ContextVP: Fully Context-Aware Video Prediction

Wonmin Byeon

NVIDIA Research

September 13, 2018

Collaborators

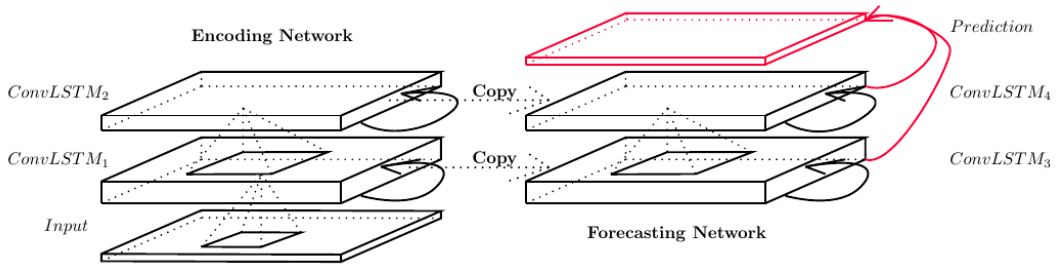
Qin Wang (ETH Zurich)

Rupesh Kumar Srivastava (NNAISENSE)

Petros Koumoutsakos (ETH Zurich)

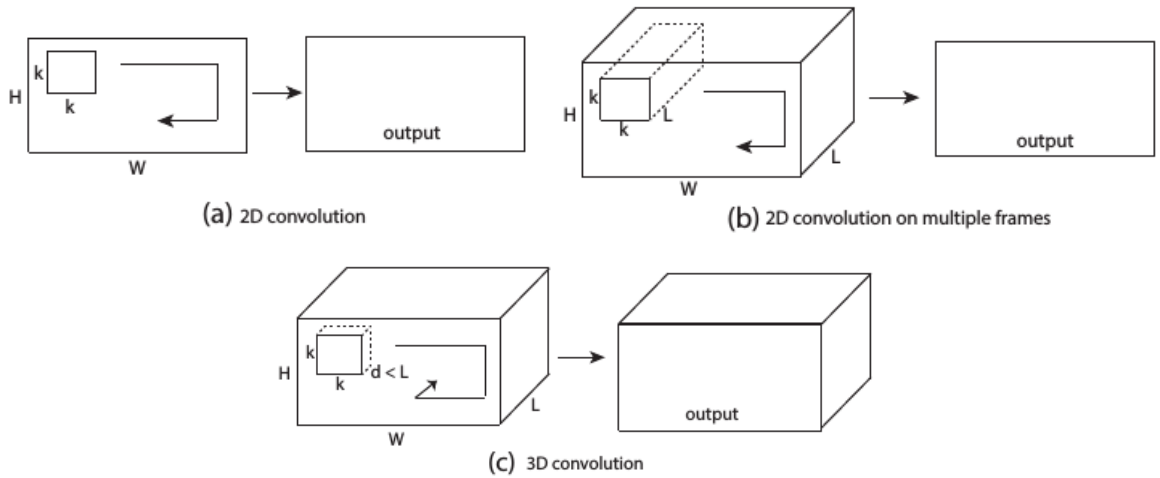
Video Prediction

t: time frame



Convolutional LSTM (ConvLSTM)

[Shi15, Finn16, Lotter16, Villegas17]



3D CNN

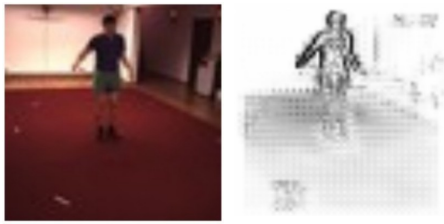
[Tran15, Mathieu15]

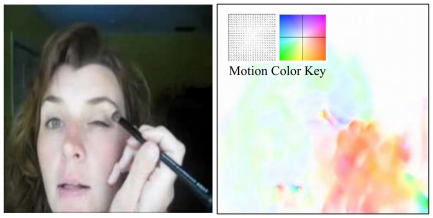
Blurry Predictions

[Ranzato14]

Solutions for Blurry Predictions?

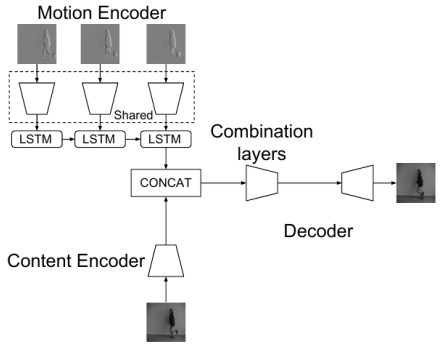
- Predicting pixel motions [Finn16]
- Estimating optical flow [Patraucean15, Liu17]





Solutions for Blurry Predictions?

