Learning to Anticipate

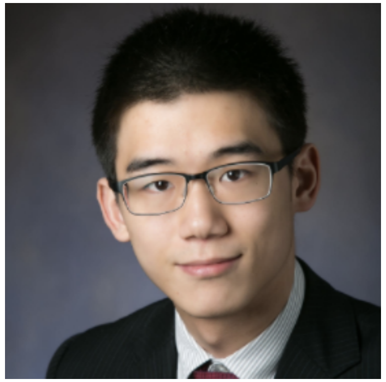
Alex Schwing

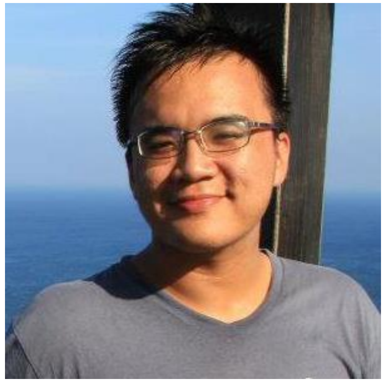
University of Illinois at Urbana Champaign

September 23, 2019

Magicians









cisco

AI – The Next Decade

Computer Vision and Machine Learning Success

Models which analyze work very well

- Image classification
- Object detection
- Semantic segmentation
- Human pose estimation

Our quest to personalize life

Models which anticipate

Combine the vision of David Marr:

"2D to 3D reasoning"

and Rodolfo Llinas, Kenneth Craik:

"a creature must anticipate outcome of movement to navigate safely"

Major Ingredients to Anticipate

- Interaction reasoning
- Revealing priors
- Holistic object understanding
- Capturing ambiguity

Major Ingredients to Anticipate

- Interaction reasoning
- Revealing priors
- **Holistic object understanding**
- Modeling ambiguity

Instance Level Video Object Segmentation

Weakly Supervised Setting

- Given objects outlined in first frame
- Predict object contours in subsequent frames

Classical Deep-Net Approaches:

- Train a classifier using first frame data
- Run on remaining frames

Concern: training at runtime is slow

VideoMatch: Matching based Video Object Segmentation

Goals:

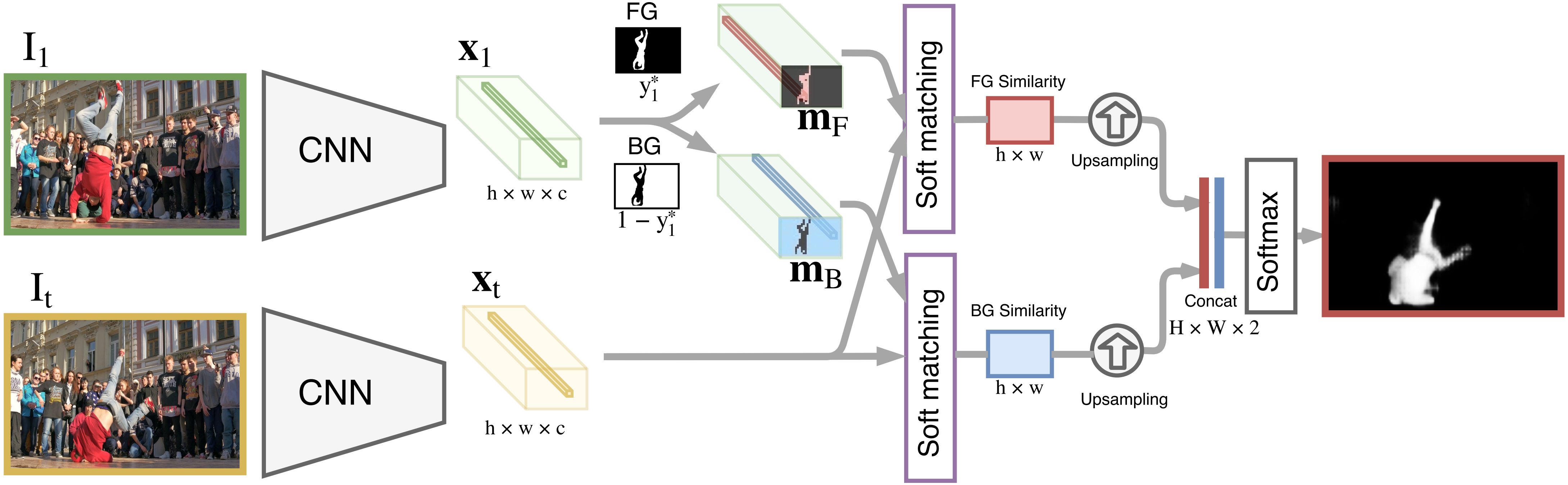
- Efficient algorithm that does not require fine-tuning
- Combination of detection and tracking
- Implicit extraction of temporal information

See also: Voigtlaender et al. 2019, Vondrick et al. 2018

Joint with
Yuan-Ting Hu, Jia-Bin Huang
ECCV 2018

VideoMatch Approach

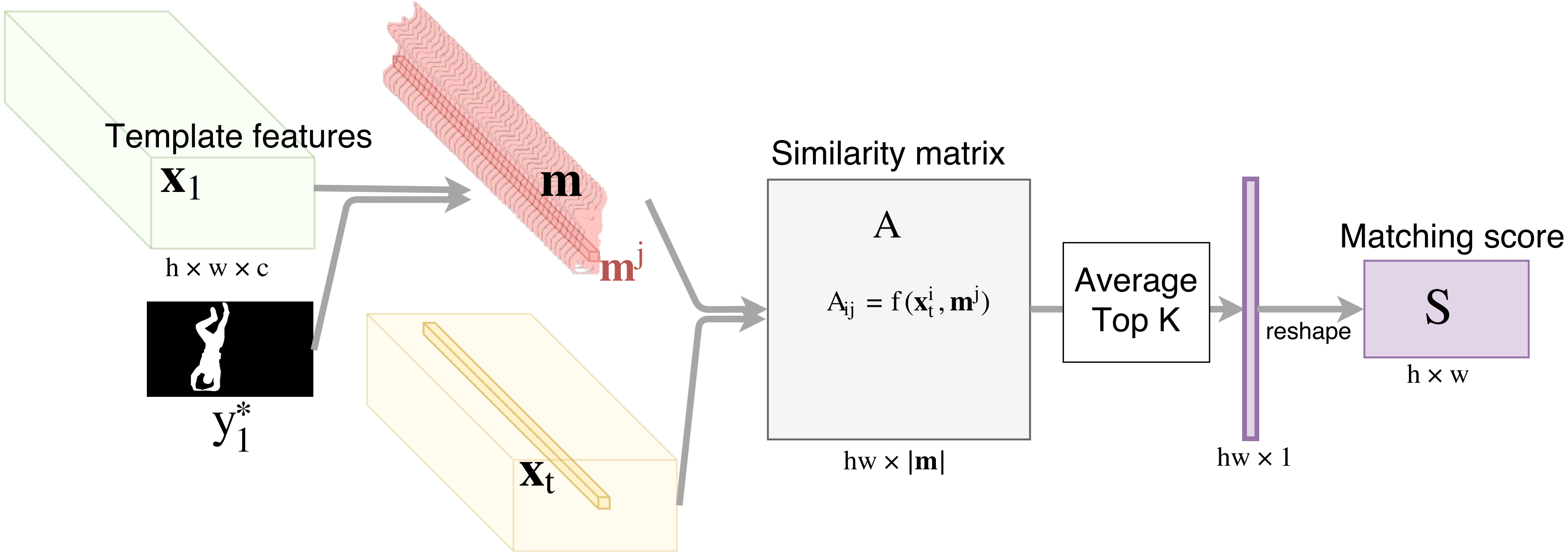
Idea: Learning to match feature representations



Joint with
Yuan-Ting Hu, Jia-Bin Huang
ECCV 2018

VideoMatch Softmatching

Idea: Learning to match feature representations



Joint with
Yuan-Ting Hu, Jia-Bin Huang
ECCV 2018

VideoMatch Extensions

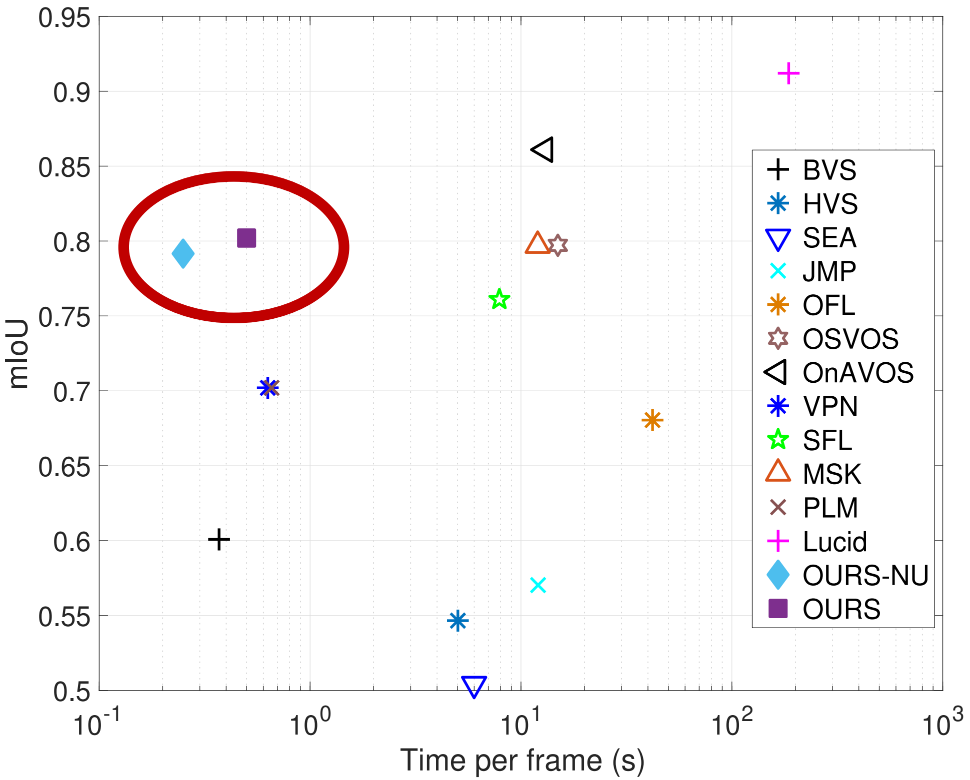
- Online Update: augment foreground and background sets
- Outlier Removal

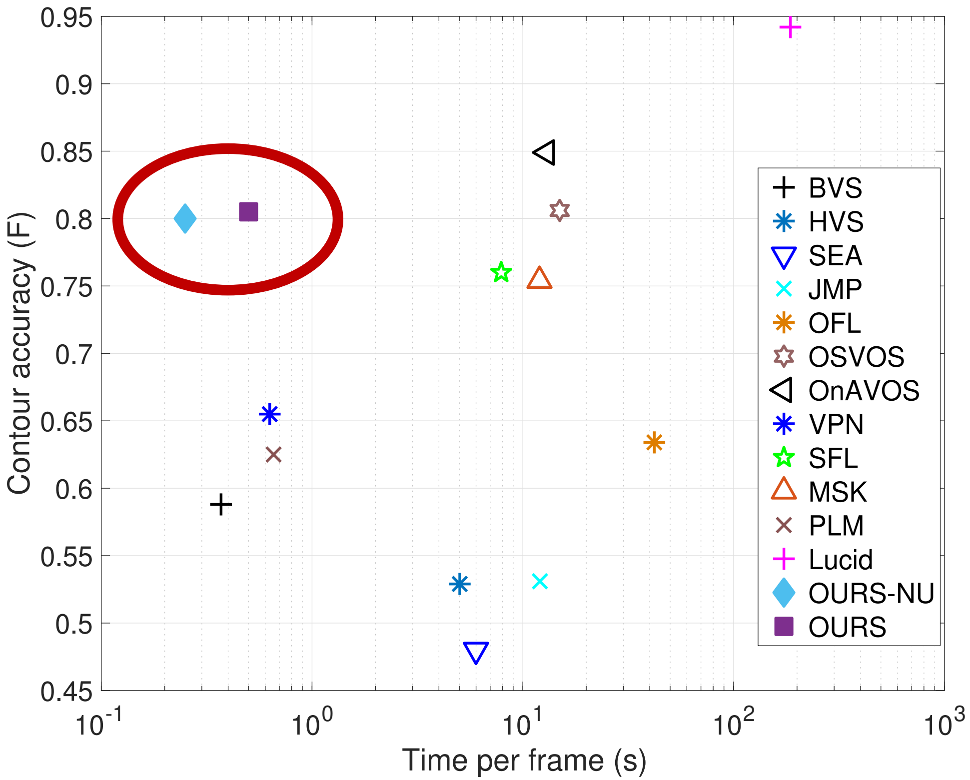
(a) FG pred. $y_{t,\text{init}}$ (b) FG pred. y_{t-1} (c) Extruded pred. $\hat{y}_{t-1}$ (d) Output pred. y_t

Joint with
Yuan-Ting Hu, Jia-Bin Huang
ECCV 2018

VideoMatch Results

Results (DAVIS-16 Validation)





