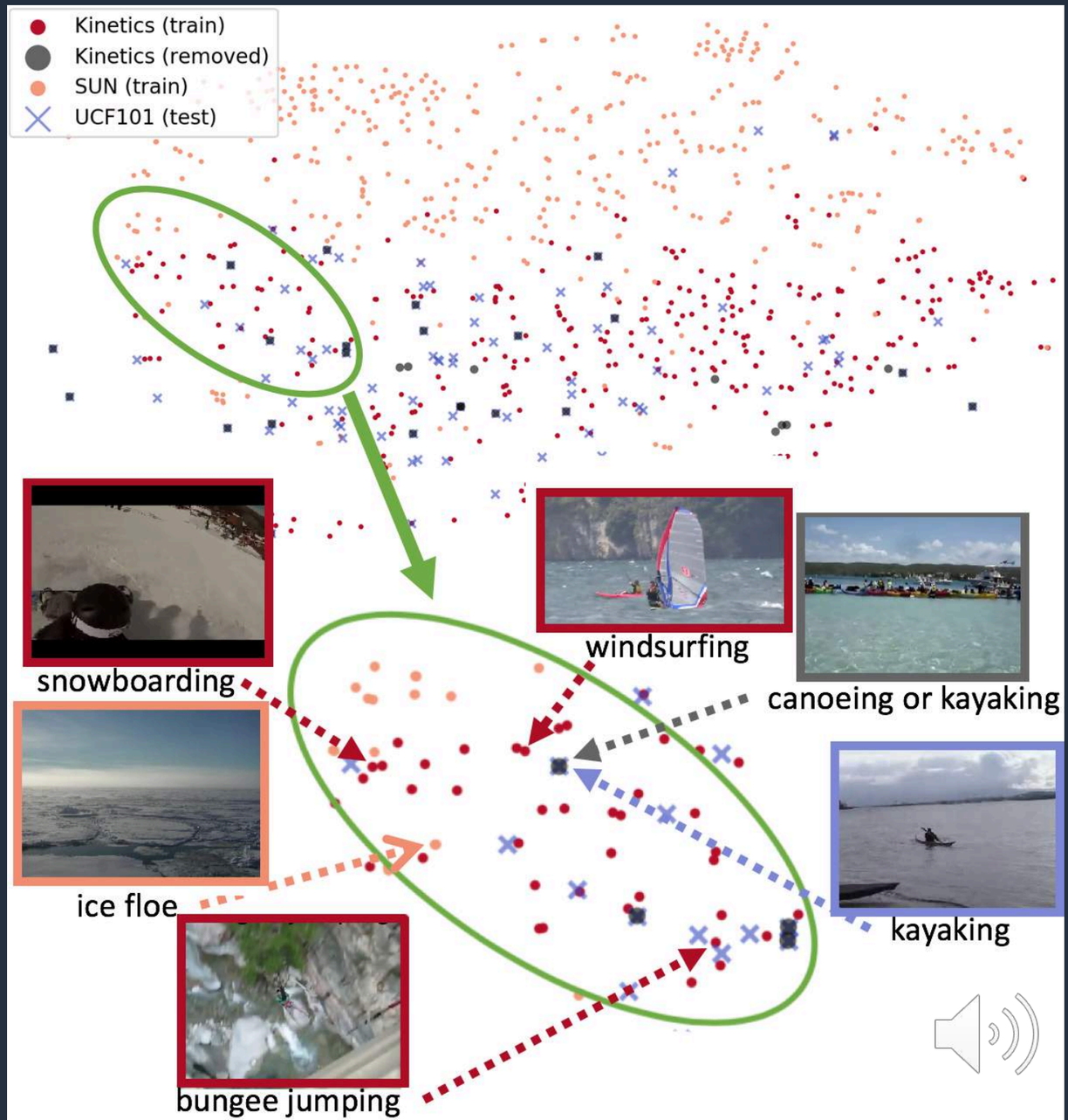
Enforcing ZSL paradigm

<u>Zero-shot learning paradigm:</u>

Training and testing classes should **NOT** overlap

Previous methods:
The ZSL paradigm is not enforced in the pre-trained CNN.

Example:
“kayaking” (UCF101) and
“canoeing and kayaking” (Kinetics)
are overlapping even though they have different names



© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Comparison: strict ZSL protocol

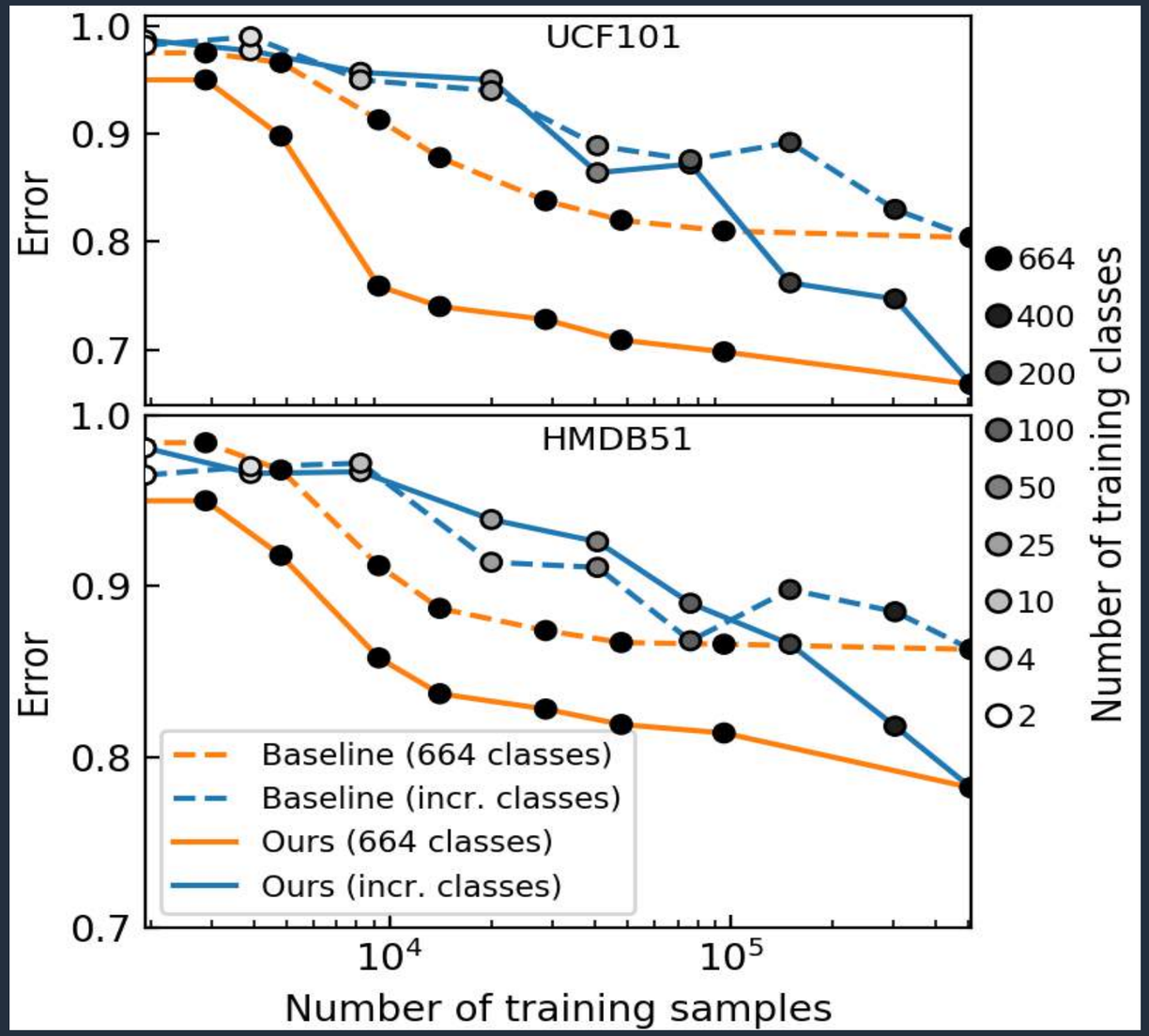
We outperform state-of-the-art while using strict ZSL protocol.