- Predicting pixel motions [Finn16]
- Estimating optical flow [Patraucean15, Liu17]
- Explicitly modeling moving foreground objects separately from the background [Finn16, Vondrick16, Simonyan14]
- Decomposing the scene content and motion [Villegas17]



Solutions for Blurry Predictions?

- Predicting pixel motions [Finn16]
- Estimating optical flow [Patraucean15, Liu17]
- Explicitly modeling moving foreground objects separately from the background [Finn16, Vondrick16, Simonyan14]
- Decomposing the scene content and motion [Villegas17]
- Changing loss functions, Adversarial training [Mathieu15, Vondrick16]

What causes blurry predictions?

What causes blurry predictions?

- The future may be ambiguous given the past

What causes blurry predictions?

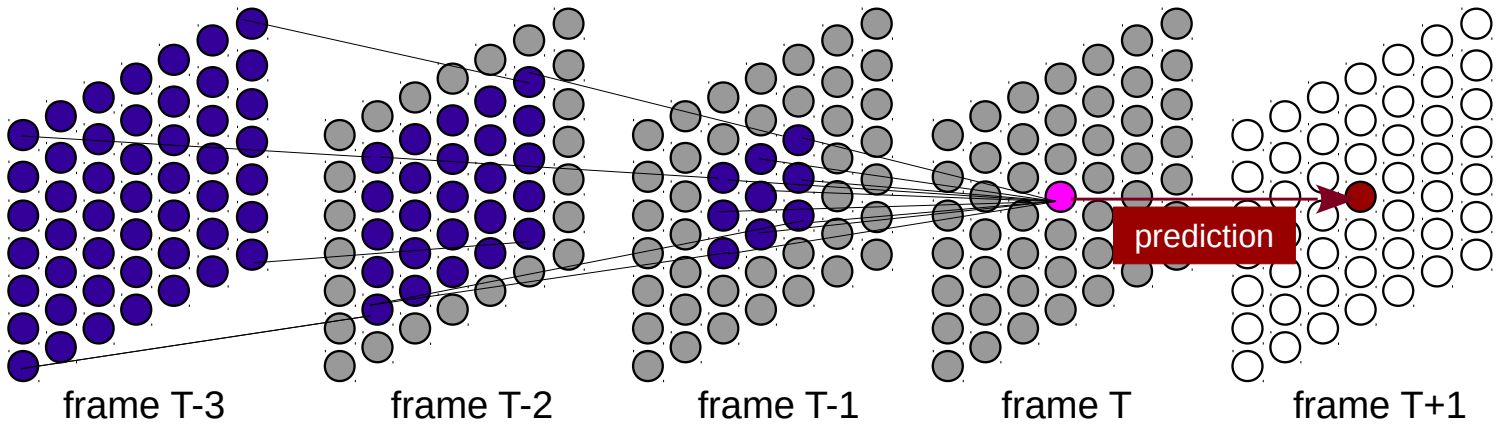
- The future may be ambiguous given the past
- The model does not utilize the all available information from the past

What causes blurry predictions?

- The future may be ambiguous given the past
- **The model does not utilize the all available information from the past**