Joint with
Yuan-Ting Hu, Jia-Bin Huang
ECCV 2018

VideoMatch Results

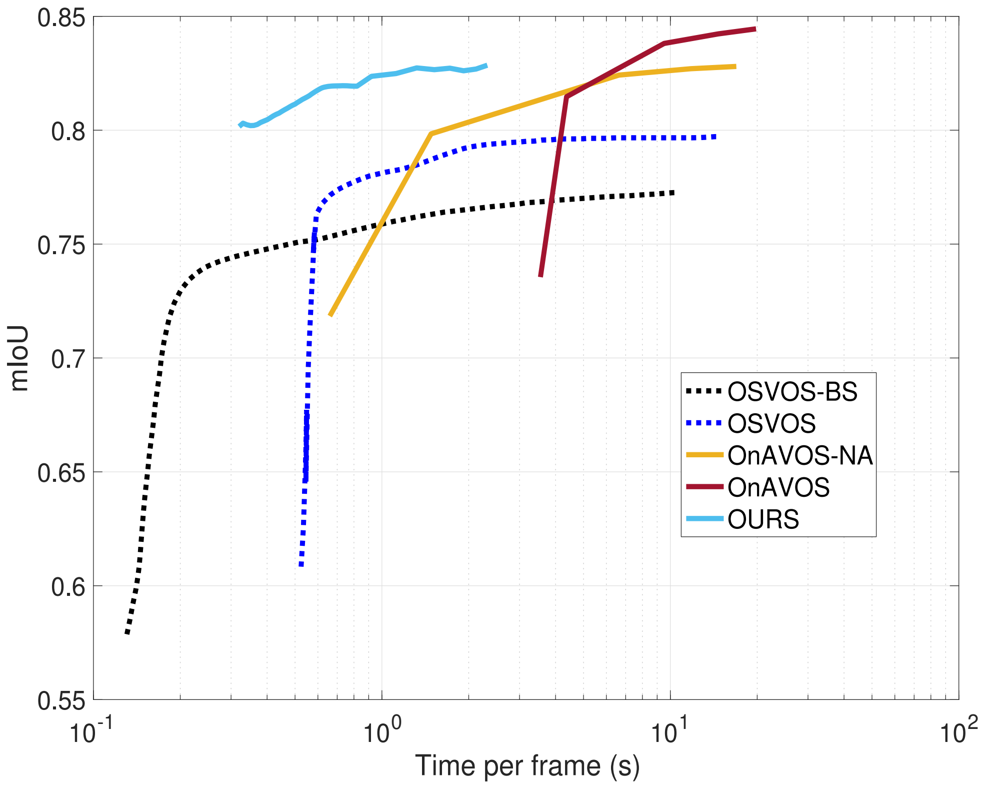
Results (DAVIS-16 Validation)

Outlier removal	BG update	FG update	mIoU
-	-	-	0.792
✓	-	-	0.796
✓	✓	-	0.799
✓	✓	✓	0.802

Joint with
Yuan-Ting Hu, Jia-Bin Huang
ECCV 2018

VideoMatch Results

Results (DAVIS-16 Validation)



Joint with
Yuan-Ting Hu, Jia-Bin Huang
ECCV 2018

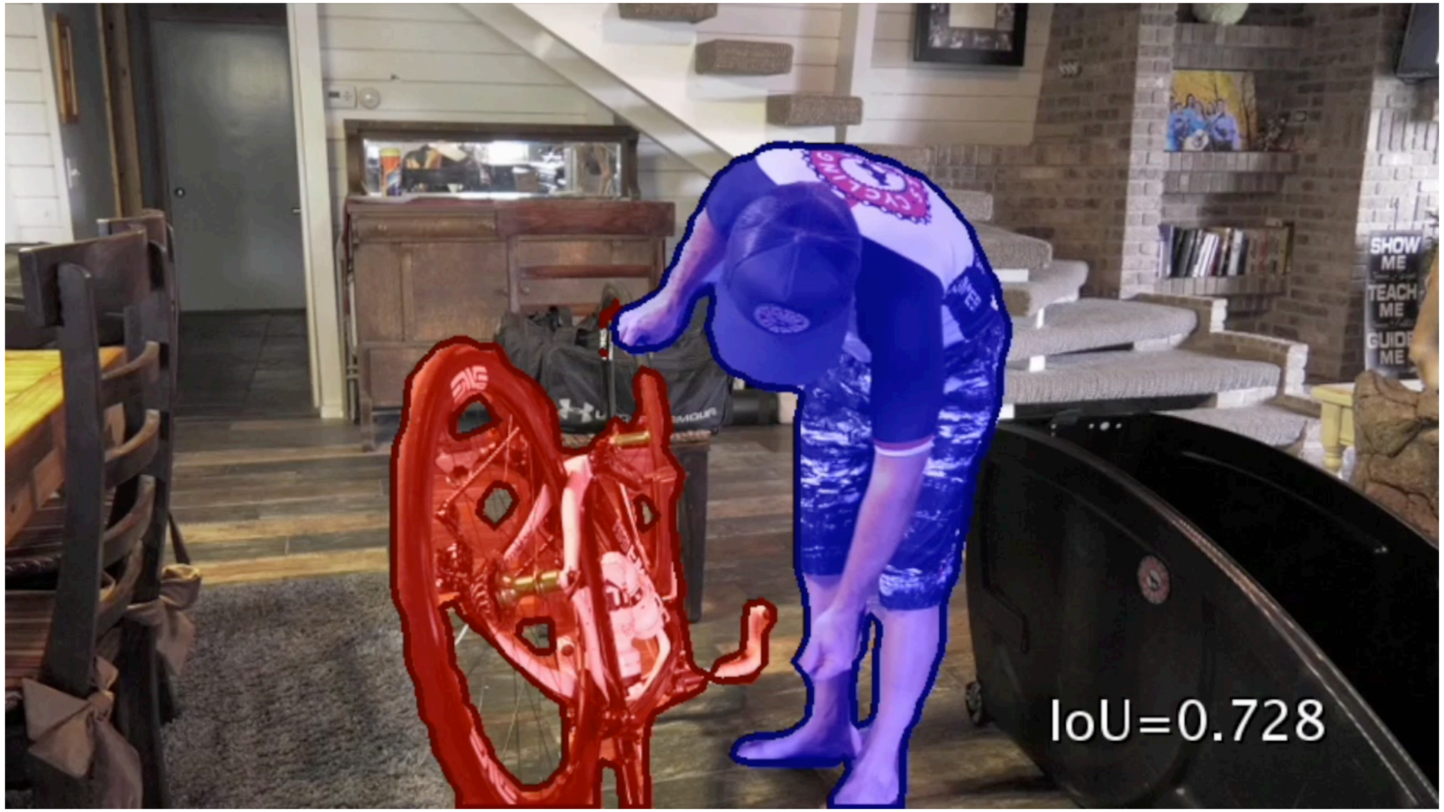
VideoMatch Results

Qualitative Results (Davis 2016)

Joint with
Yuan-Ting Hu, Jia-Bin Huang
ECCV 2018

VideoMatch Results

Qualitative Results (Davis 2017)



Joint with
Yuan-Ting Hu, Jia-Bin Huang
ECCV 2018

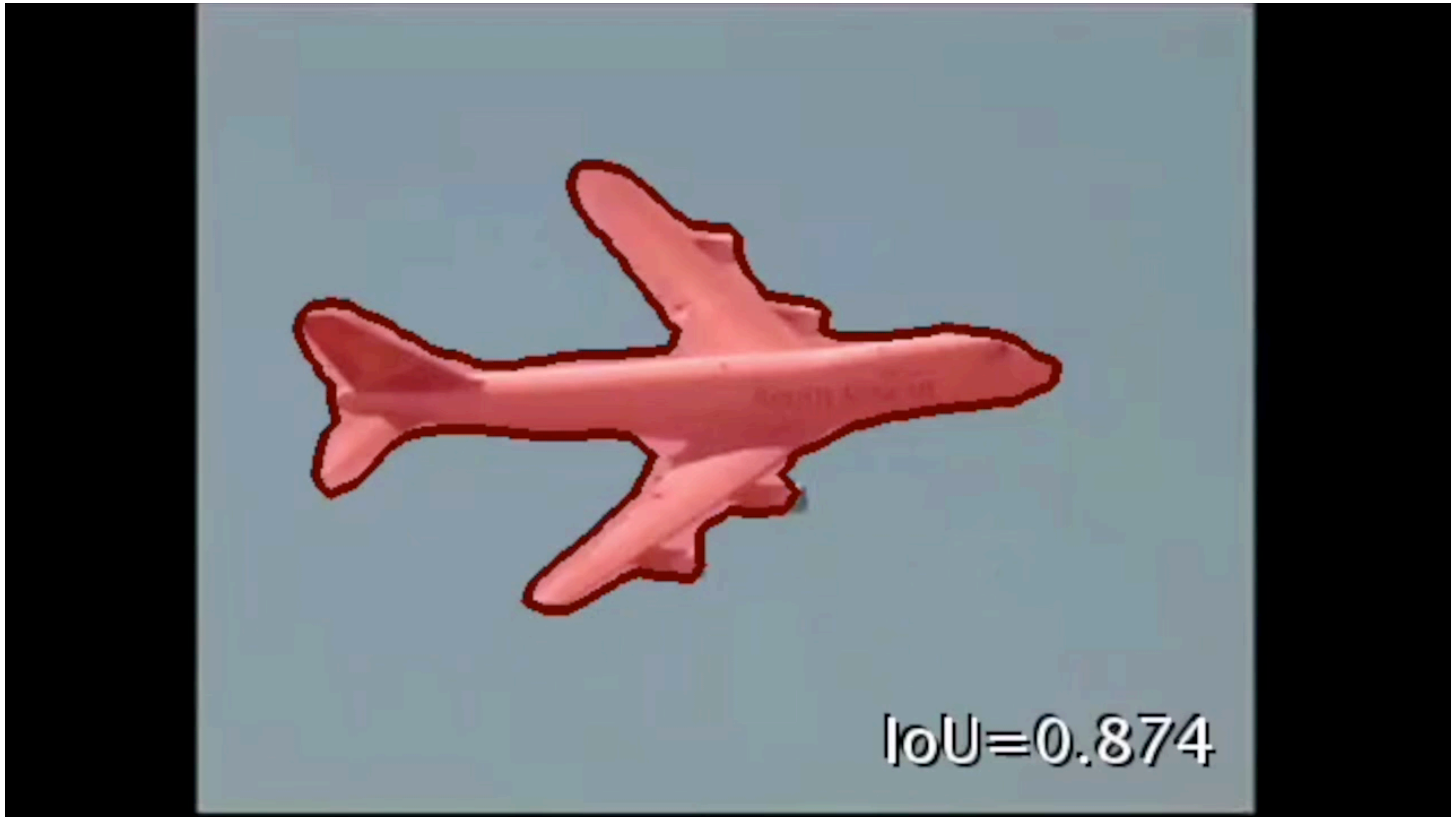
VideoMatch Results

Qualitative Results (Jumpcut - Trained on Davis 17)

Joint with
Yuan-Ting Hu, Jia-Bin Huang
ECCV 2018

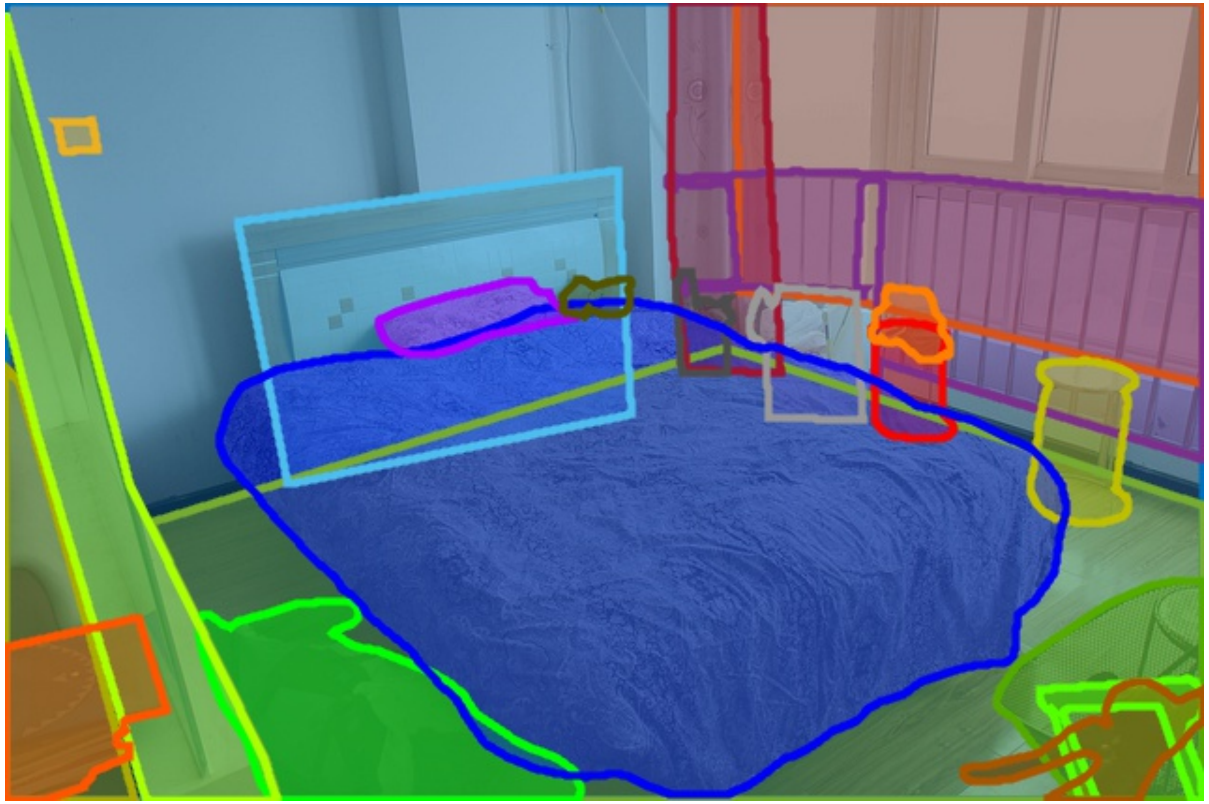
VideoMatch Results

Qualitative Results (YouTube-Objects – Trained on Davis 2017)



Amodal Segmentation

- Recognizing the full extent of the object



Amodal Segmentation

- Recognizing the full extent of the object

Difficulties

- Humans are capable of amodal segmentation
- A challenging task for AI systems
  - Occlusion reasoning
  - Predicting the invisible part
  - Expensive to get the groundtruth

Current Amodal Segmentation Datasets

- COCOA

Real data

A subset of MS-COCO dataset

- D2S

Real data

Groceries on the table

- DYCE

Synthetic data

Indoor static scene

Current Amodal Segmentation Datasets

- COCOA

Real data

A subset of MS-COCO dataset

- D2S

Real data

Groceries on the table

- DYCE

Synthetic data

Indoor static scene

All of them are image datasets

Temporal Information is Missing!