Dataset	VisualFeat	UCF	HMDB	Activity
URL [64]	ResNet200	42.5	51.8	-
DataAug [60]	-	18.3	19.7	-
InfDem [39]	I3D	17.8	21.3	-
Bidirectional [55]	IDT	21.4	18.9	-
FairZSL [40]	-	-	23.1	-
TARN [4]	C3D	19	19.5	-
Action2Vec [18]	C3D	22.1	23.5	-
Ours(605classes)	C3D	41.5	25.0	24.8
Ours(664classes)	C3D	43.8	24.7	-
Ours(605classes)	R(2+1)D_18	44.1	29.8	**26.6**
Ours(664classes)	R(2+1)D_18	**48**	**32.7**	-

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Weakness/Strenghts study

Better more training samples or classes?

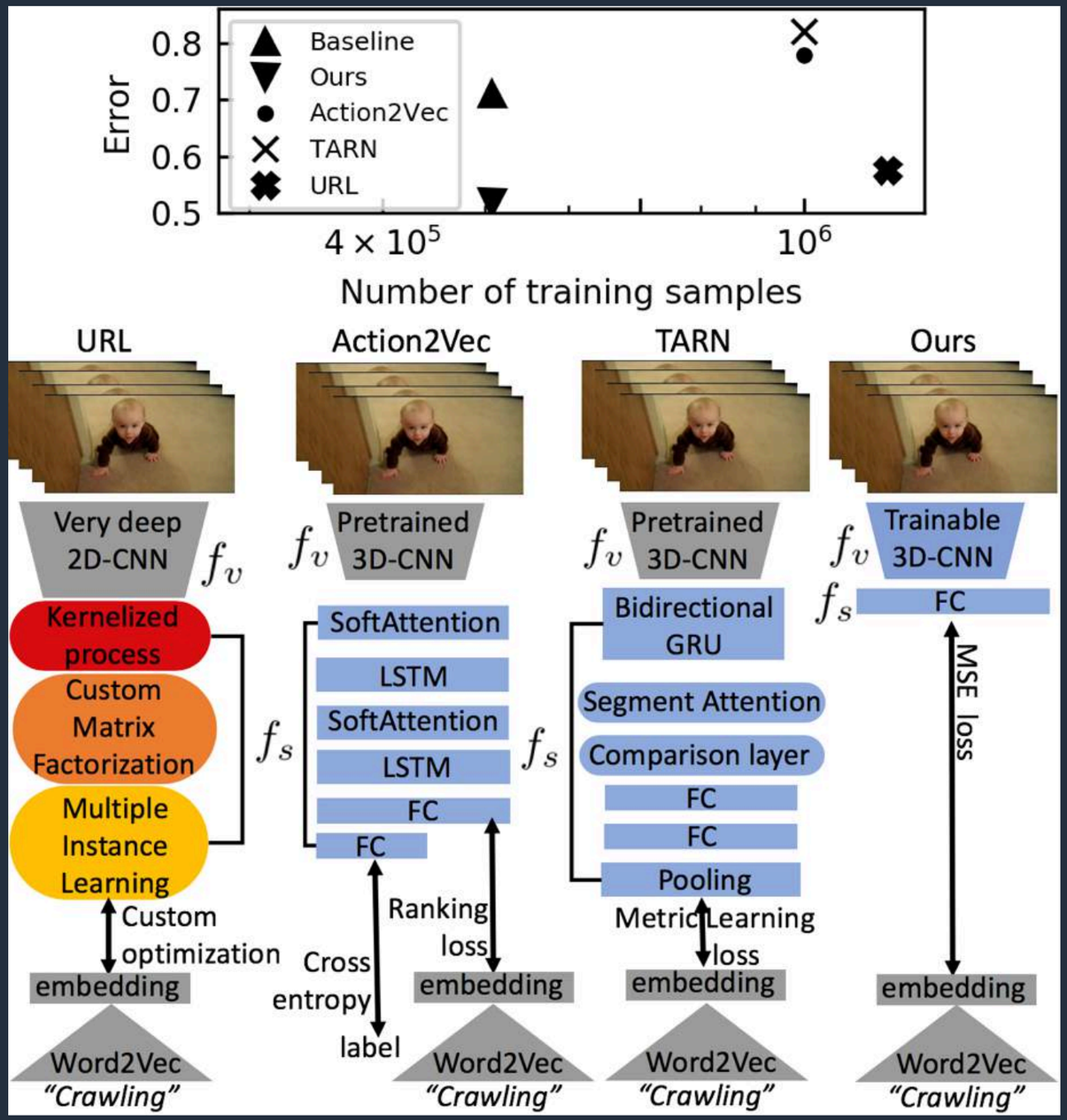


© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Conclusion

Contributions:

- First end-to-end model for ZSL video classification
- Strong and simple baseline for future work
- Cheap method for increasing training classes for a more robust model.
- Deep study of ZSL weaknesses and strengths



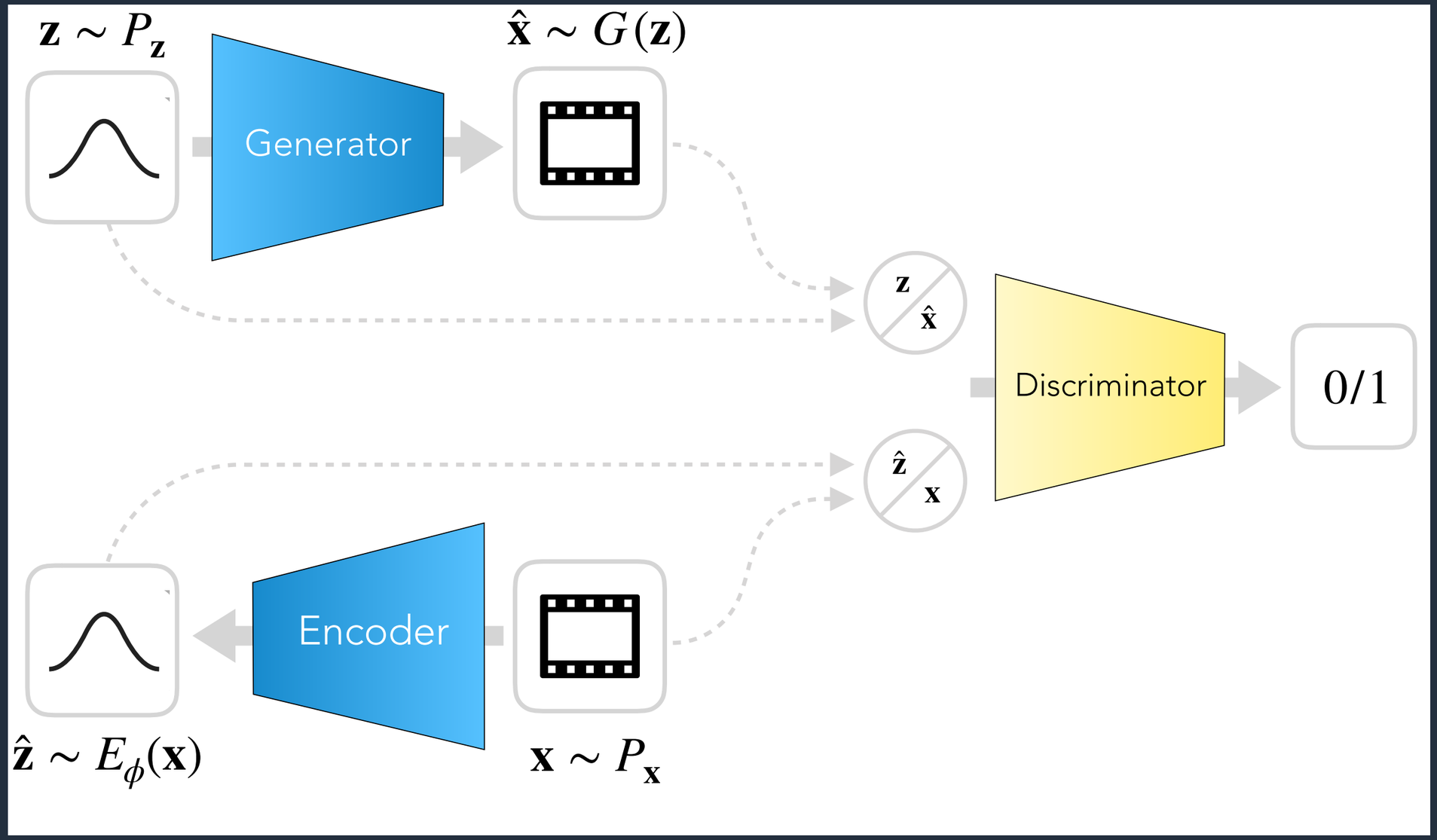
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Bidirectional GANs for Unsupervised Video Representation Learning