Convolutional LSTM for Video Prediction

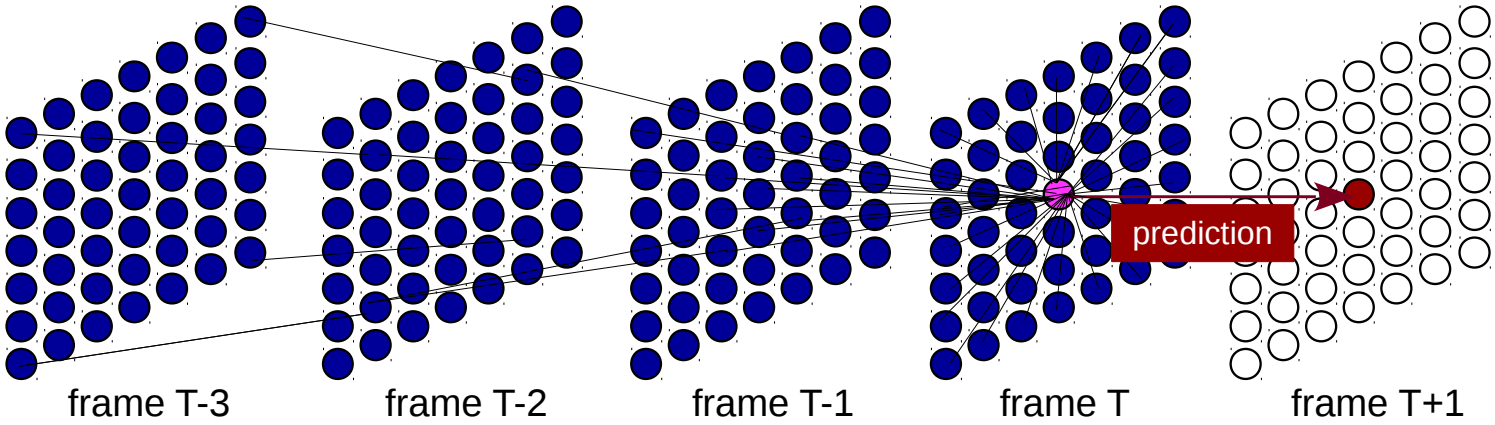
Blind Spot Problem



red: current pixel blue: context covered gray: blind spots

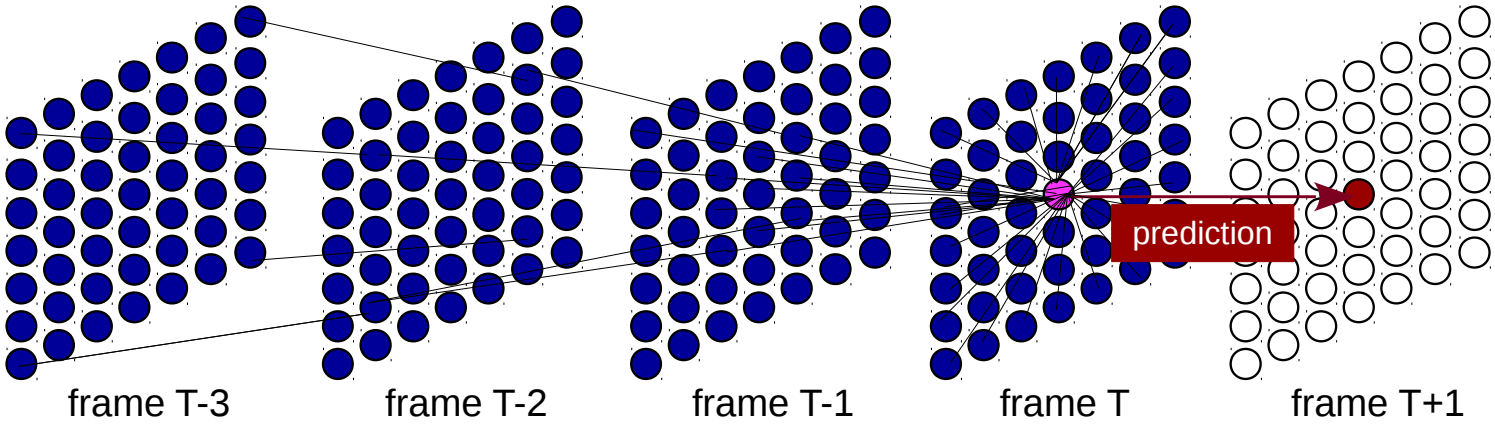
Convolutional LSTM for Video Prediction

Solution?



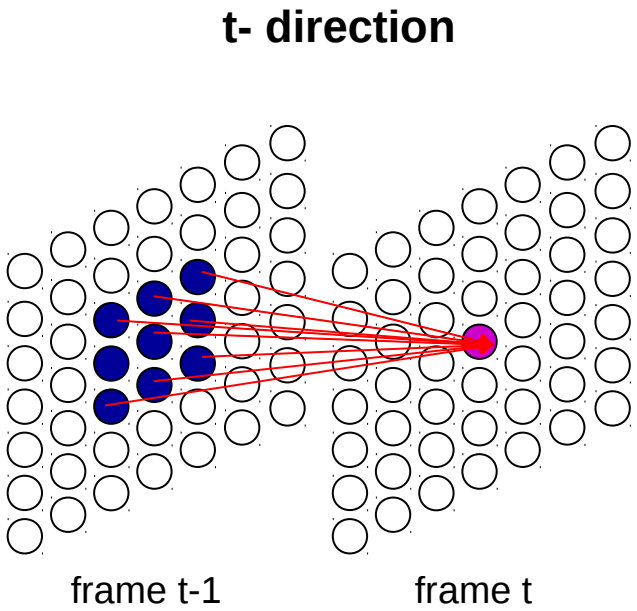
red: current pixel blue: context covered gray: blind spots