- Temporal context helps to predict amodal segmentation

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

A Dataset for Amodal Video Segmentation

- A synthetic dataset using Grand Theft Auto V (GTA-V)
  - Realistic rendering
  - Various scenarios
  - Different weather/lighting condition
  - Groundtruth annotations from the game



Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

A Dataset for Amodal Video Segmentation

- A synthetic dataset using Grand Theft Auto V (GTA-V)
  - Realistic rendering
  - Various scenarios
  - Different weather/lighting condition
  - Groundtruth annotations from the game



Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

A Dataset for Amodal Video Segmentation

- A synthetic dataset using Grand Theft Auto V (GTA-V)
  - Realistic rendering
  - Various scenarios
  - Different weather/lighting conditions
  - Groundtruth annotations from the game



Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

A Dataset for Amodal Video Segmentation

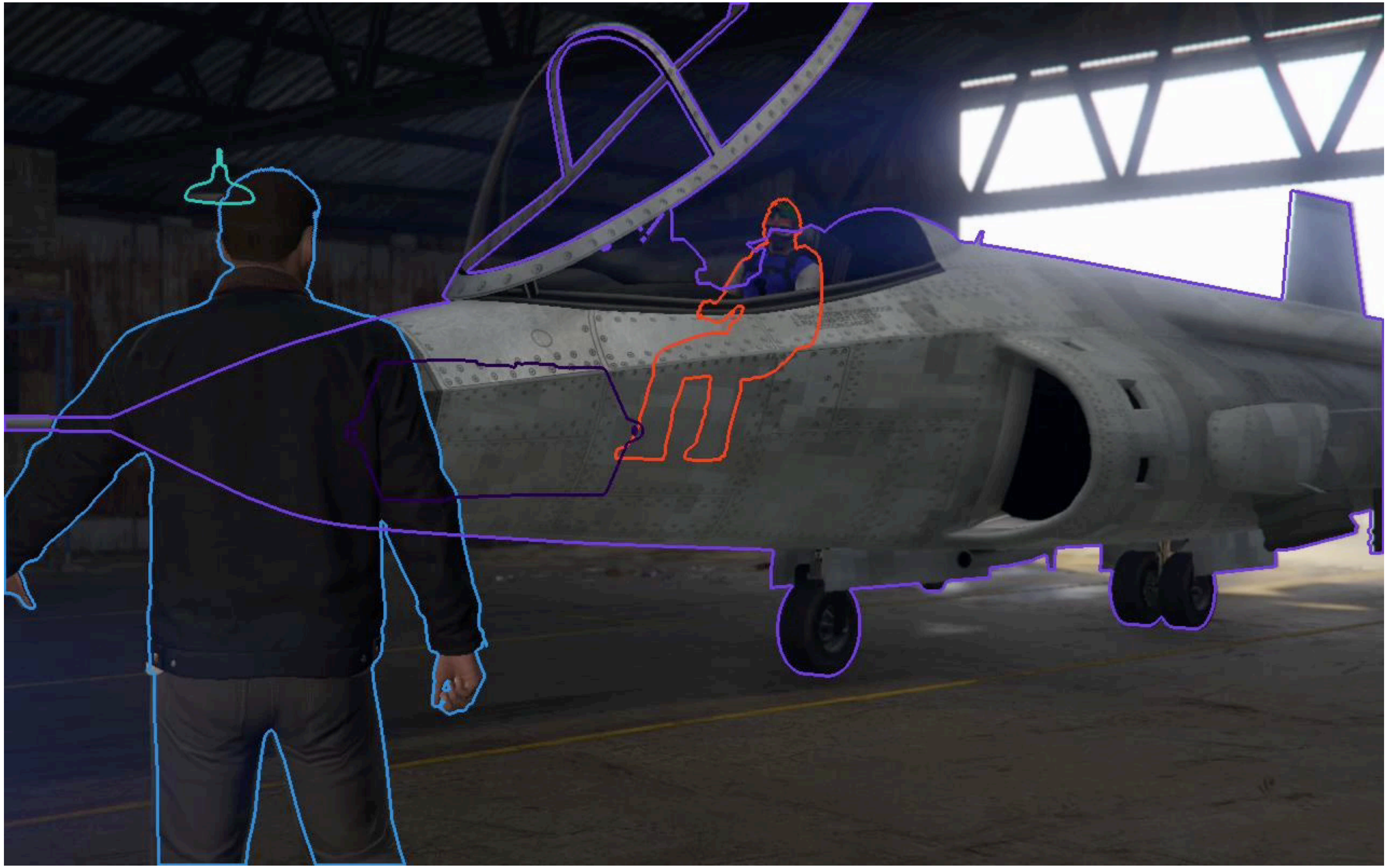
- A synthetic dataset using Grand Theft Auto V (GTA-V)
  - Realistic rendering
  - Various scenarios
  - Different weather/lighting condition
  - Groundtruth annotations from the game

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Example Video

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Example Video



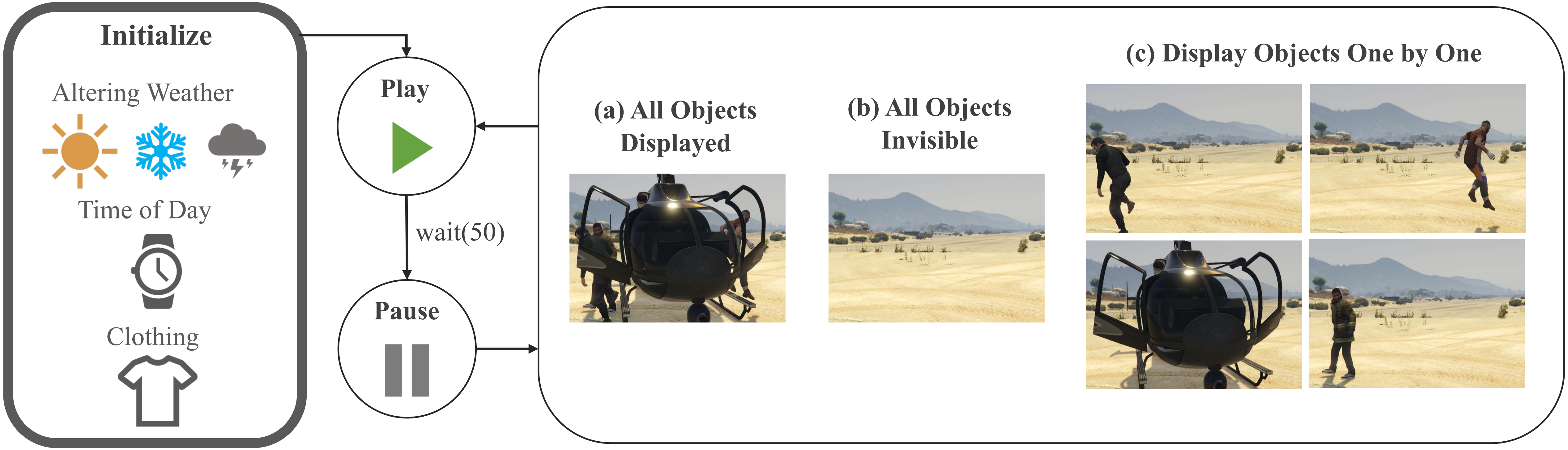
Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Controlling the Game

- We use ScriptHook V to control the game
  - Altering the weather, time of day and clothing
  - Pausing the game
  - Toggling the visibility of objects

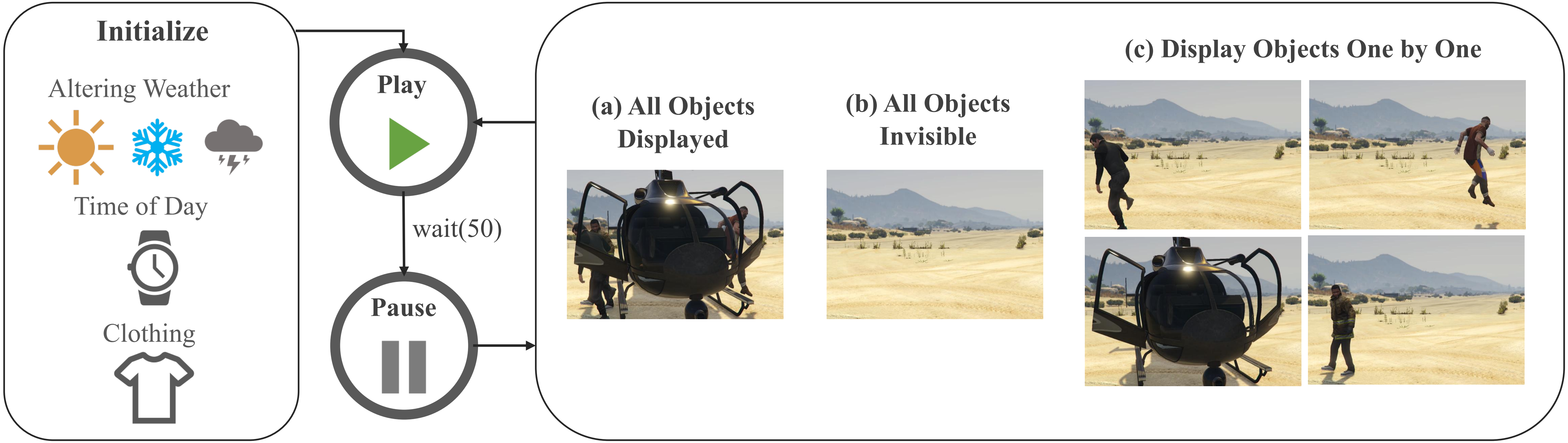
Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Dataset Collection Pipeline



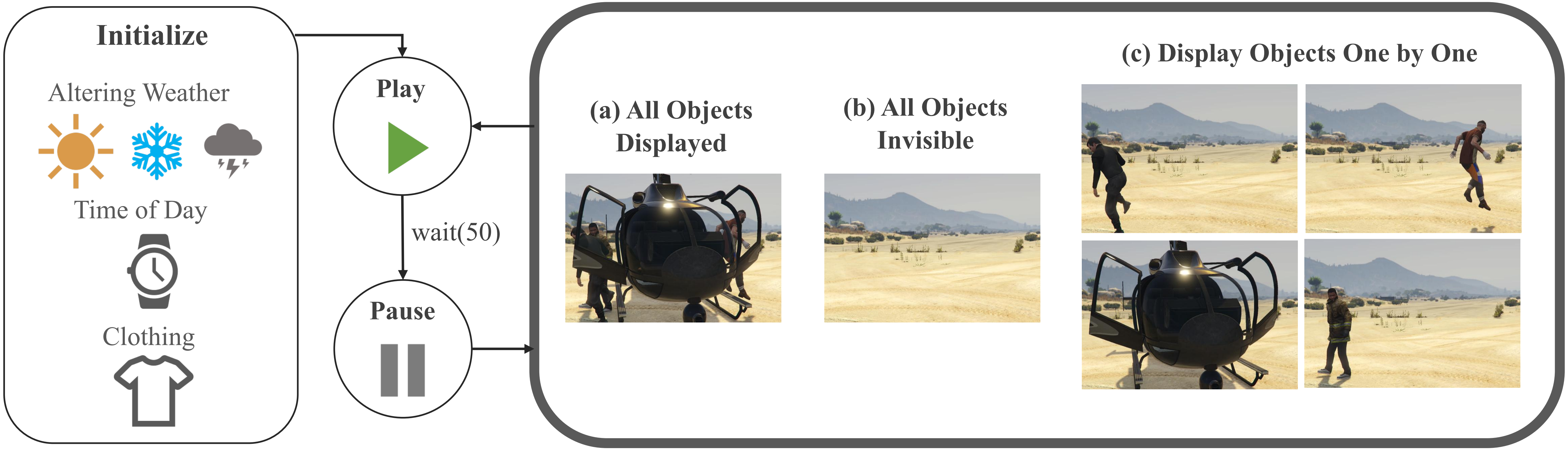
Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Dataset Collection Pipeline



Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Dataset Collection Pipeline



Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

How to Compute the Amodal Segmentation

- Comparing the RGB pixels wouldn't be robust enough due to rendering

All Objects Displayed

All Objects Invisible

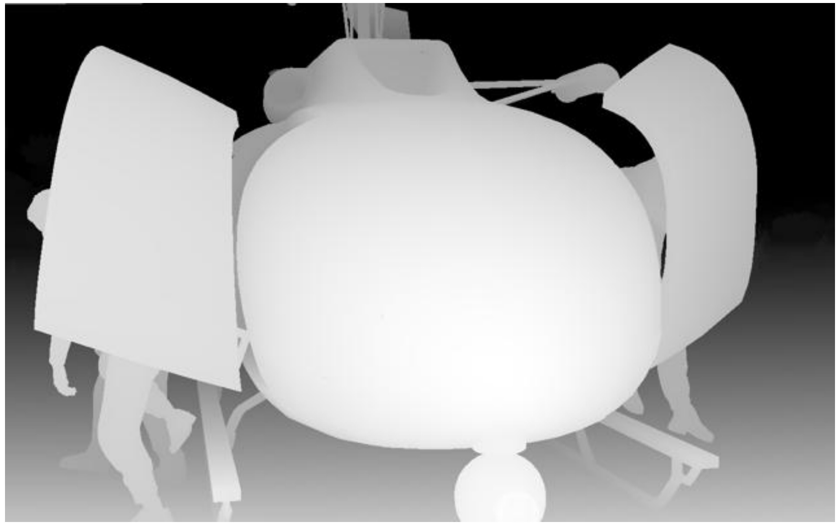
Display Objects One by One

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Depth Buffer and Stencil Buffer

- Along with the RGB images, we also record the corresponding depth buffer and stencil buffer by hooking into DirectX functions.
- All objects displayed

RGB image



Depth buffer

Stencil buffer

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Depth Buffer and Stencil Buffer