Tom F. H. Runia, Andrew Berneshawi, Rahul Rama Varior, Davide Modolo, Joseph Tighe

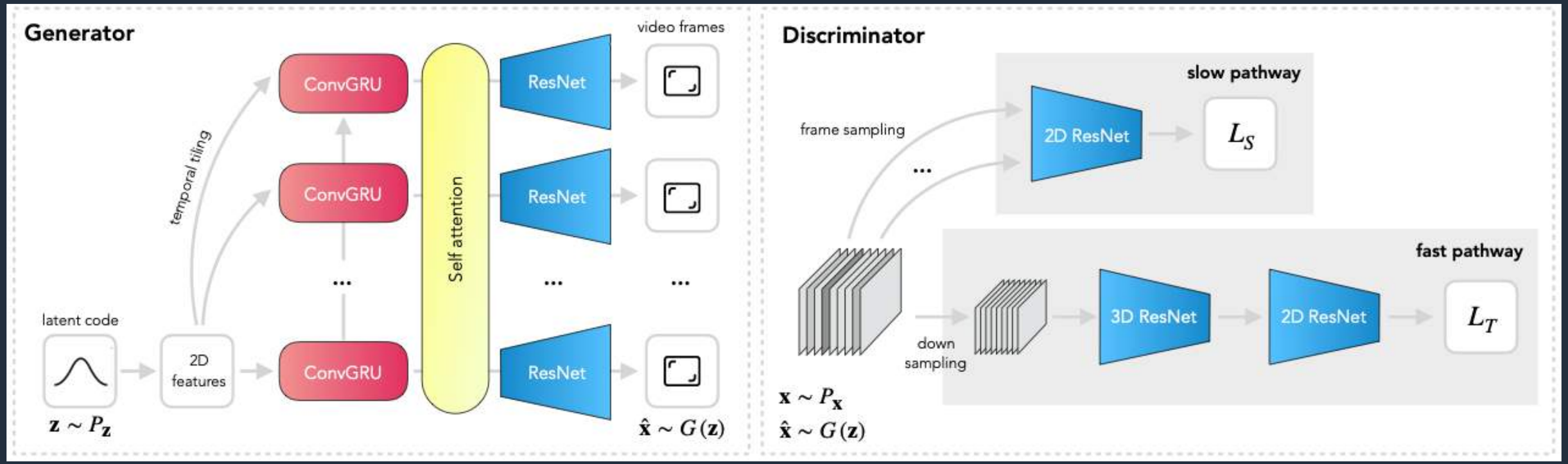
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

BigBiGAN like Setup For Videos



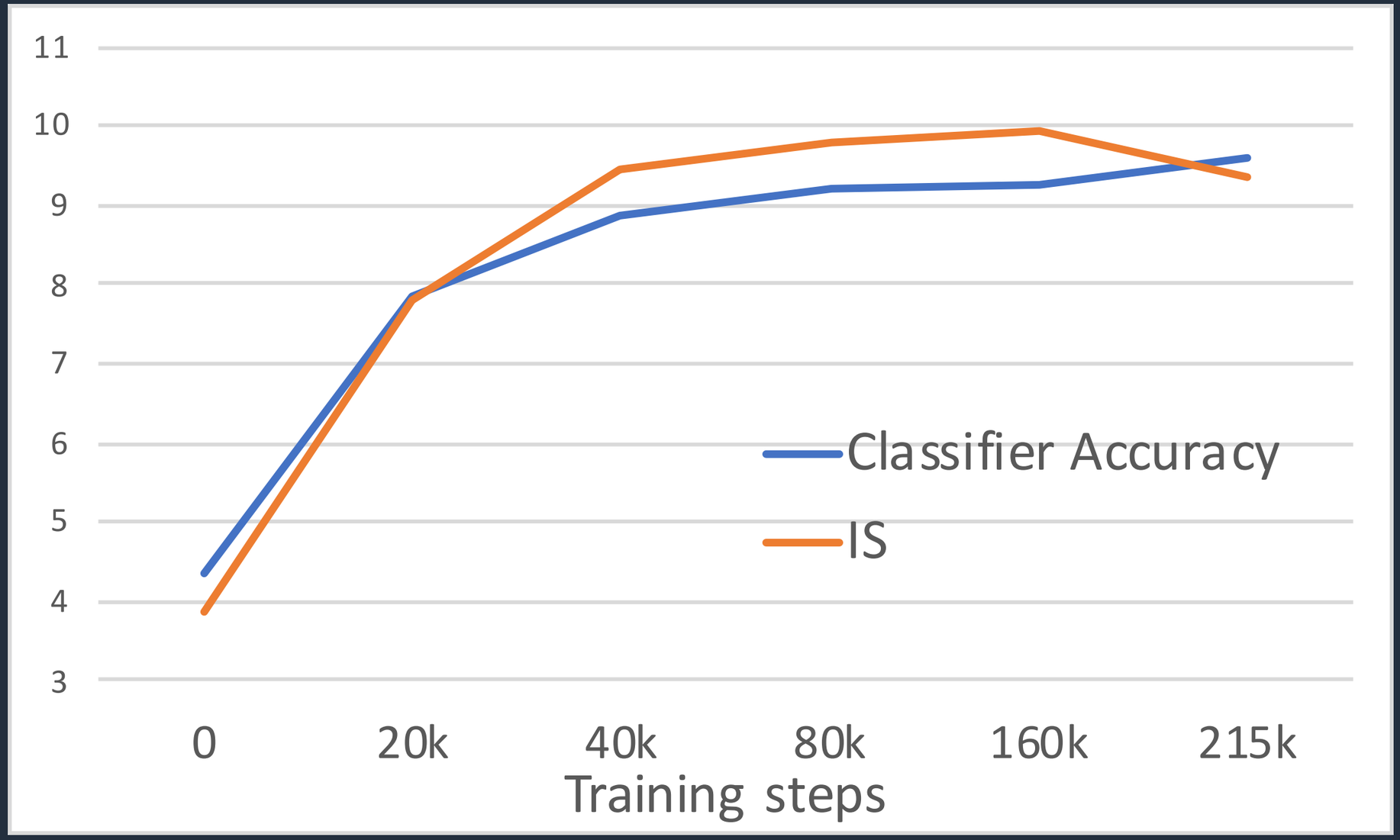
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