Convolutional LSTM for Video Prediction



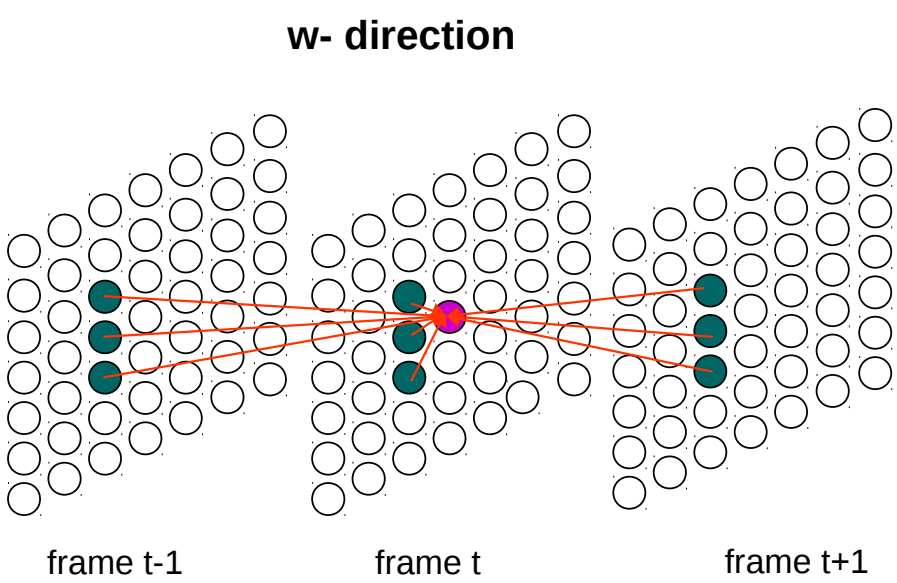
Fully Context-Aware Video Prediction!

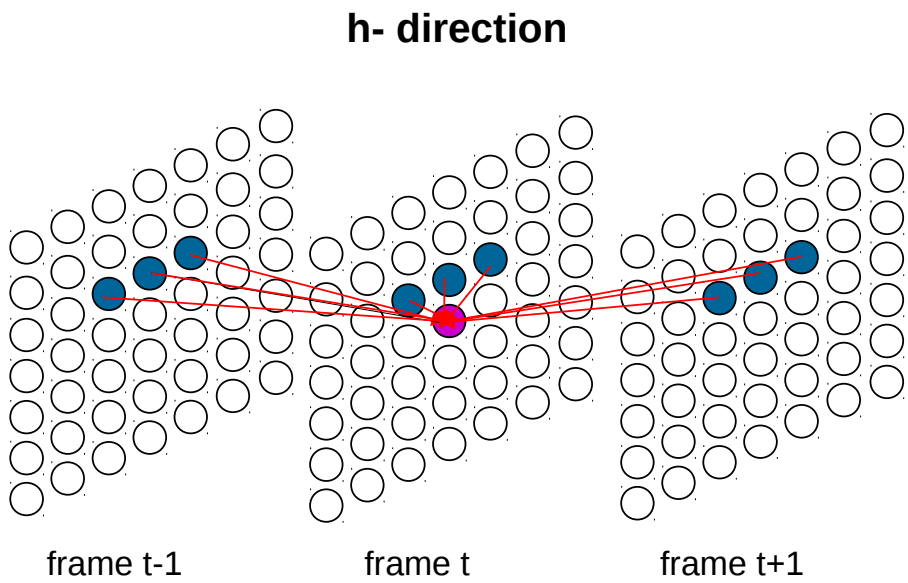
Fully Context-Aware Video Prediction: ContextVP