- Along with the RGB images, we also record the corresponding depth buffer and stencil buffer by hooking into DirectX functions.
- Background

RGB image

Depth buffer

Stencil buffer

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Depth Buffer and Stencil Buffer

- Along with the RGB images, we also record the corresponding depth buffer and stencil buffer by hooking into DirectX functions.
- One object displayed

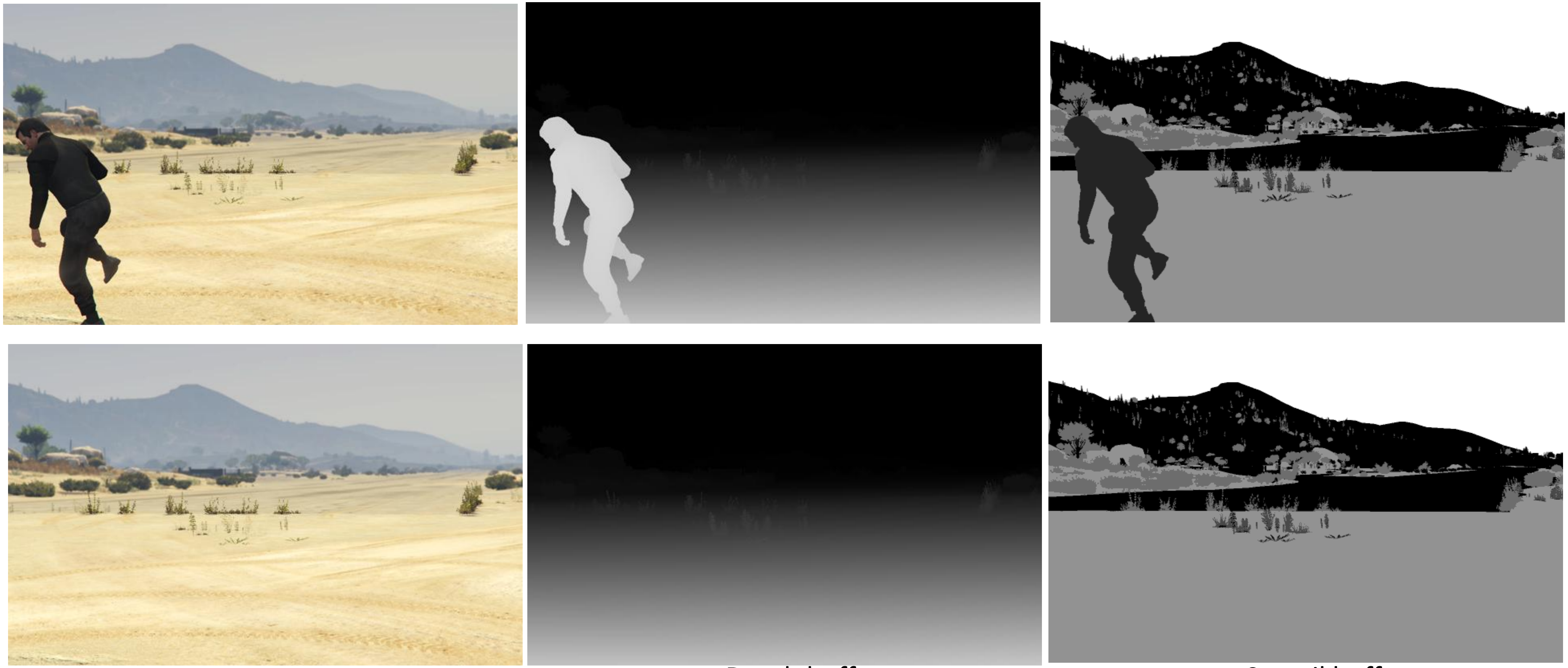
RGB image

Depth buffer

Stencil buffer

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Computing the Amodal Segmentation



RGB image | Depth buffer | Stencil buffer

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Computing the Amodal Segmentation

RGB image

Amodal segmentation

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Amodal Segmentation

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Visible Mask

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Tracking Objects

- The game assigns a unique ID to each object
- We can track the objects based on the IDs

Amodal

Visible

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Semantic Class Labels

- We are able to obtain the name of the 3D model of each object
- We merge the objects with similar names into **162** classes
- 60% of the classes in MS-COCO can be found in the proposed dataset

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Pose Information

- Amodal 2D/3D pose information for human


DO NOT
ENTER

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Dataset Statistics

Dataset	COCOA	COCOA-cls	D2S	DYCE	Ours
Image/Video	Image	Image	Image	Image	Video
Resolution	275K pix	275K pix	3M pix	1M pix	1M pix
	-	-	1440×1920	1000×1000	800×1280
Synthetic/Real	Real	Real	Real	Synthetic	Synthetic
# of images	5,073	3499	5,600	5,500	111,654
# of classes	-	80	60	79	162
# of instances	46,314	10,562	28,720	85,975	1,896,295
# of occluded instances	28,106	5,175	16,337	70,766	1,653,980
Avg. occlusion rate	18.8%	10.7%	15.0%	27.7%	56.3%

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Baselines

- MaskRCNN
- MaskAmodal: a variant of MaskRCNN predicting the amodal mask
- MaskJoint: jointly predicting modal and amodal masks

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Baselines

- MaskRCNN
- MaskAmodal: a variant of MaskRCNN predicting the amodal mask
- MaskJoint: jointly predicting modal and amodal masks using two output heads

	Modal mask							Amodal mask						
	AP_{50}	AP	AP^P_{50}	AP^H_{50}	AP^L_{50}	AP^M_{50}	AP^S_{50}	AP_{50}	AP	AP^P_{50}	AP^H_{50}	AP^L_{50}	AP^M_{50}	AP^S_{50}
MaskRCNN [38]	**40.6**	**28.0**	**51.2**	**13.5**	**74.6**	**20.2**	**5.6**	-	-	-	-	-	-	-
MaskAmodal [30]	-	-	-	-	-	-	-	40.4	**26.6**	**51.2**	14.8	72.9	**20.6**	6.8
MaskJoint	38.8	26.0	49.5	11.9	70.4	17.4	6.4	**40.8**	26.4	**51.2**	**15.8**	**73.1**	19.6	**7.5**

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Quantitative Results

DAVIS fraction	0%	10%	20%	30%	50%	100%
VideoMatch-S	**0.74**	**0.77**	**0.78**	**0.78**	**0.78**	0.79
VideoMatch	0.55	0.66	0.73	0.74	**0.78**	**0.81**

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Qualitative Results – Amodal Segmentation

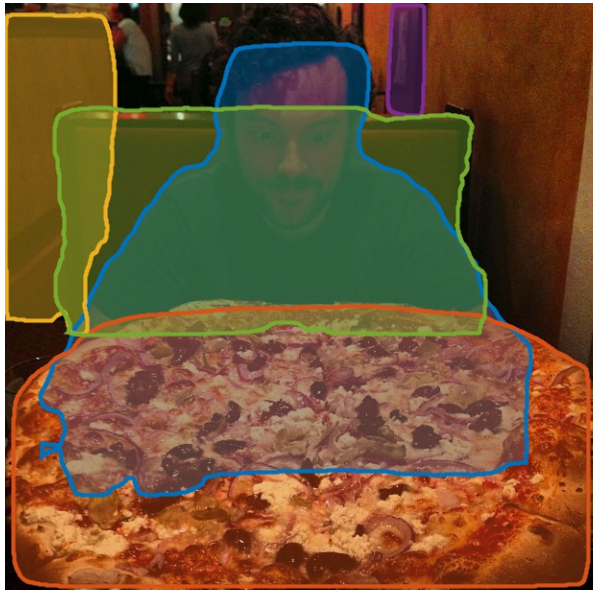
Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

Qualitative Results


Image


Groundtruth


Pretrained model on our dataset


Finetuned model on COCOA

Joint with
Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang
CVPR 2019

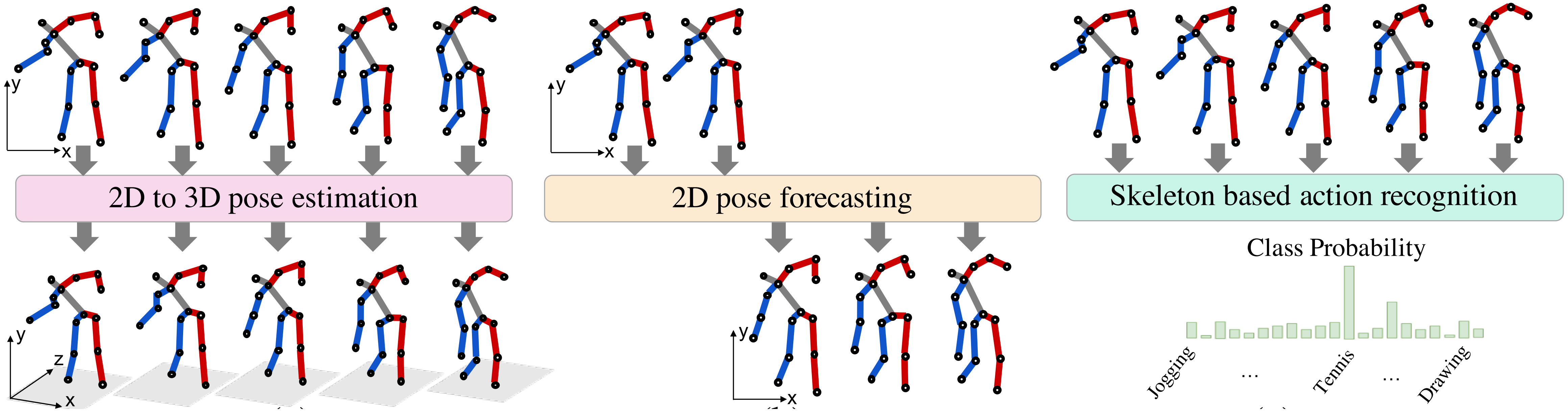
Video Results

Summary

- A dataset for semantic amodal instance-level video object segmentation
- Groundtruth annotations include modal segmentation, amodal segmentation, semantic class labels and human pose information
- Transfer to real world COCOA dataset training with the proposed dataset

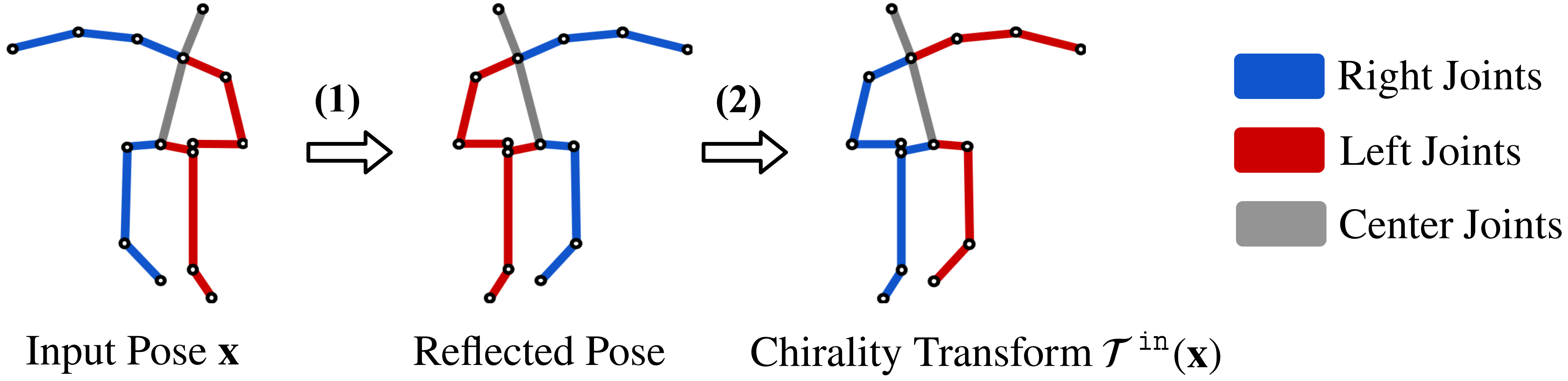
http://sailvos.web.illinois.edu

Pose regression tasks



Data augmentation for pose regression tasks?

Data augmentation for pose regression tasks



Disadvantages of Data Augmentation

- Deep nets need to learn equivariance from data
- Sample-inefficient
- Computationally more demanding

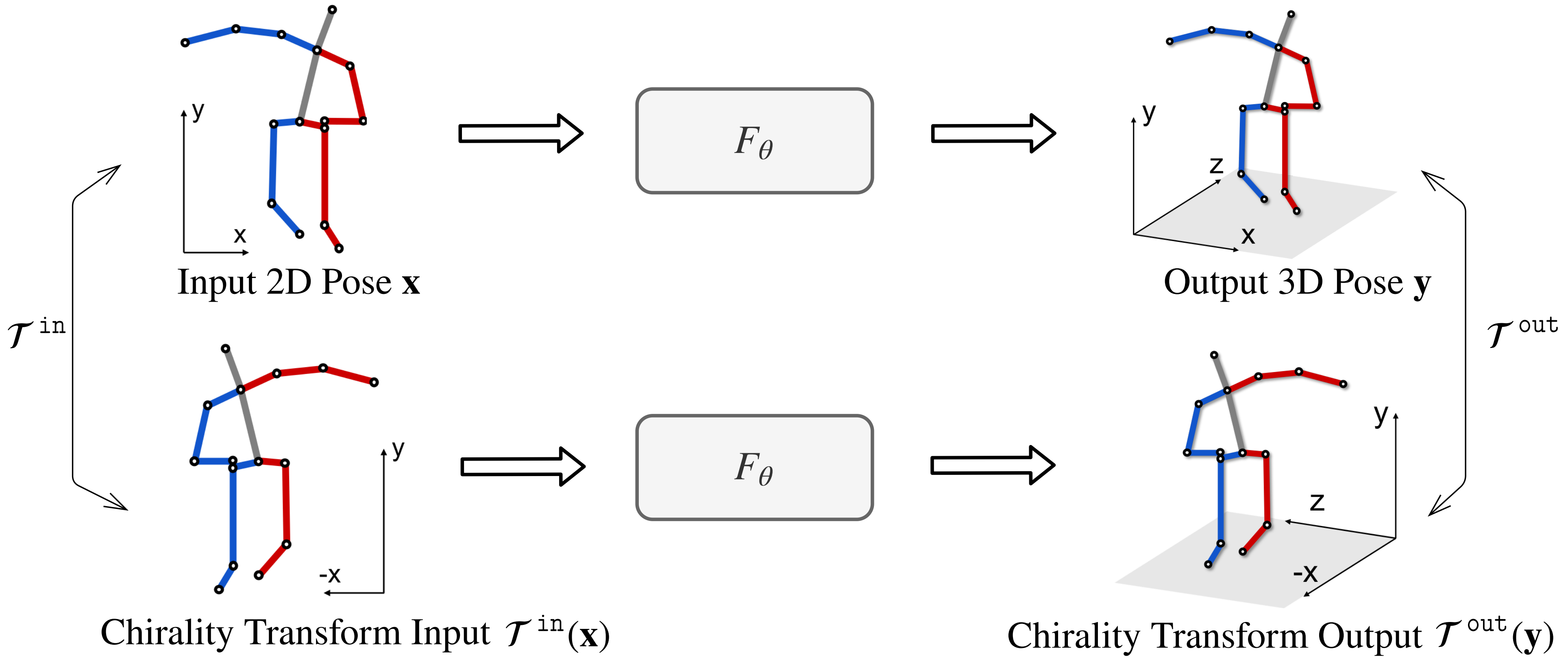
Question:

Can we develop deep nets that are equivariant w.r.t. pose transforms?