BigBiGAN like Setup For Videos



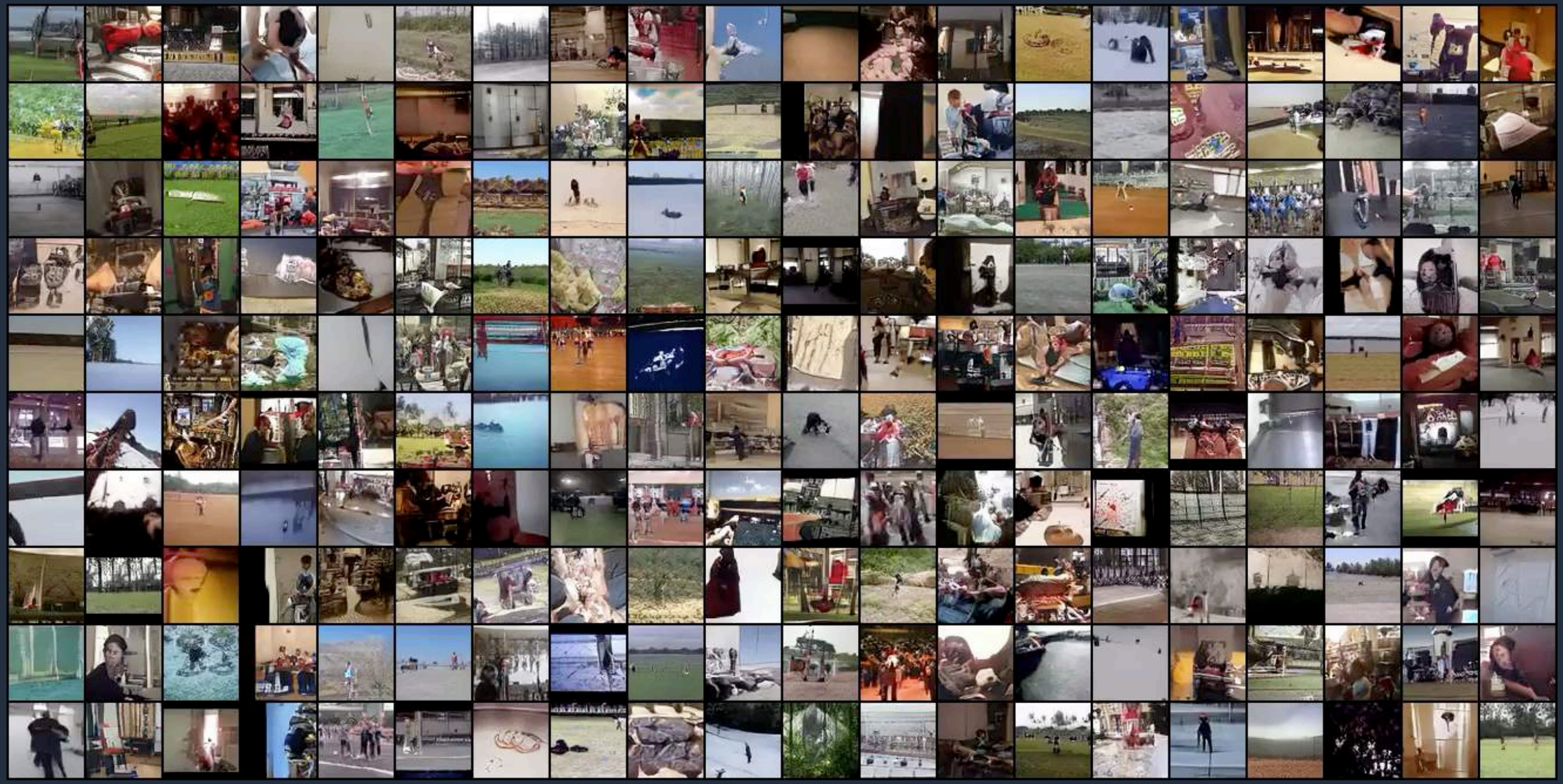
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Embedding Performance



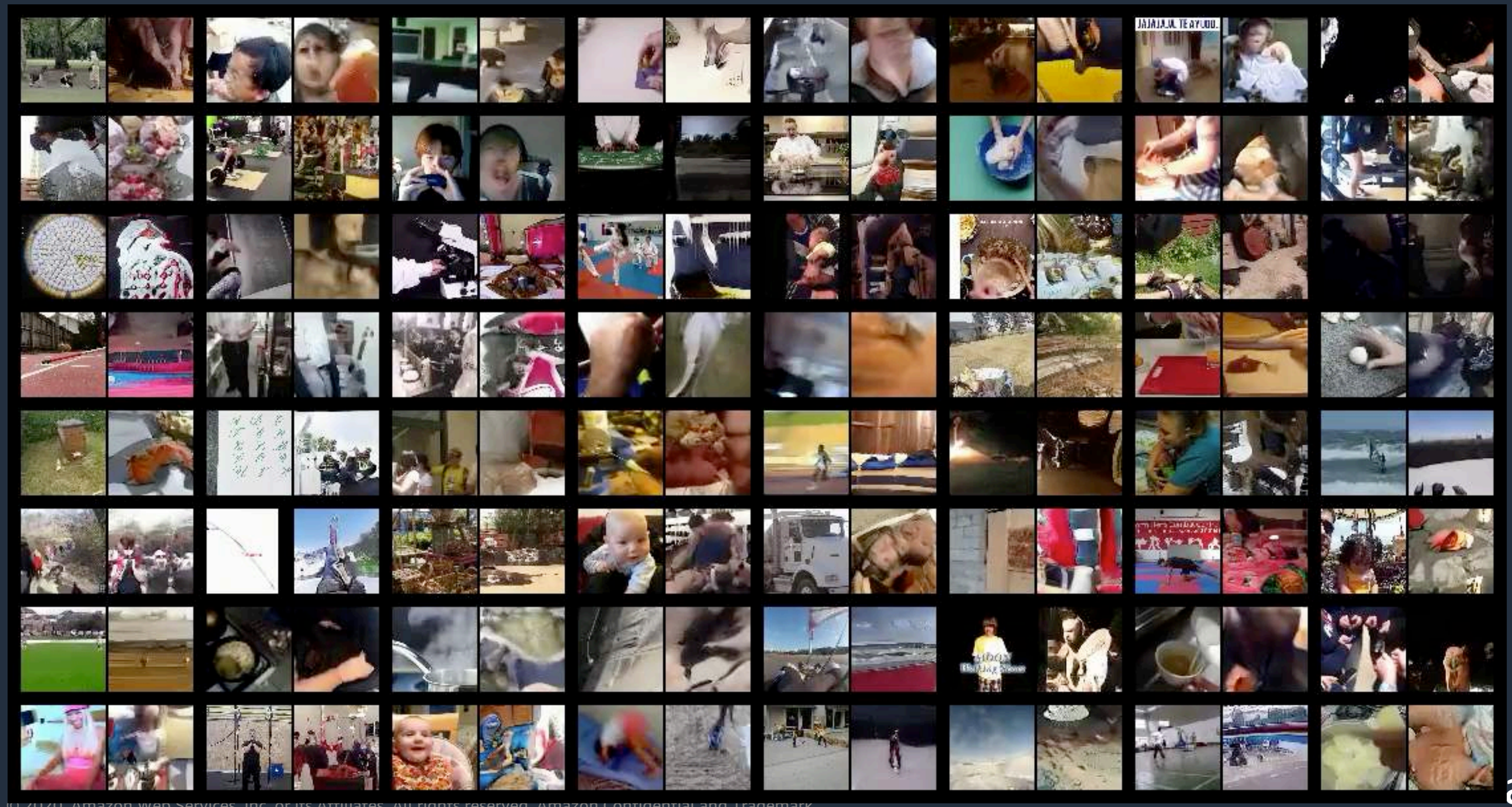
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Generator Examples