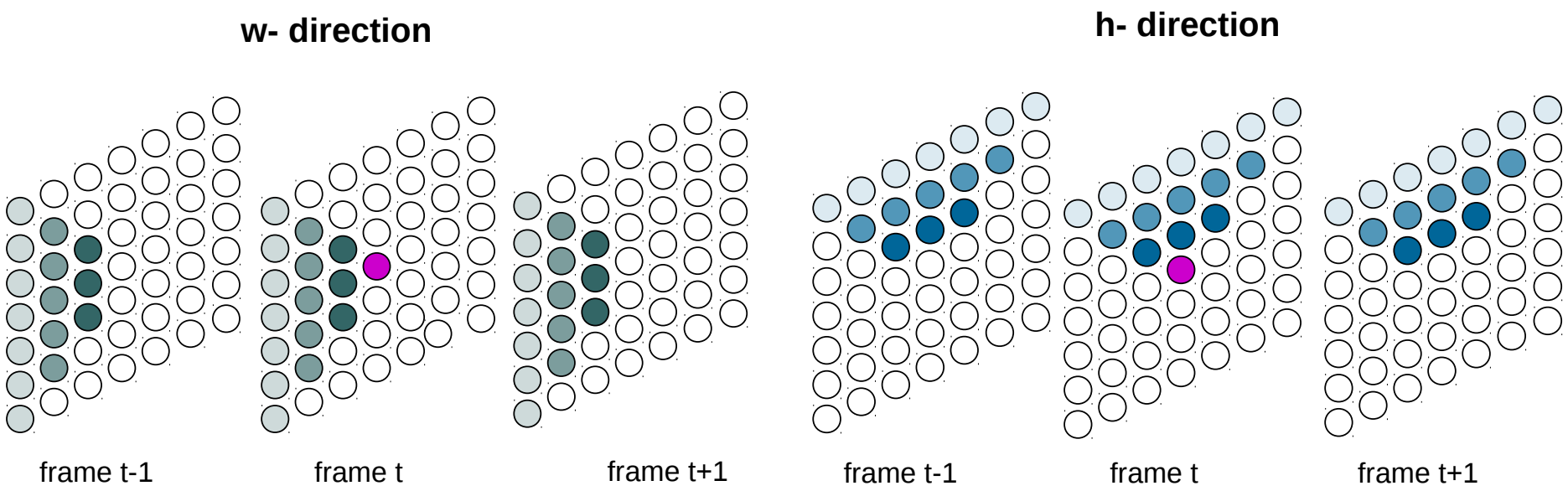
- Convolutional LSTM [Shi15]
- Parallel Multi-Dimensional LSTM (PMD-LSTM) along the time dimension (t-) [Stollenga&Byeon15]