Joint with
Raymond Yeh, Yuan-Ting Hu
NeurIPS 2019

Chirality Nets

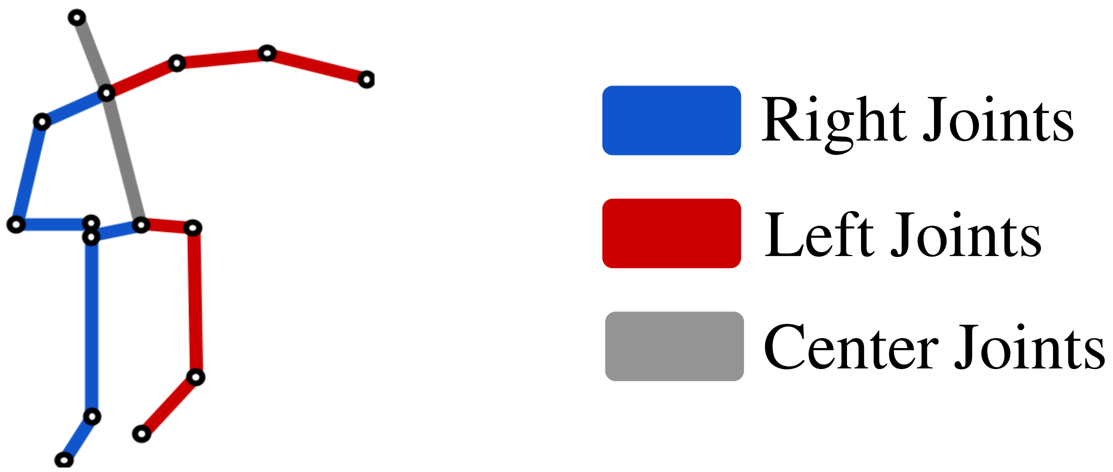
- Deep nets that guarantee the equivariant output



Joint with
Raymond Yeh, Yuan-Ting Hu
NeurIPS 2019

How to do it

- Define groups (right, left, center)
- Order sample data

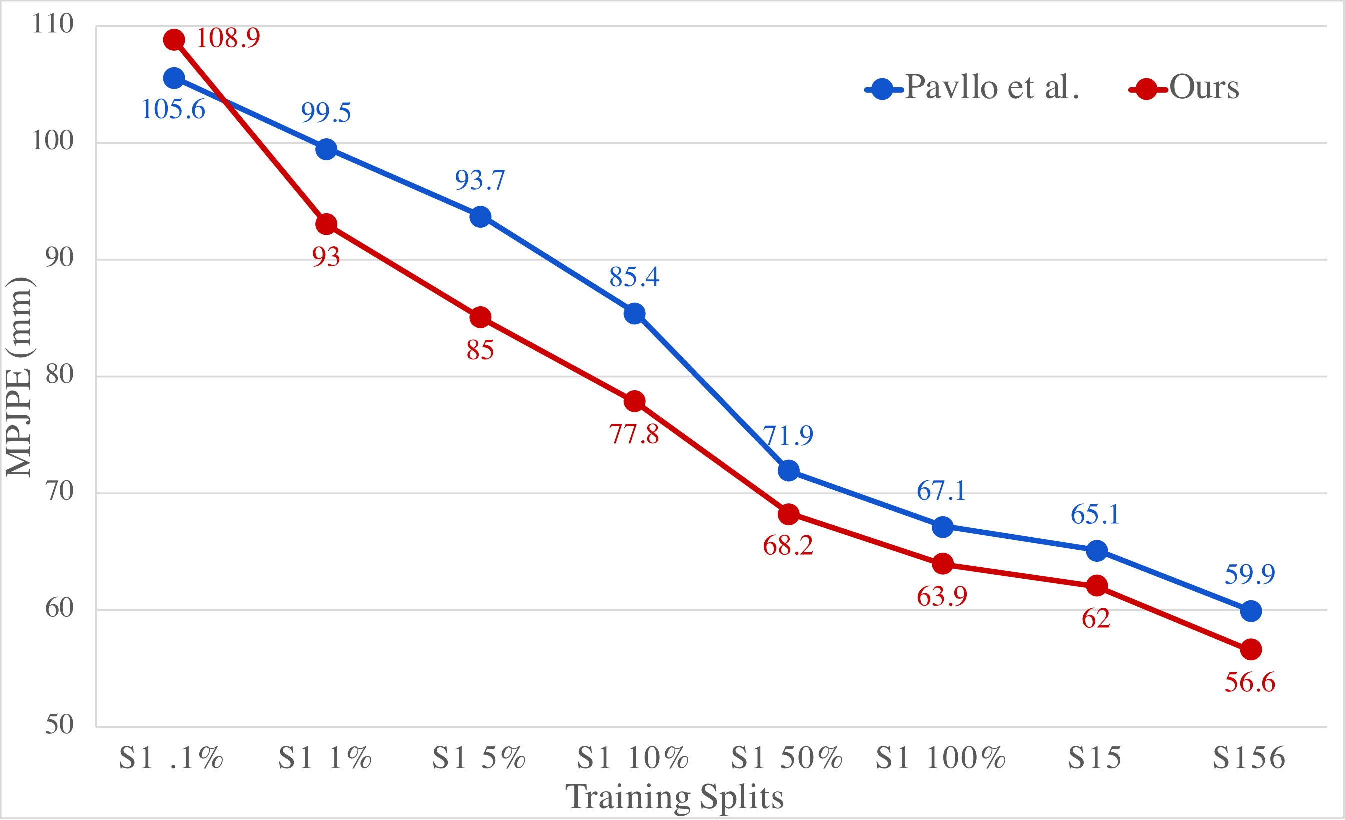


$$W = \begin{bmatrix} \begin{bmatrix} W_{\mathtt{ln},\mathtt{ln}} & W_{\mathtt{ln},\mathtt{lp}} \\ W_{\mathtt{lp},\mathtt{ln}} & W_{\mathtt{lp},\mathtt{lp}} \end{bmatrix} & \begin{bmatrix} W_{\mathtt{ln},\mathtt{rn}} & W_{\mathtt{ln},\mathtt{rp}} \\ W_{\mathtt{lp},\mathtt{rn}} & W_{\mathtt{lp},\mathtt{rp}} \end{bmatrix} & \begin{bmatrix} W_{\mathtt{ln},\mathtt{cn}} & W_{\mathtt{ln},\mathtt{cp}} \\ W_{\mathtt{lp},\mathtt{cn}} & W_{\mathtt{lp},\mathtt{cp}} \end{bmatrix} \\ \begin{bmatrix} W_{\mathtt{ln},\mathtt{rn}} & -W_{\mathtt{ln},\mathtt{rp}} \\ -W_{\mathtt{lp},\mathtt{rn}} & W_{\mathtt{lp},\mathtt{rp}} \end{bmatrix} & \begin{bmatrix} W_{\mathtt{ln},\mathtt{ln}} & -W_{\mathtt{ln},\mathtt{lp}} \\ -W_{\mathtt{lp},\mathtt{ln}} & W_{\mathtt{lp},\mathtt{lp}} \end{bmatrix} & \begin{bmatrix} W_{\mathtt{ln},\mathtt{cn}} & -W_{\mathtt{ln},\mathtt{cp}} \\ -W_{\mathtt{lp},\mathtt{cn}} & W_{\mathtt{lp},\mathtt{cp}} \end{bmatrix} \\ \begin{bmatrix} W_{\mathtt{cn},\mathtt{ln}} & W_{\mathtt{cn},\mathtt{lp}} \\ \mathbf{0} & W_{\mathtt{cp},\mathtt{lp}} \end{bmatrix} & \begin{bmatrix} W_{\mathtt{cn},\mathtt{ln}} & -W_{\mathtt{cn},\mathtt{lp}} \\ \mathbf{0} & W_{\mathtt{cp},\mathtt{lp}} \end{bmatrix} & \begin{bmatrix} W_{\mathtt{cn},\mathtt{cn}} & \mathbf{0} \\ \mathbf{0} & W_{\mathtt{cp},\mathtt{cp}} \end{bmatrix} \end{bmatrix}, \quad b = \begin{bmatrix} \begin{bmatrix} b_{\mathtt{ln}} \\ b_{\mathtt{lp}} \end{bmatrix} \\ \begin{bmatrix} -b_{\mathtt{ln}} \\ b_{\mathtt{lp}} \end{bmatrix} \\ \begin{bmatrix} \mathbf{0} \\ b_{\mathtt{cp}} \end{bmatrix} \end{bmatrix}$$

Joint with
Raymond Yeh, Yuan-Ting Hu
NeurIPS 2019

2D to 3D Benefits of Chirality Nets

- No data-augmentation necessary
- More sample efficient (Human3.6M)



Joint with
Raymond Yeh, Yuan-Ting Hu
NeurIPS 2019

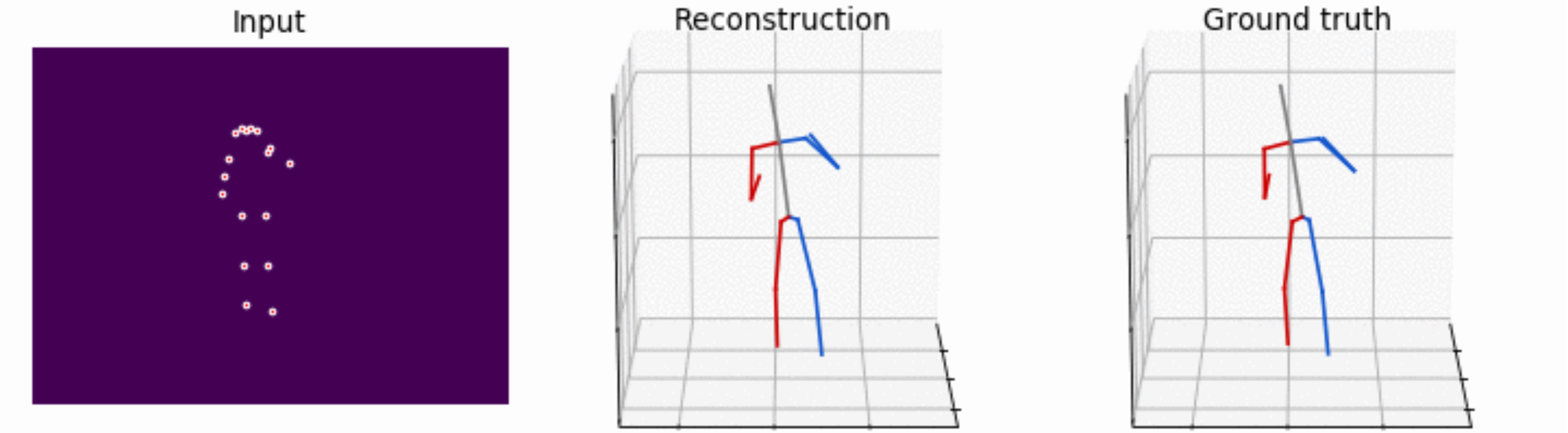
2D to 3D Benefits of Chirality Nets

- Human3.6M dataset

Approach	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Pavlakos [35] (CVPR'18)	48.5	54.4	54.4	52.0	59.4	65.3	49.9	52.9	65.8	71.1	56.6	52.9	60.9	44.7	47.8	56.2
Yang [52] (CVPR'18)	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	**43.6**	60.1	47.7	58.6
Luvizon [28] (CVPR'18) (◇)	49.2	51.6	47.6	50.5	51.8	60.3	48.5	51.7	61.5	70.9	53.7	48.9	57.9	44.4	48.9	53.2
Hossain [17] (ECCV'18)(†, ◇)	48.4	50.7	57.2	55.2	63.1	72.6	53.0	51.7	66.1	80.9	59.0	57.3	62.4	46.6	49.6	58.3
Lee [25] (ECCV'18)(†, ◇)	**40.2**	49.2	47.8	52.6	50.1	75.0	50.2	**43.0**	**55.8**	73.9	54.1	55.6	58.2	43.3	43.3	52.8
Pavllo [36] (CVPR'19)	47.1	50.6	49.0	51.8	53.6	61.4	49.4	47.4	59.3	67.4	52.4	49.5	55.3	39.5	42.7	51.8
Pavllo [36] (CVPR'19)(†)	45.9	47.5	44.3	46.4	50.0	56.9	45.6	44.6	58.8	66.8	47.9	44.7	49.7	33.1	34.0	47.7
Pavllo [36] (CVPR'19)(†, ‡)	45.2	46.7	**43.3**	**45.6**	**48.1**	**55.1**	**44.6**	44.3	57.3	65.8	**47.1**	44.0	49.0	32.8	33.9	46.8
Ours, single-frame	47.4	49.9	47.4	51.1	53.8	61.2	48.3	45.9	60.4	67.1	52.0	48.6	54.6	40.1	43.0	51.4
Ours (†)	44.8	**46.1**	**43.3**	46.4	49.0	55.2	**44.6**	44.0	58.3	**62.7**	**47.1**	**43.9**	48.6	**32.7**	**33.3**	**46.7**