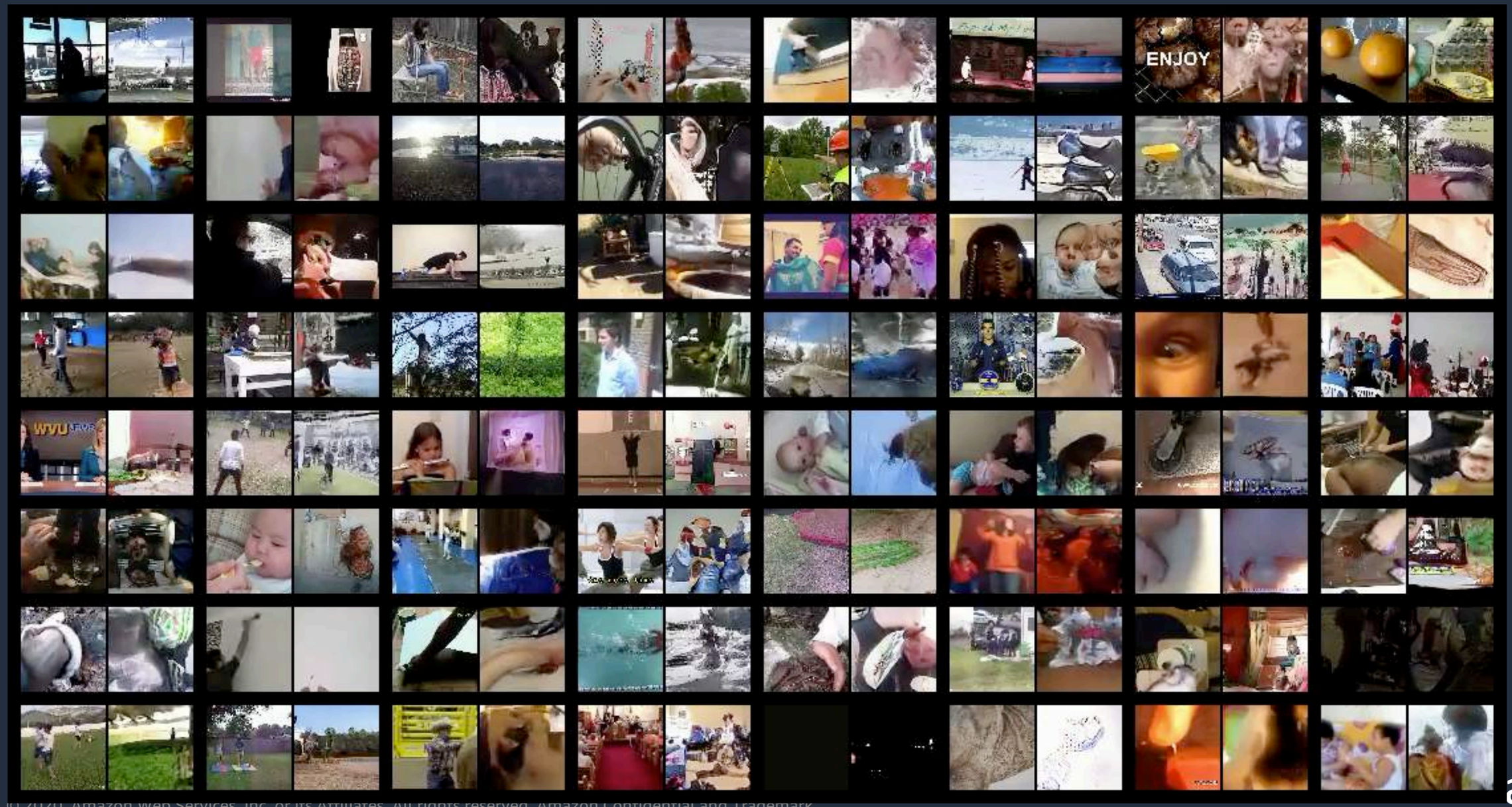
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Cycle Examples



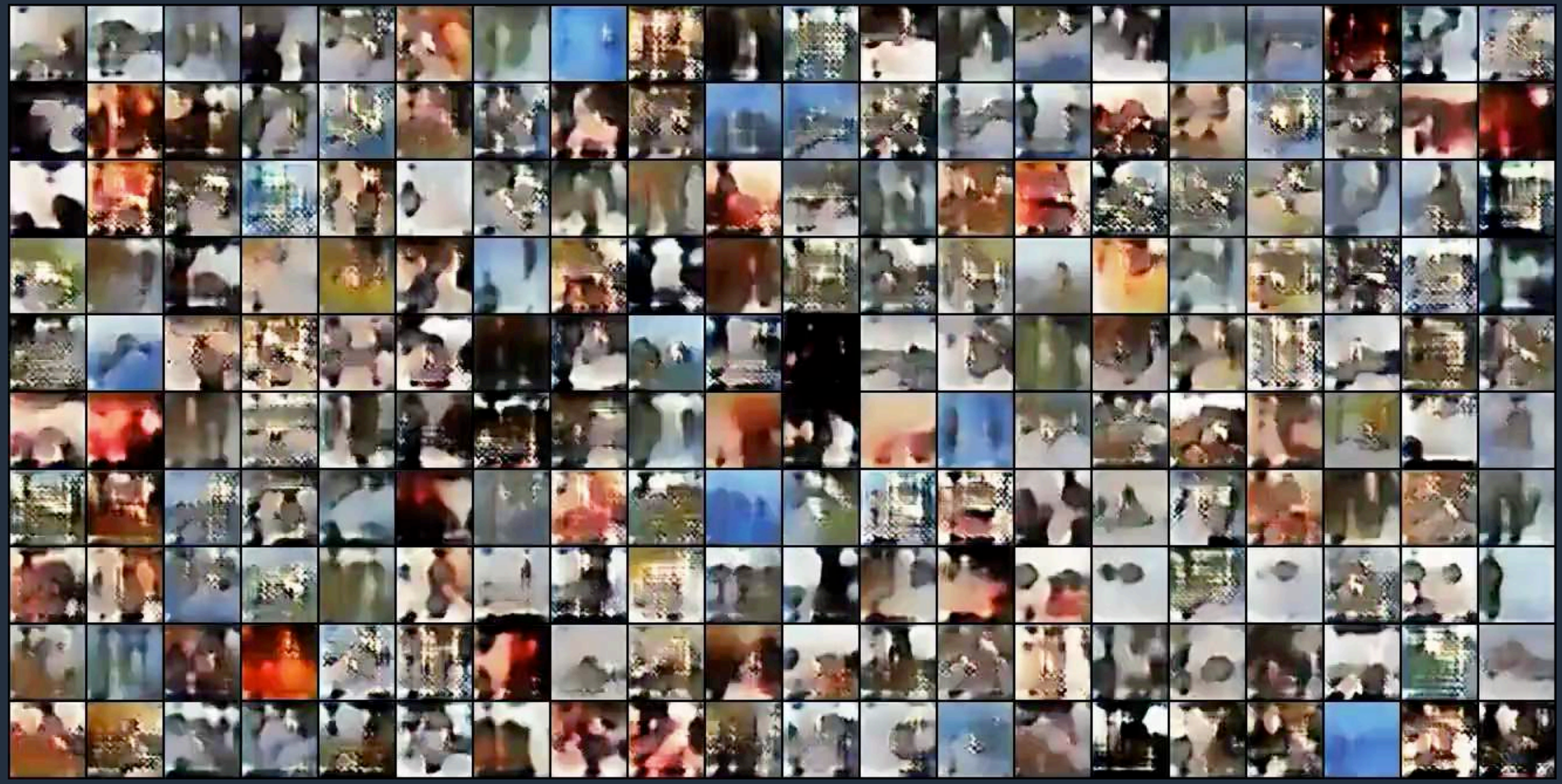
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Cycle Examples



© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

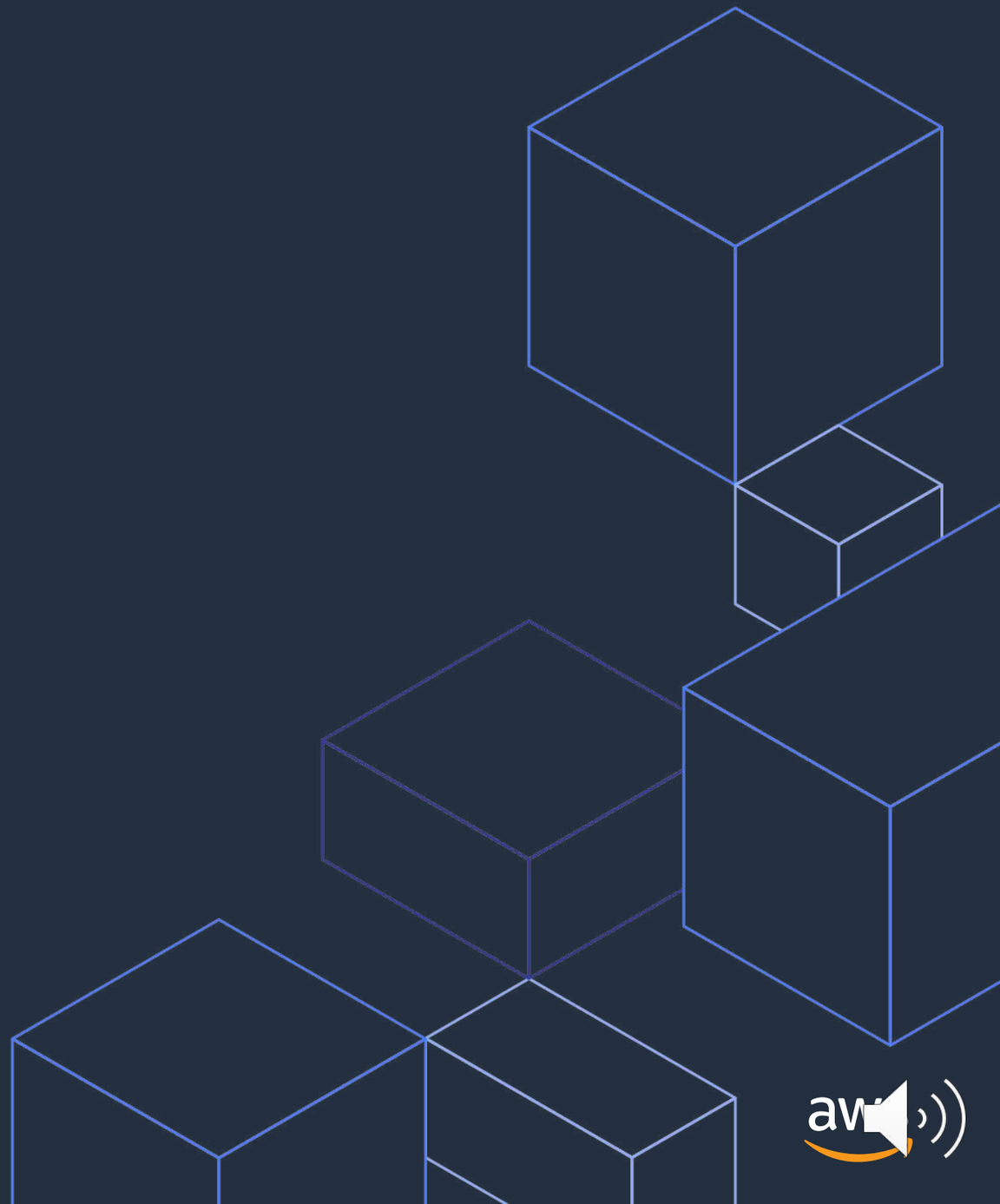
Training Evolution



© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Fast Training and Inference

Chunhui Liu, Xinyu Li, Joseph Tighe



© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

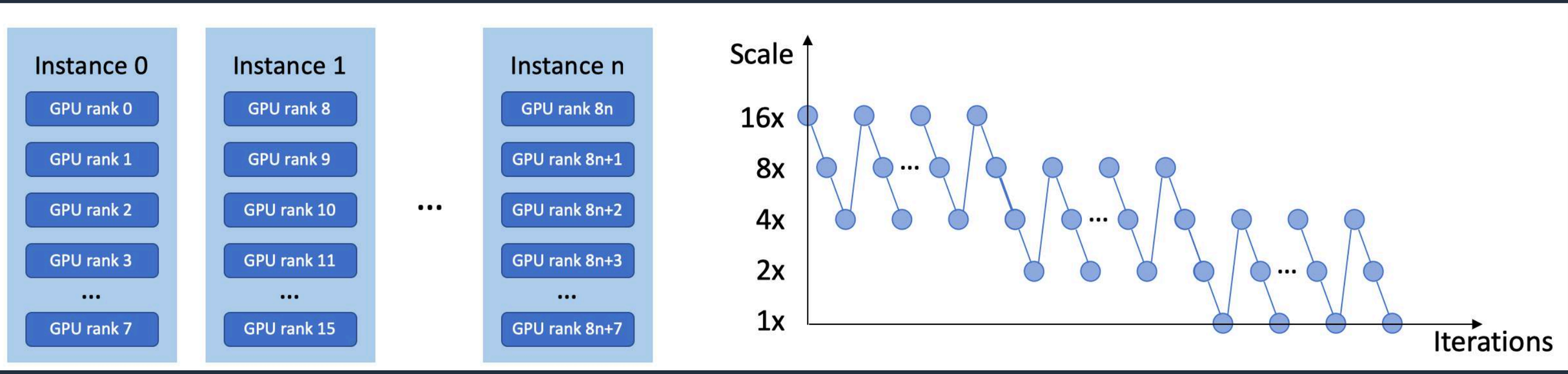
Video Model Fast Training Effort

Cross Instance Distributed Data Parallel (DDP) for Videos

- Easy deployment on multiple AWS instances
- Support both video and image model

Multi Grid Training[1]