Fully Context-Aware Video Prediction: ContextVP

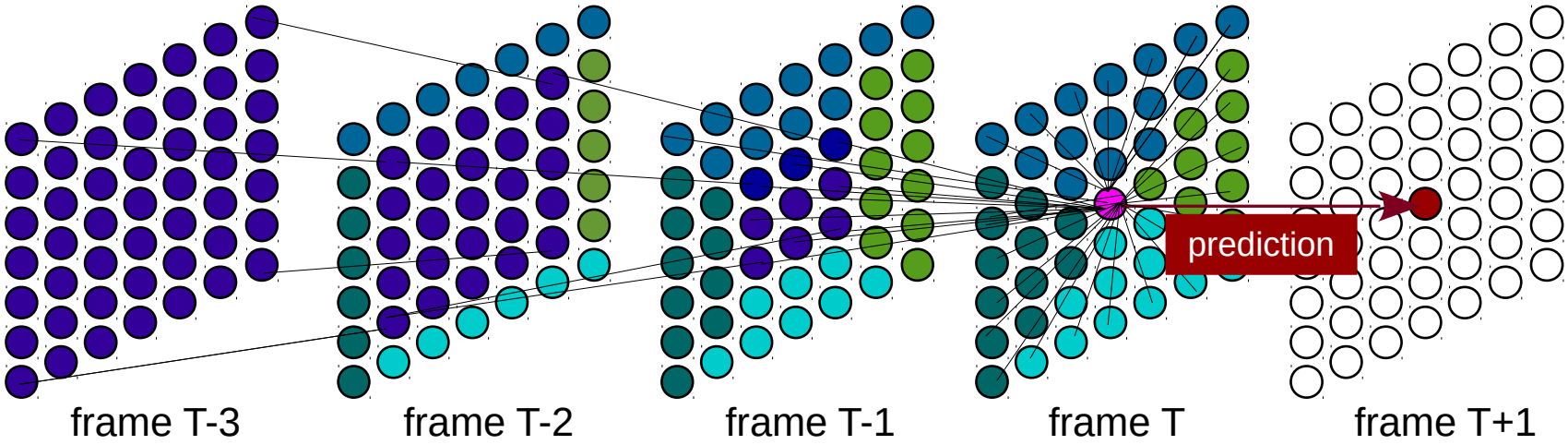




Fully Context-Aware Video Prediction: ContextVP



Fully Context-Aware Video Prediction: ContextVP

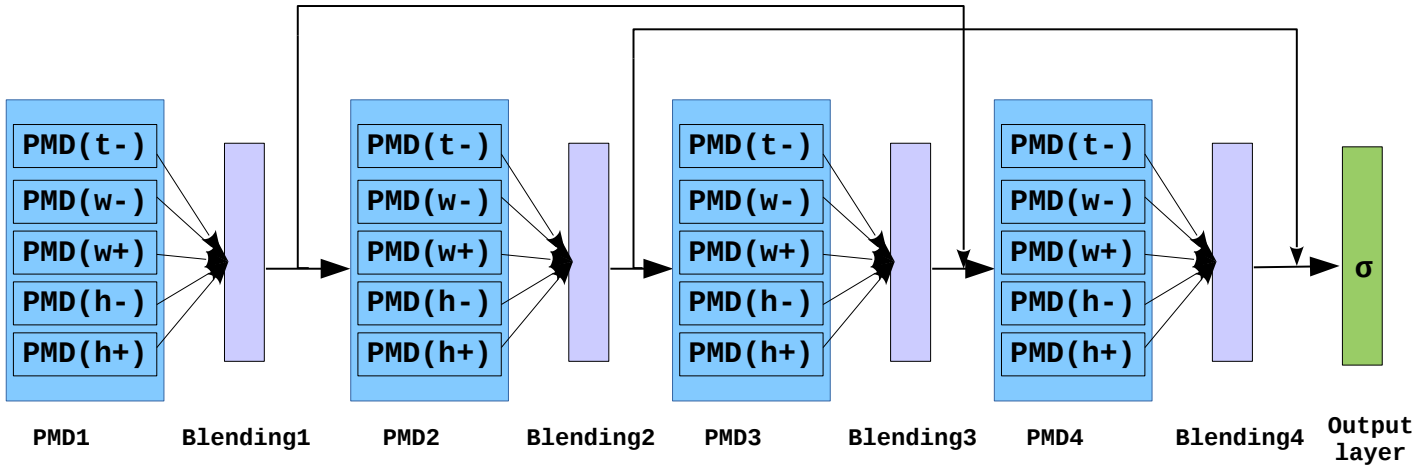


PMD-LSTM for all possible directions

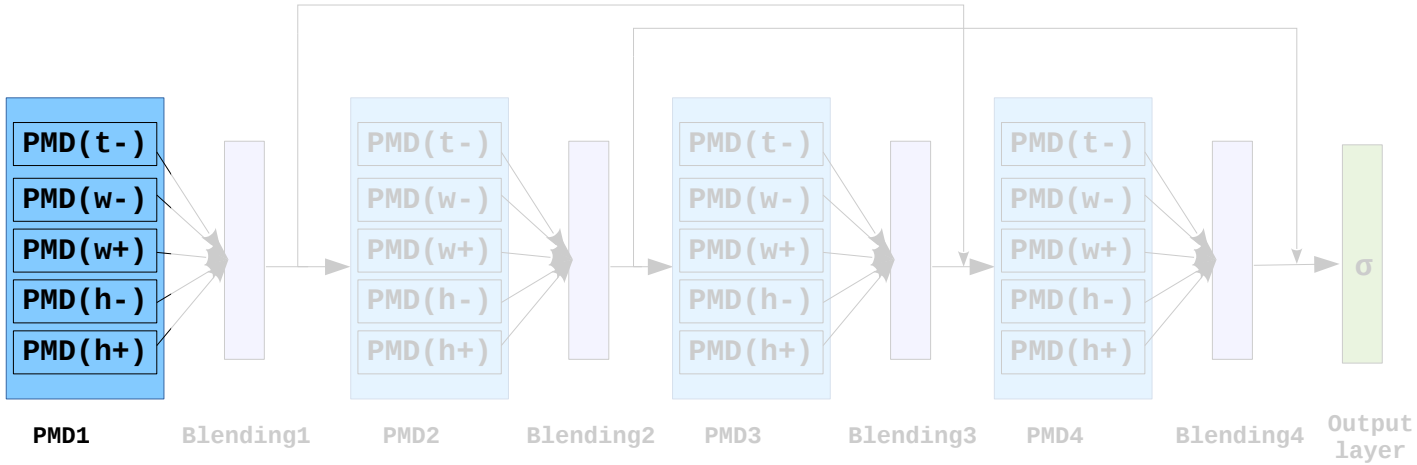
$(t\text{-}, w\text{-}, w+, h\text{-}, h+)$