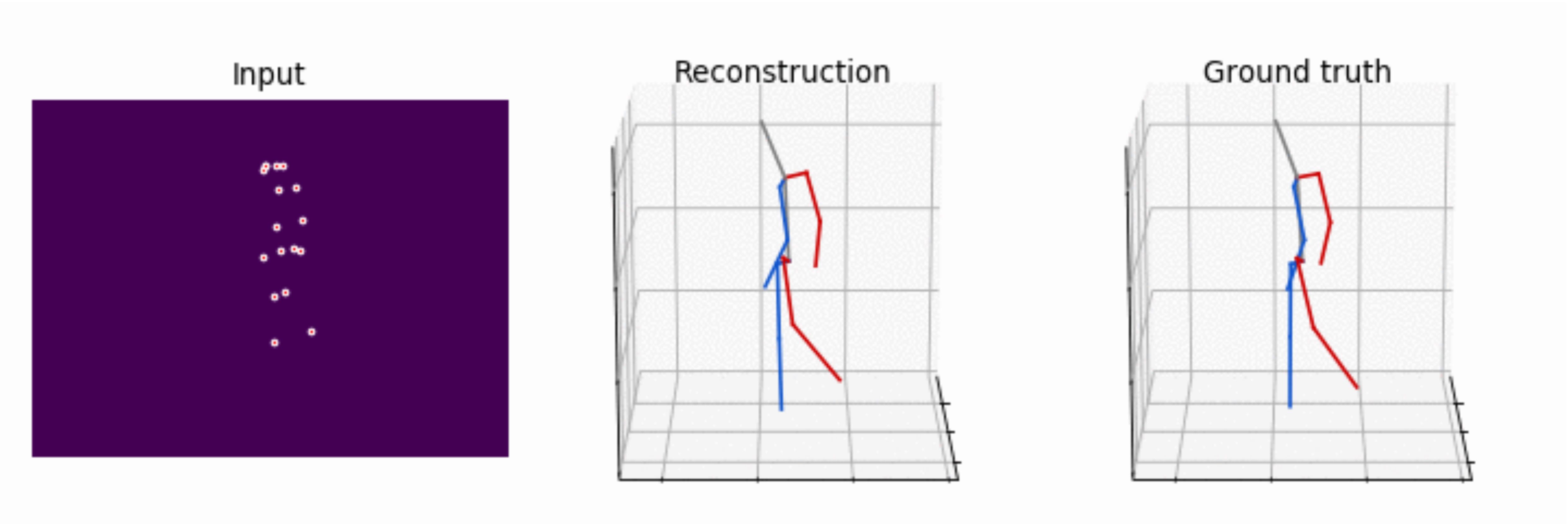
Joint with
Raymond Yeh, Yuan-Ting Hu
NeurIPS 2019

Qualitative Results



Joint with
Raymond Yeh, Yuan-Ting Hu
NeurIPS 2019

Qualitative Results



Joint with
Raymond Yeh, Yuan-Ting Hu
NeurIPS 2019

Forecasting Benefits of Chirality Nets

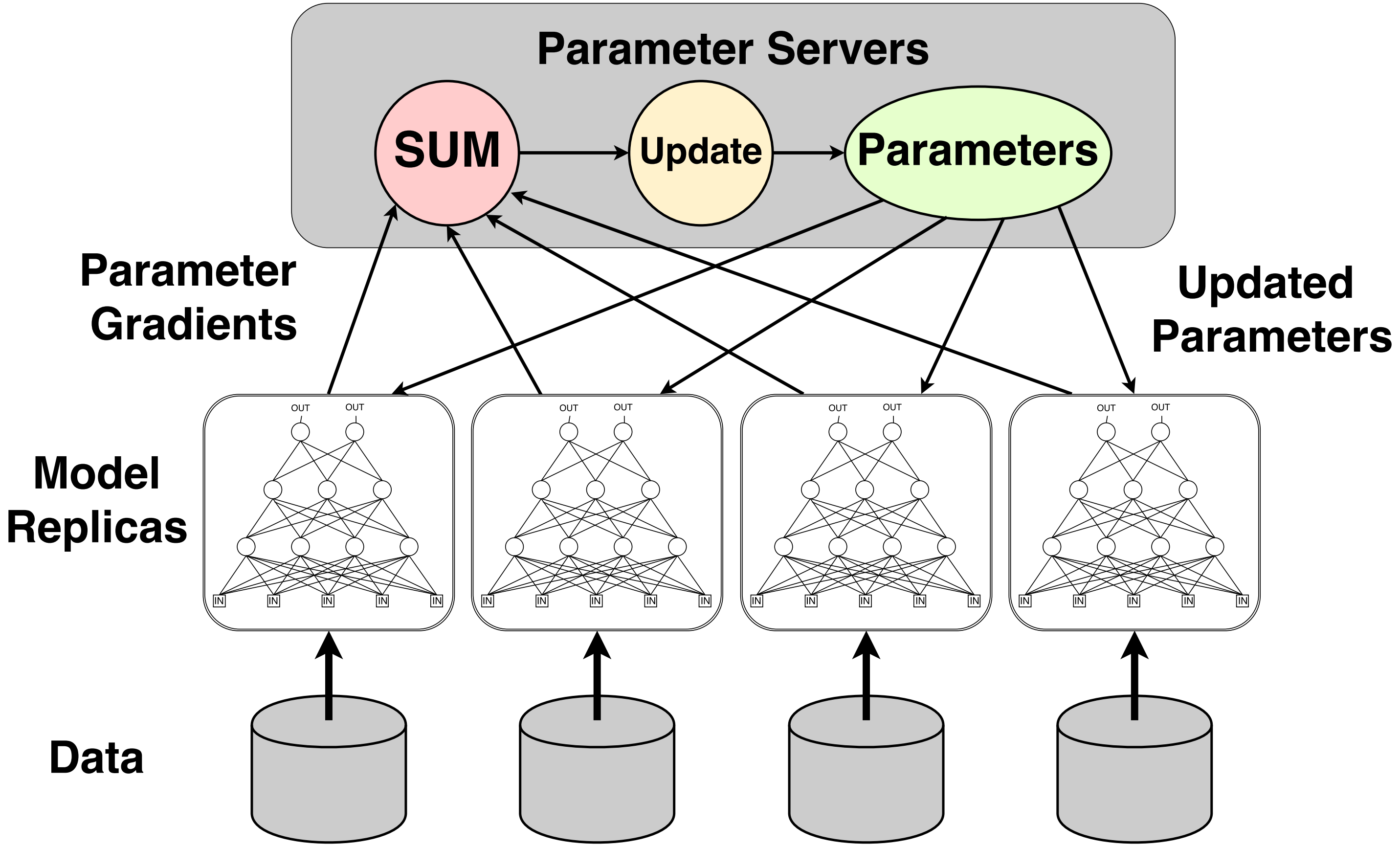
	Prediction Steps																Avg.
Approach	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	-
Residual [29] (CVPR‘17)	82.4	68.3	58.5	50.9	44.7	40.0	36.4	33.4	31.3	29.5	28.3	27.3	26.4	25.7	25.0	24.5	39.5
3D-PFNet [3](CVPR‘17)	79.2	60.0	49.0	43.9	41.5	40.3	39.8	39.7	40.1	40.5	41.1	41.6	42.3	42.9	43.2	43.3	45.5
TP-RNN [5] (WACV‘19)	84.5	72.0	64.8	60.3	57.2	55.0	53.4	52.1	50.9	50.0	49.3	48.7	48.3	47.9	47.6	47.3	55.6
Baseline w/o aug.	87.3	75.7	68.5	64.0	61.0	59.1	**57.6**	56.3	55.4	54.9	54.5	54.5	54.4	54.5	54.6	**54.7**	60.4
Baseline w/ aug.	86.9	75.2	67.9	63.5	60.4	58.4	57.0	55.8	55.1	54.5	54.1	54.0	53.9	53.9	54.0	54.0	59.9
Baseline w/ aug.(‡)	87.0	75.5	68.4	64.1	61.0	59.1	57.5	56.3	55.5	55.0	**54.7**	**54.7**	**54.6**	**54.7**	**54.7**	**54.7**	60.5
Ours	**87.5**	**77.0**	**68.7**	**64.2**	**61.2**	**59.2**	**57.6**	**56.5**	**55.7**	**55.1**	**54.7**	54.6	54.4	54.5	54.5	54.5	**60.6**

Deep Net Training

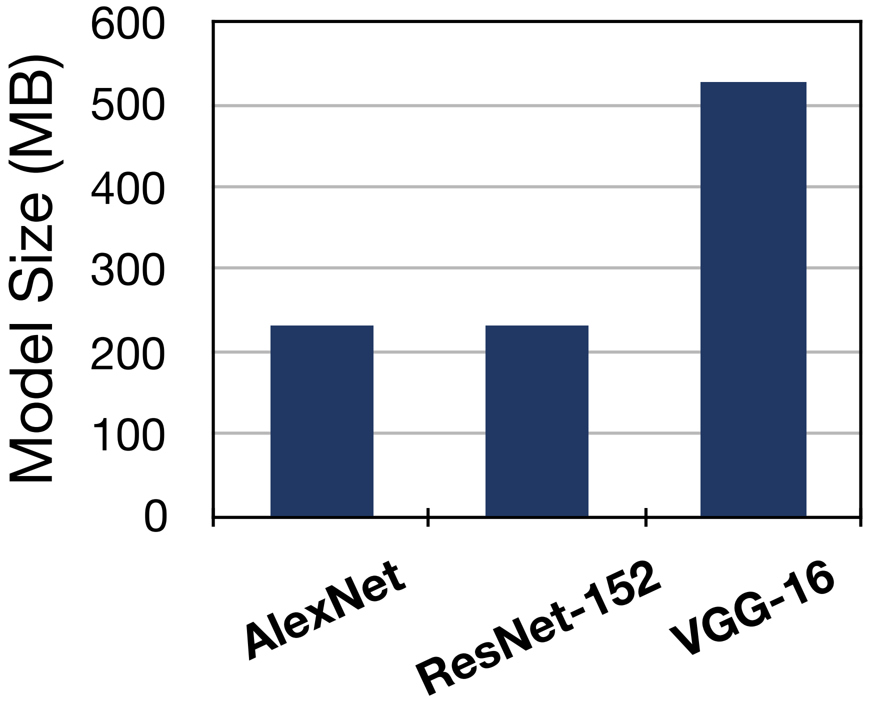
Algorithm:

- Load a batch of samples
- Compute predictions for every sample
- Compare predictions to groundtruth
- Backpropagate error
- Update parameters