- Speedup training by rescaling input size



1 A Multigrid Method for Efficiently Training Video Models
Chao-Yuan Wu, Ross Girshick, Kaiming He, **Christoph Feichtenhofer**, Philipp Krähenbühl CVPR 2020

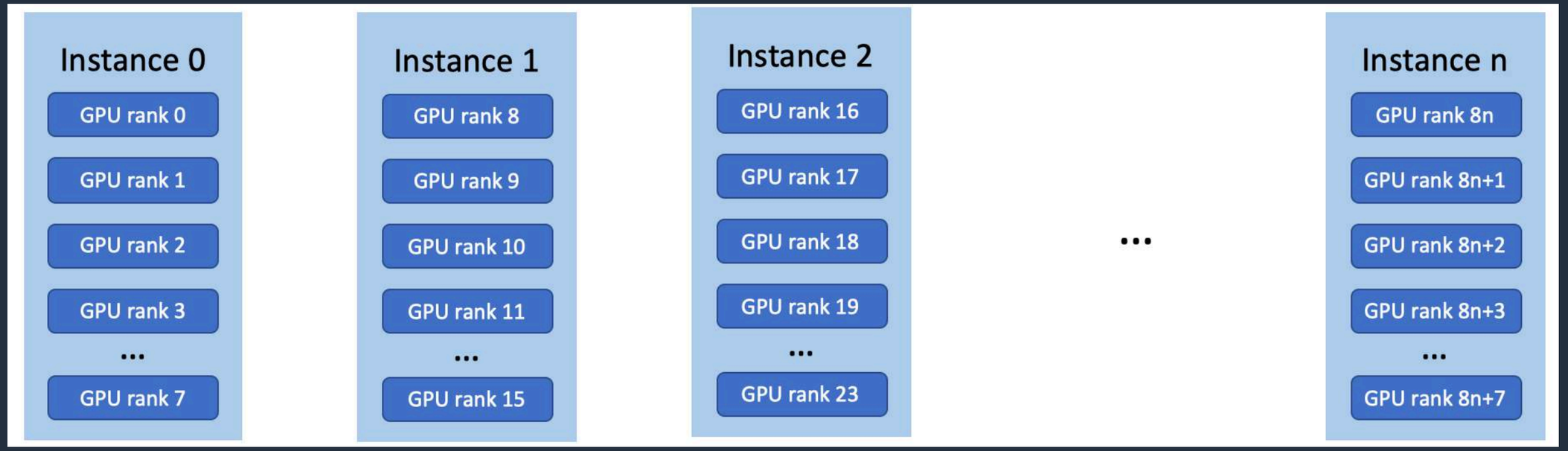
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Cross Instance DDP Training for Videos

Distributed Data Parallel (DDP) make it possible to use multiple instances on AWS for model training.

Each GPU has its own forward-backward process. Then, DDP uses collective functions to synchronize gradients and buffers among GPUs.

The tricky part is to fully balance GPU utilization for running data flow and CPU utilization for loading data. Decord make this possible.



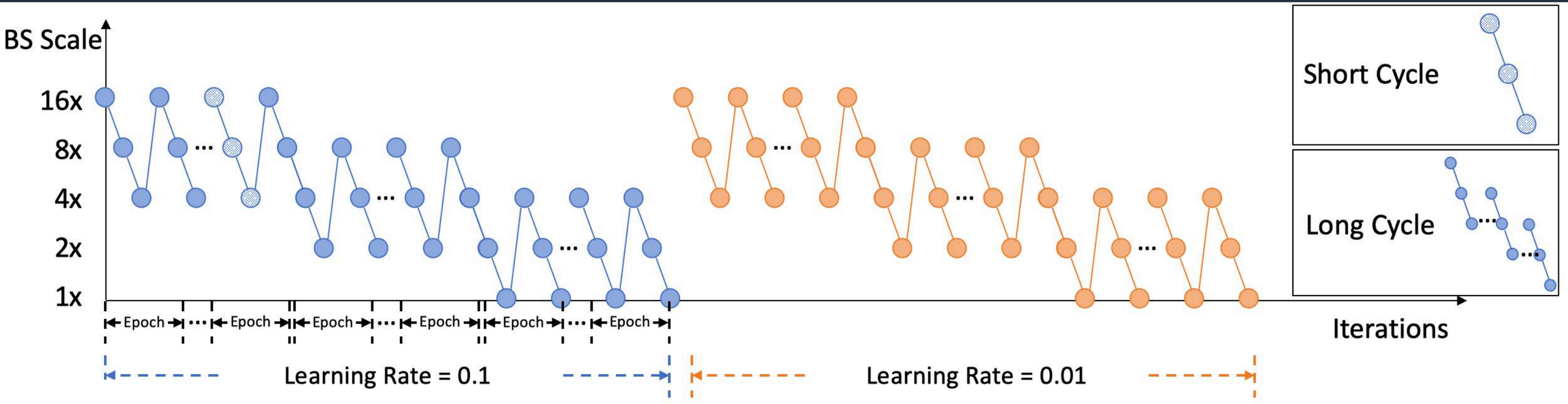
© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Multi Grid Training

Down-sample input data so that we can scale-up batch size, and thus speedup training process.

Short-cycle Scaling Strategy happens for each iteration.

Long-cycle Scaling Strategy happens in each learning rate stage.



© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Video Model Fast Training: A result

Distributed Data Parallel (DDP)

- Test on Kinetics 400 dataset, with 8x p3.16xlarge instances (64x Tesla V100).
- 4 epoch/hour, 12 hour to fully train an I3D model.

Multi Grid Training + DDP

- 3 times speedup using short-cycle only,
- 4 times speedup using long-short cycle.
- **3 hour** to fully train an I3D model.

Important Learnings

- Short-cycle is more stable than long-short cycle.
- Long-short cycle becomes unstable if temporal down-sampling is too aggressive.
- In Multi Grid training, we found that training is more stable without using learning rate - batch scaling rule.
- Single grid finetuning can help with ~0.5% accuracy.

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Video Model Fast Inference Effort

Current Approaches:

- Efficient Neural Architecture Search (NAS)
- Model Compression and Distillation
- Model on Compressed Data

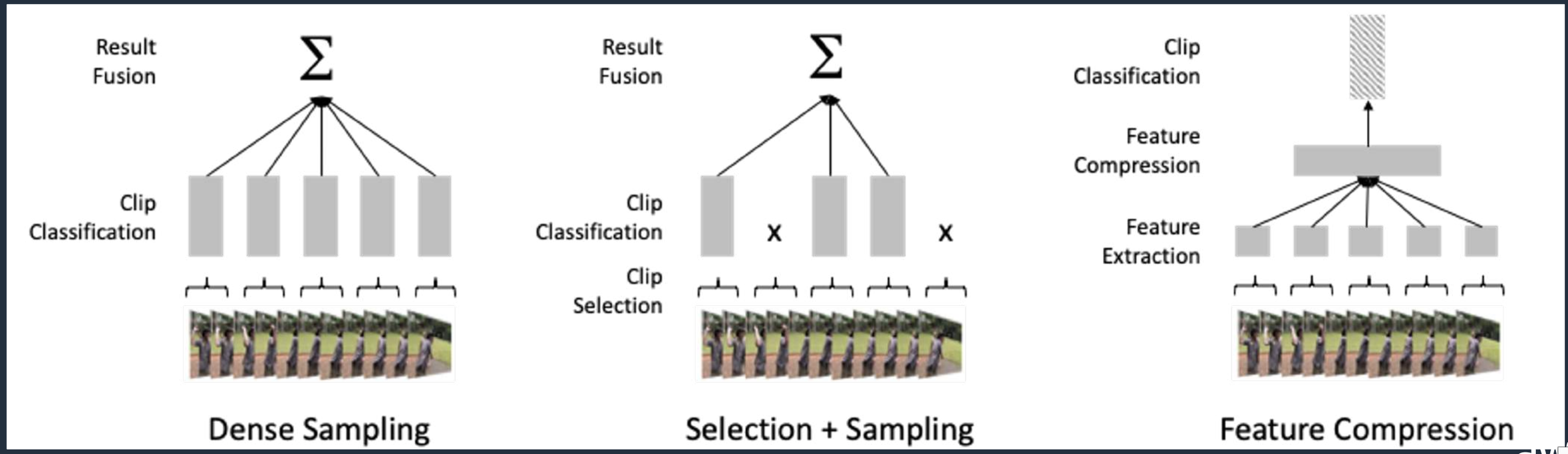
Our effort:

- Single Pass Feature Selection Model for Video Fast Inference (Ongoing)

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Motivation: From Multi Crop to One Pass

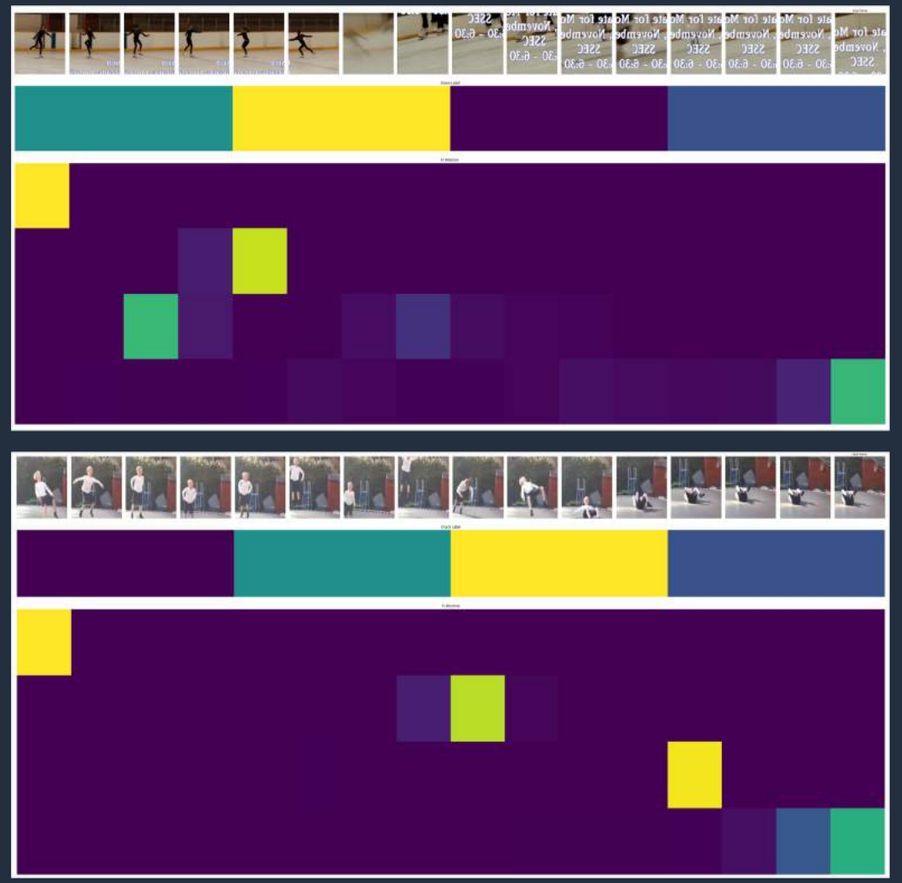
- Most state-of-the-art methods use short clips at training time and thus at test require dense oversampling to achieve high performance numbers.
- We want to avoid dense sampling during inference time and maintains this performance



© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Results on R50 I3D

Method	Top-1 Acc	Top-5 Acc	FLOPS
64x30 I3D baseline	73.8	91.0	998G
256 frame randomly choose one 64 clip	62.3	78.38	33G
256 frame forward	69.2	88.2	133G
256 frame Feature Selection	71.83	89.6	154G



© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

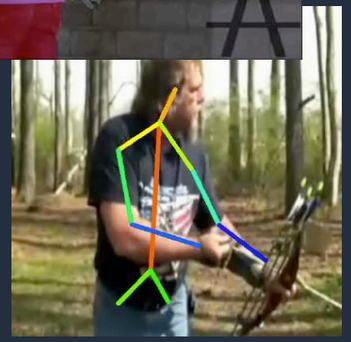
Future of Video Research

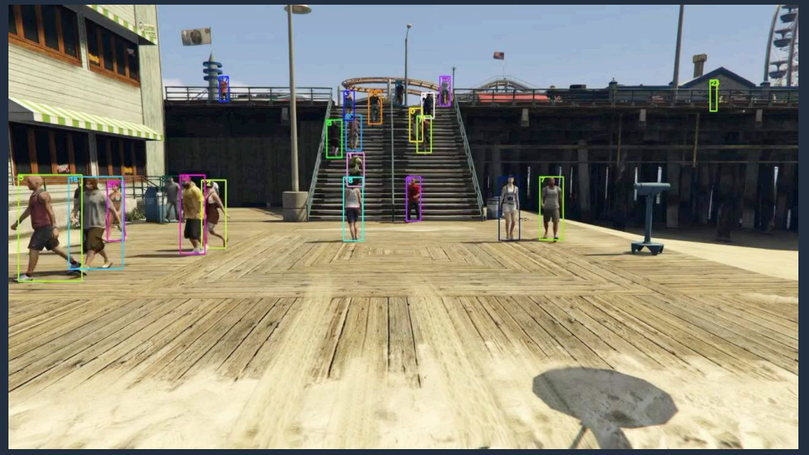
- Datasets

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Future of Video Research

- Datasets
- Tasks



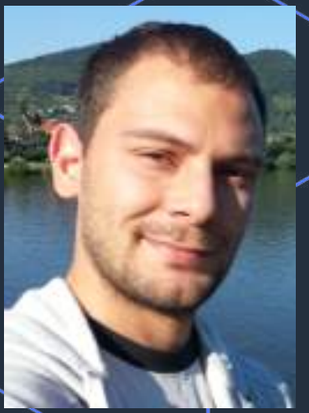


© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.

Thanks

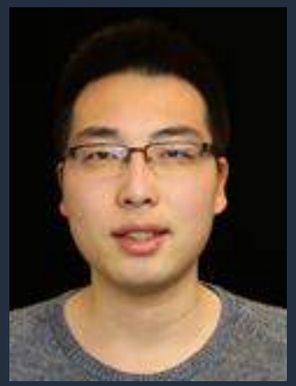













© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and

Links to Work Presented

This video includes descriptions of the following work:

Combining detection and tracking for human pose estimation in videos: https://assets.amazon.science/9c/68/ce3ec91b41c1b6a20ee1e793709d/scipub-1326.pdf

Multi-Object Tracking with Siamese Track-RCNN: https://arxiv.org/pdf/2004.07786

Action Recognition With Spatial-Temporal Discriminative Filter Banks: http://openaccess.thecvf.com/content_ICCV_2019/papers/Martinez_Action_Recognition_With_Spatial-Temporal_Discriminative_Filter_Banks_ICCV_2019_paper.pdf

Rethinking Zero-shot Video Classification: End-to-end Training for Realistic Applications: https://assets.amazon.science/b3/28/0702b5ec441aaadcb79040b58128/scipub-1328.pdf

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.