Covering full available context without having a deep model

Fully Context-Aware Video Prediction: ContextVP

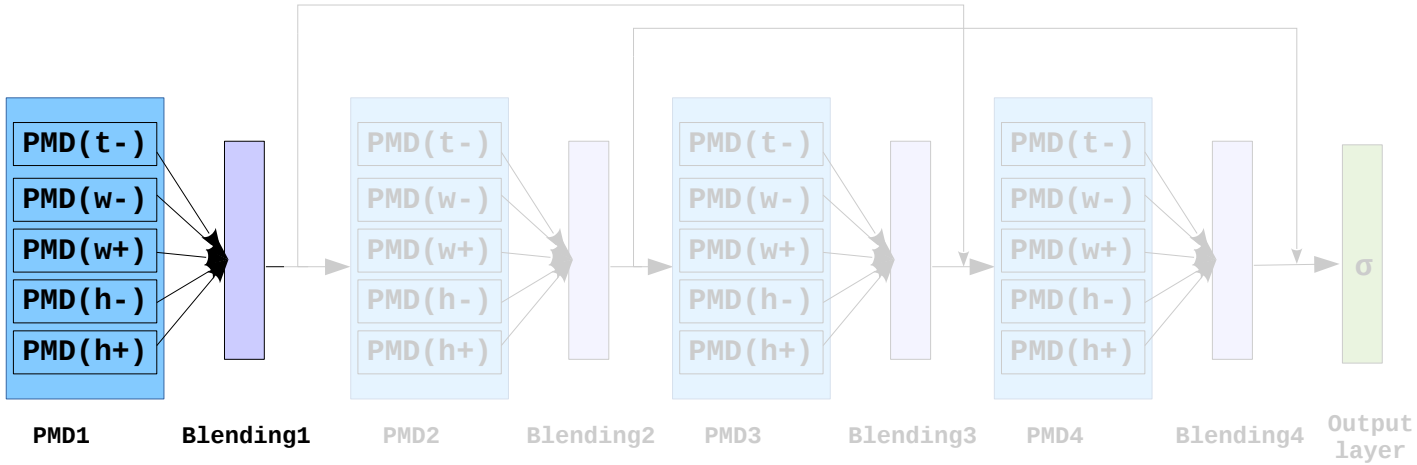


ContextVP: Architecture



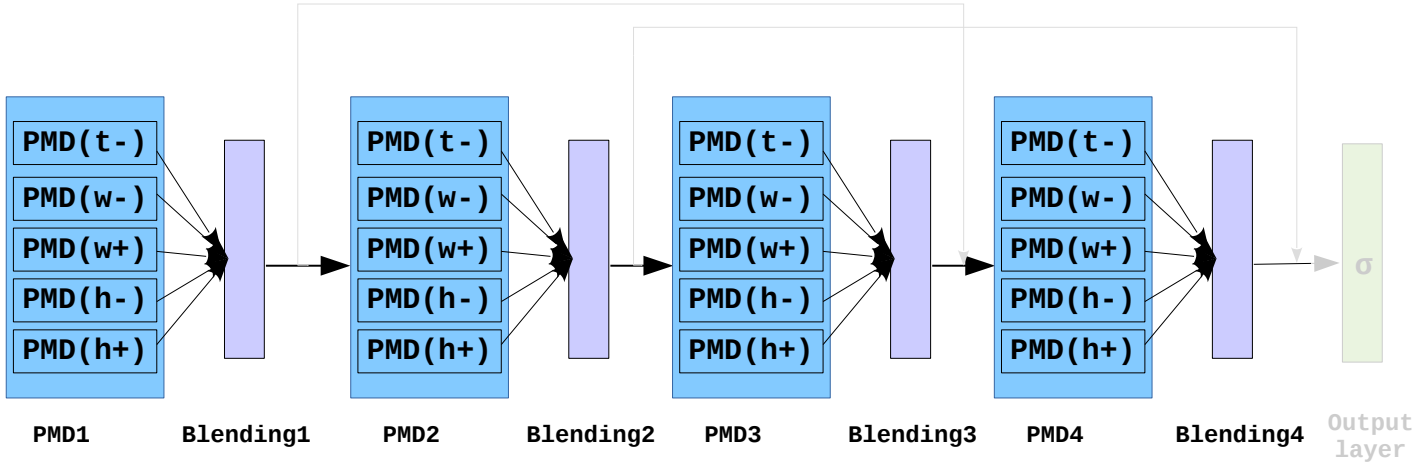
- **Parallel MD-LSTM (PMD) Block**
  - outputs with five recurrence directions: s^d, $d \in D = \{h\text{-}, h\text{+}, w\text{-}, w\text{+}, t\text{-}\}$