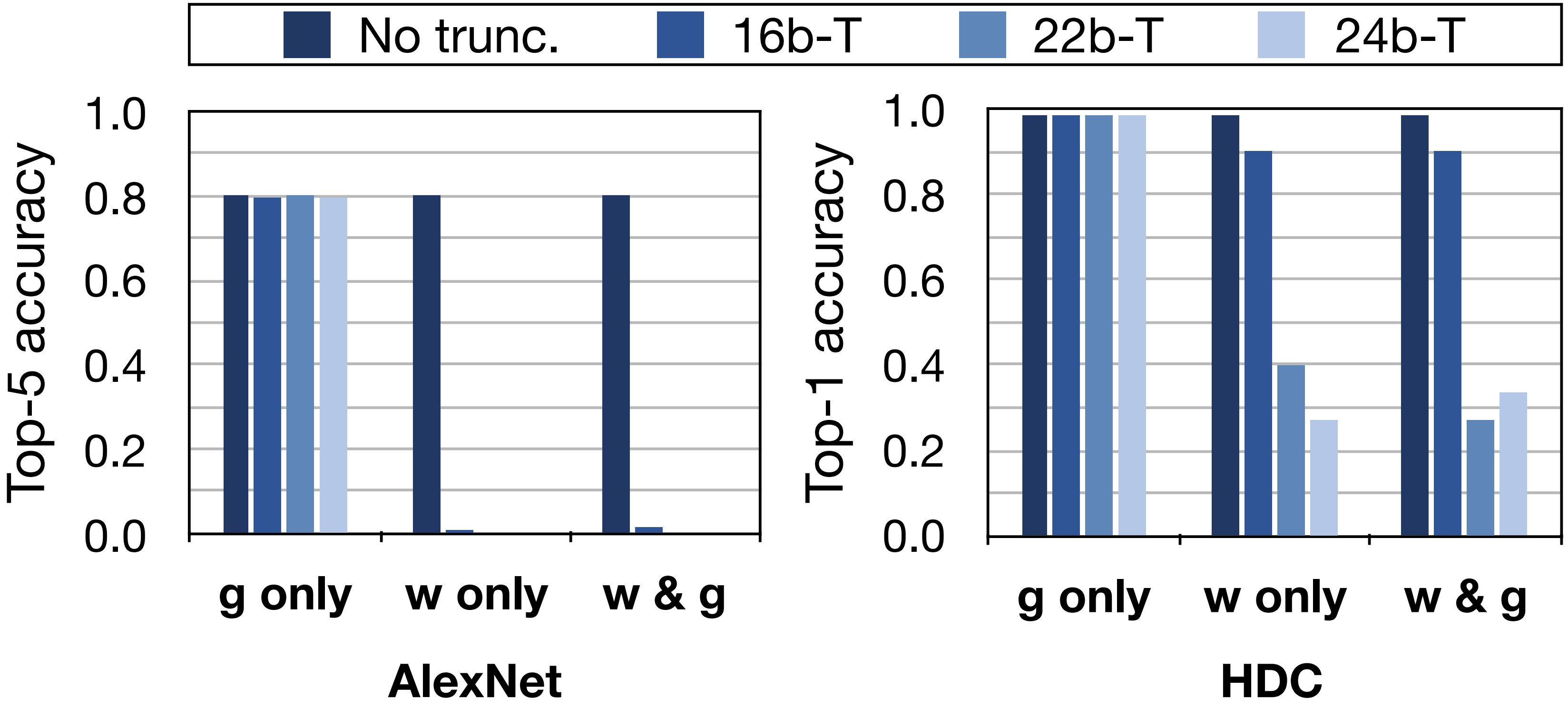
Distributed Deep Net Training



Communication is expensive

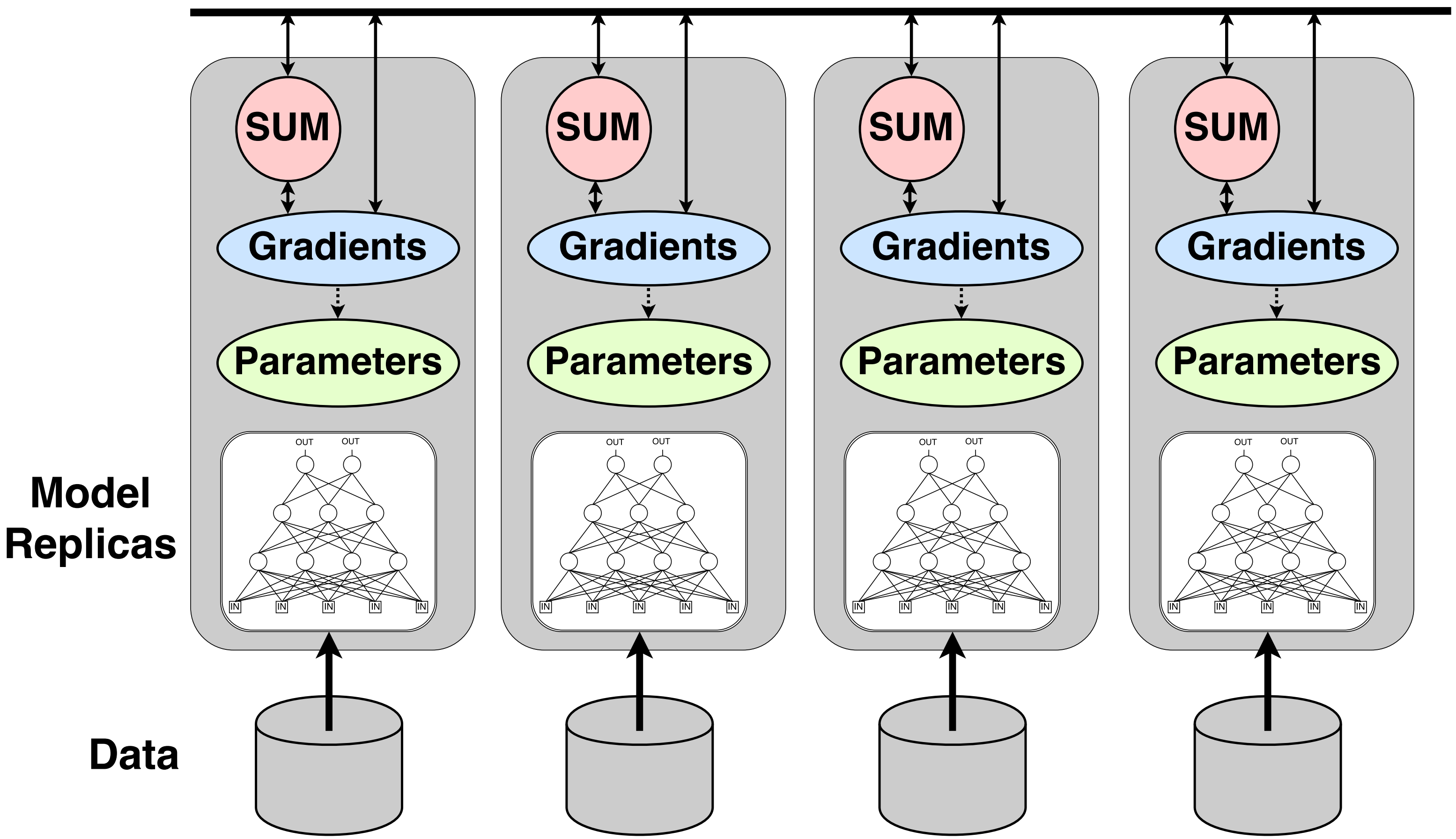


Parameters are not suitable for lossy communication

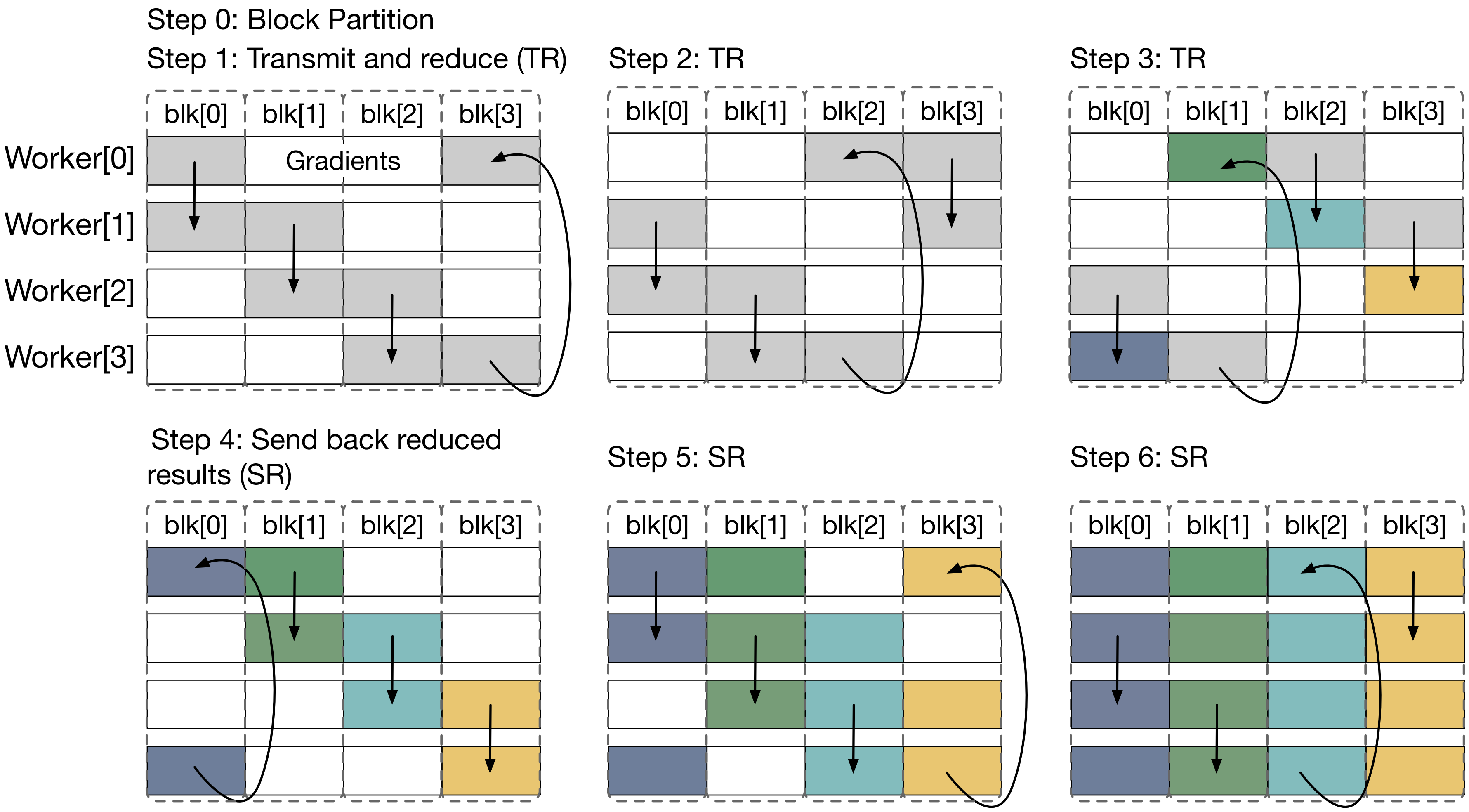


Joint with
Youjie Li, Mingchao Yu, Songze Li,
Salman Avestimehr, Nam Sung Kim
NeurIPS 2018

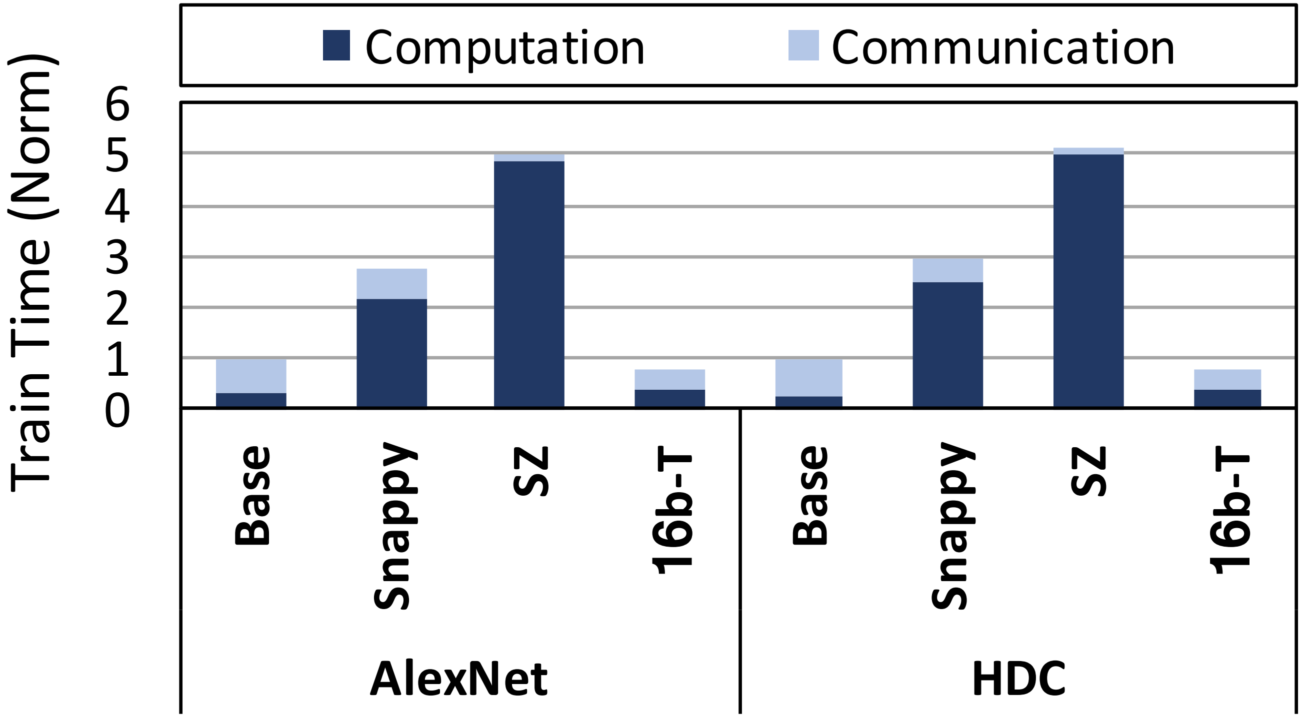
Idea



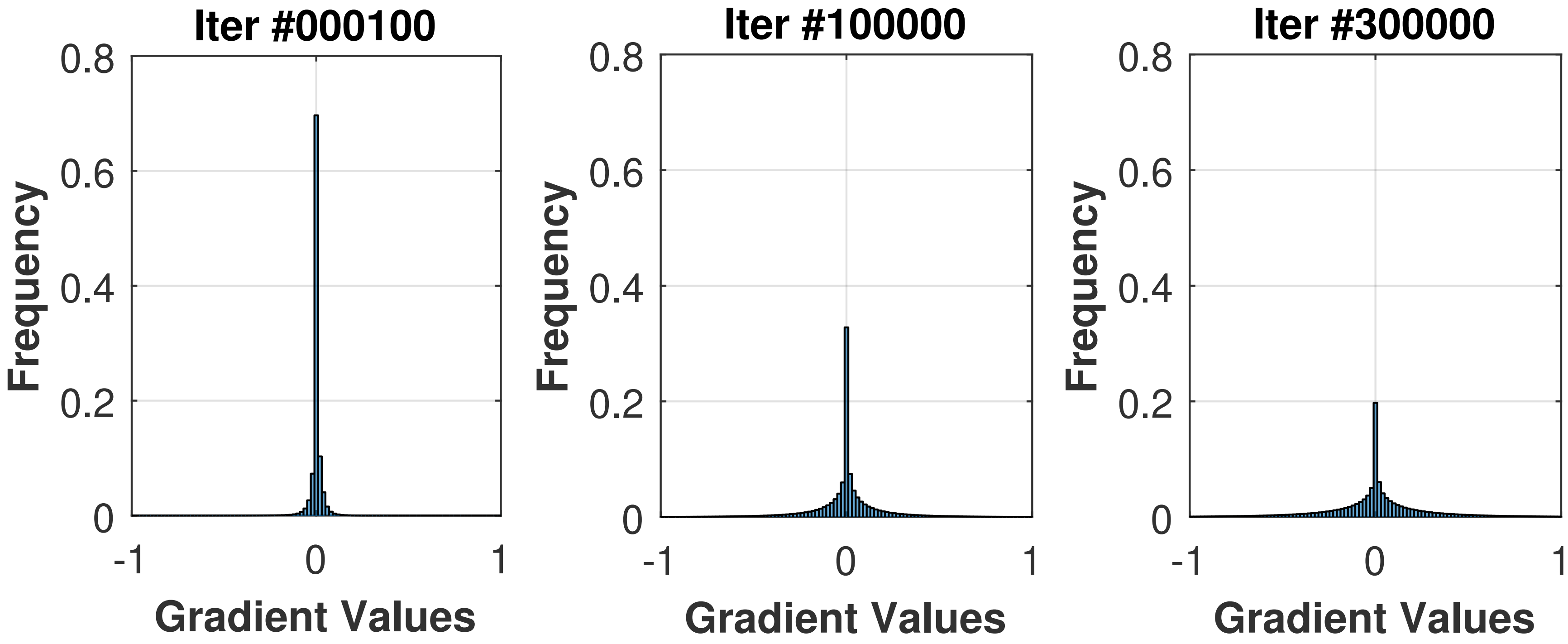
Efficient communication



Standard Compression

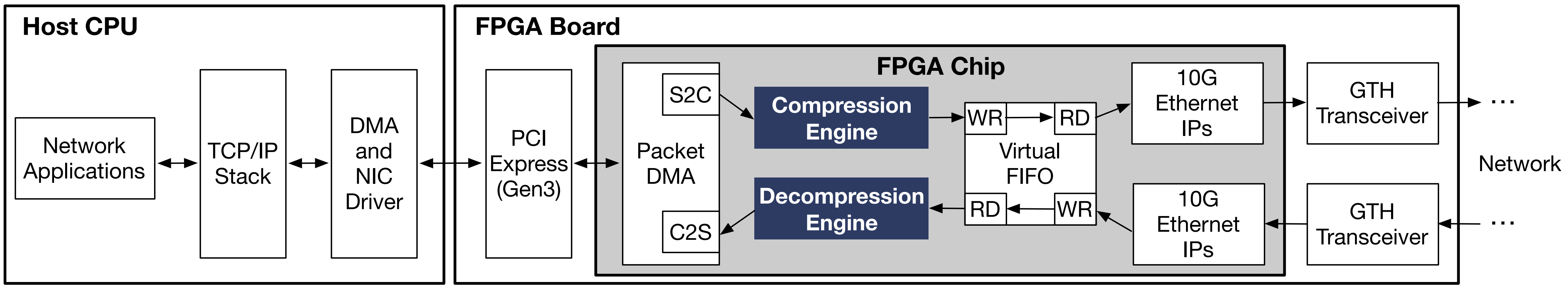


Dedicated Compression



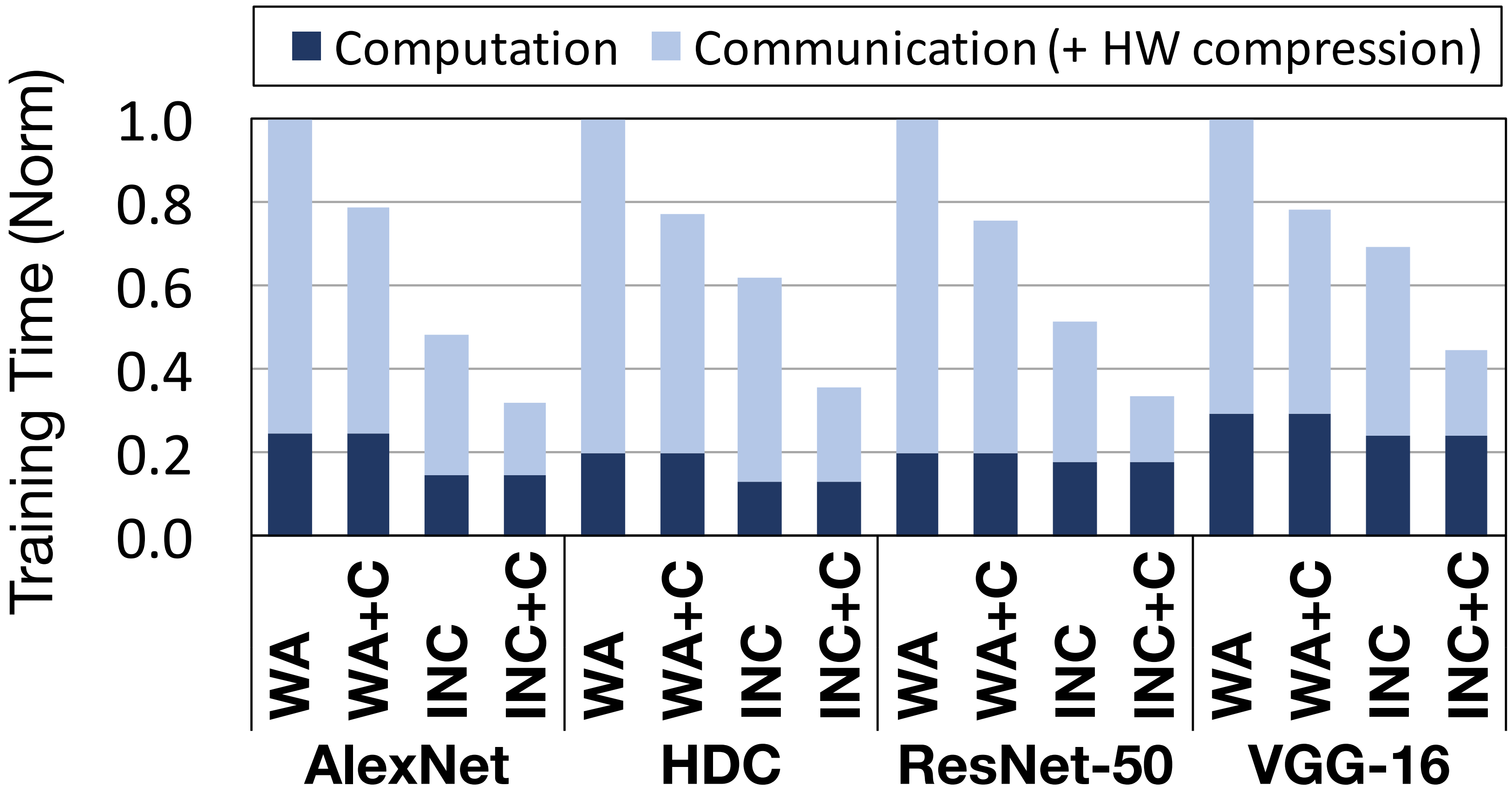
Compression in NIC

Joint with
Youjie Li, Mingchao Yu, Songze Li,
Salman Avestimehr, Nam Sung Kim
NeurIPS 2018



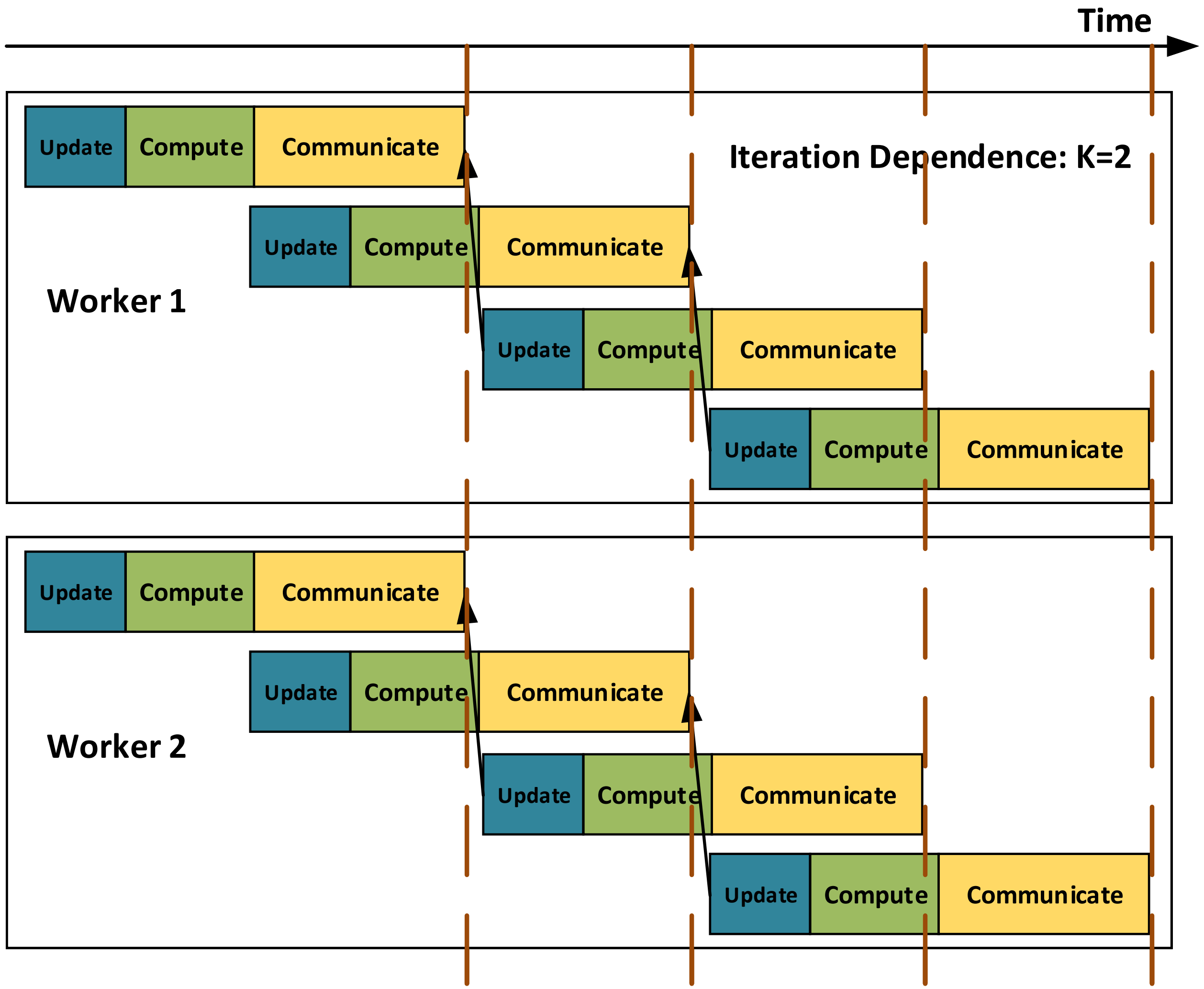
Joint with
Youjie Li, Mingchao Yu, Songze Li,
Salman Avestimehr, Nam Sung Kim
NeurIPS 2018

Results



Joint with
Youjie Li, Mingchao Yu, Songze Li,
Salman Avestimehr, Nam Sung Kim
NeurIPS 2018

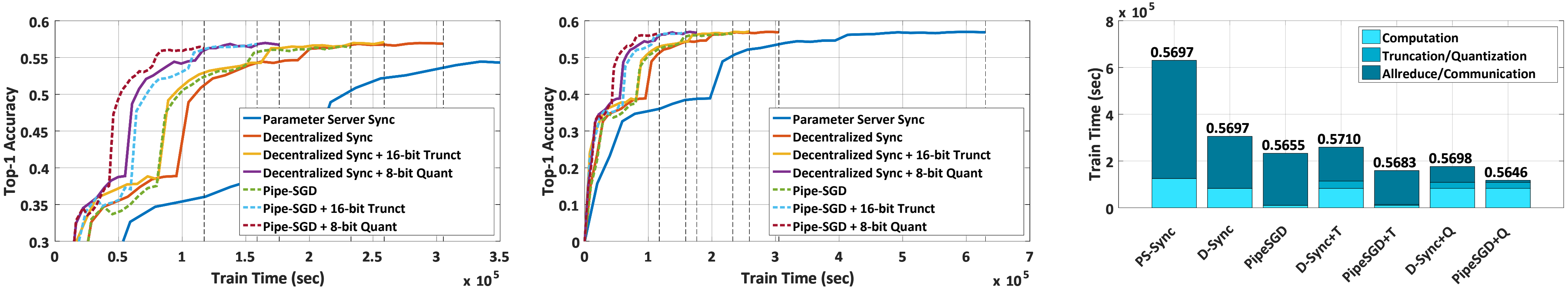
Pipelining is feasible



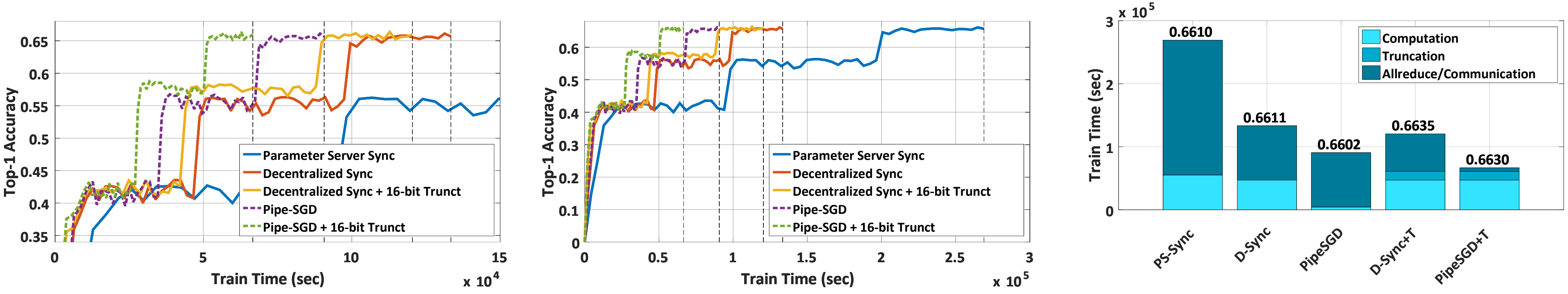
Joint with
Youjie Li, Mingchao Yu, Songze Li,
Salman Avestimehr, Nam Sung Kim
NeurIPS 2018

Results

AlexNet



ResNet18



Learning to Anticipate

Thanks

NSF SAMSUNG 3M CISCO Adobe® IBM

Contact:

- http://aschwing.ece.Illinois.edu
- http://alexander-schwing.de
- aschwing@Illinois.edu