Fully Context-Aware Video Prediction: ContextVP



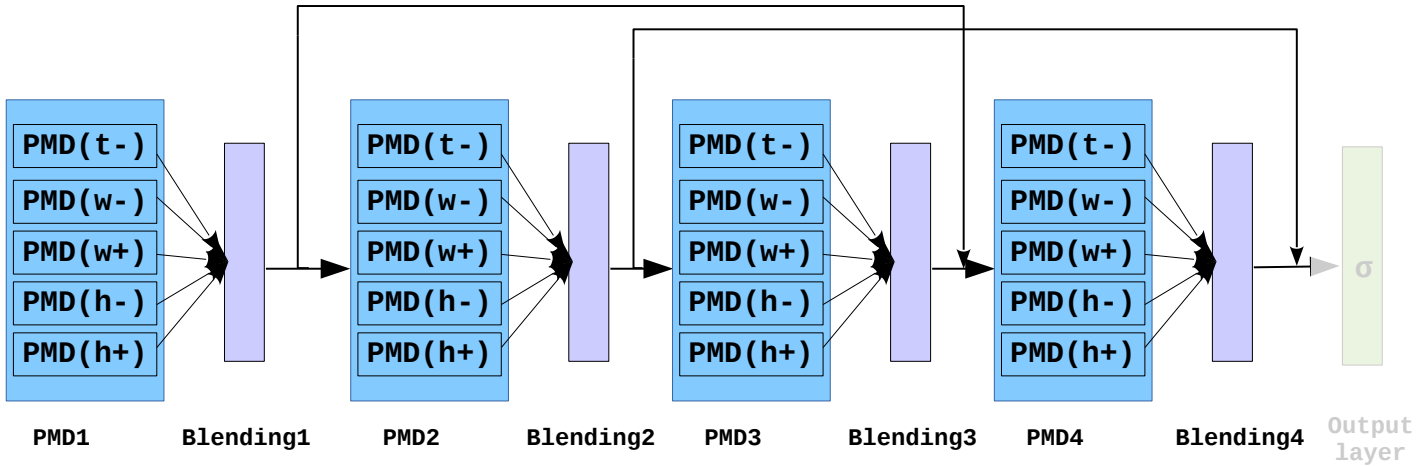
- Parallel MD-LSTM (PMD) Block: $s^d, \quad d \in D = \{h\text{-}, h\text{+}, w\text{-}, w\text{+}, t\text{-}\}$
- **Context Blending Block**
  - **Uniform Blending**: $m = f((\sum_{d \in D} s^d) \cdot W + b)$
  - **Weighted Blending**:
    $m = f(S \cdot W + b), \quad S = \begin{bmatrix} s^{t-} & s^{h-} & s^{h+} & s^{w-} & s^{w+} \end{bmatrix}^T$

Fully Context-Aware Video Prediction: ContextVP



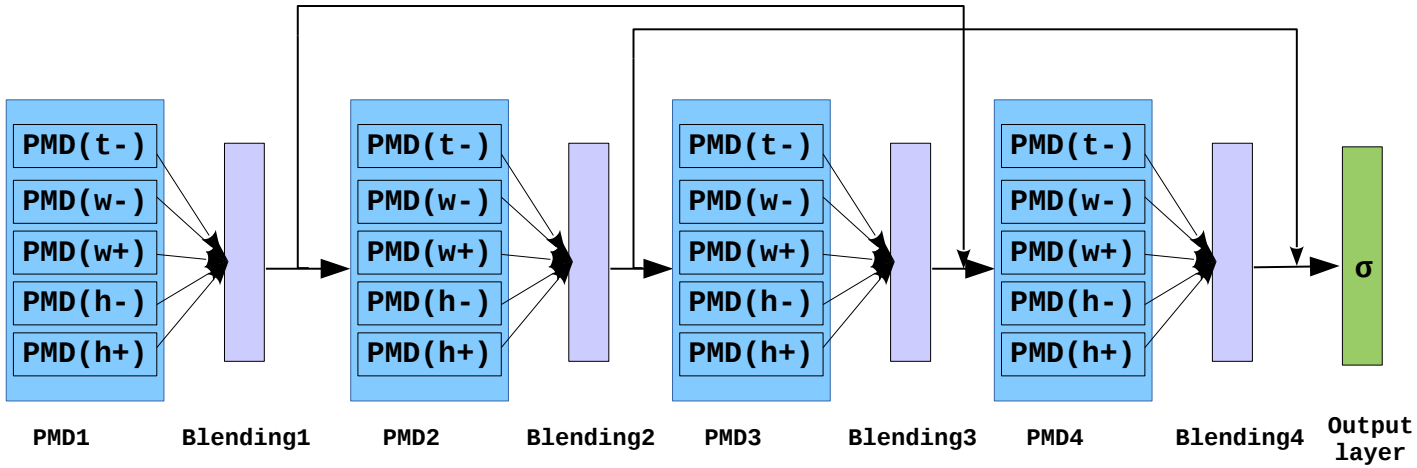
- Parallel MD-LSTM (PMD) Block: s^d, $d \in D = \{h\text{-}, h\text{+}, w\text{-}, w\text{+}, t\text{-}\}$
- Context Blending Block
  - Uniform Blending
  - Weighted Blending
- **Architecture: A stack of 4 PMD and Context Blending Blocks**

Fully Context-Aware Video Prediction: ContextVP



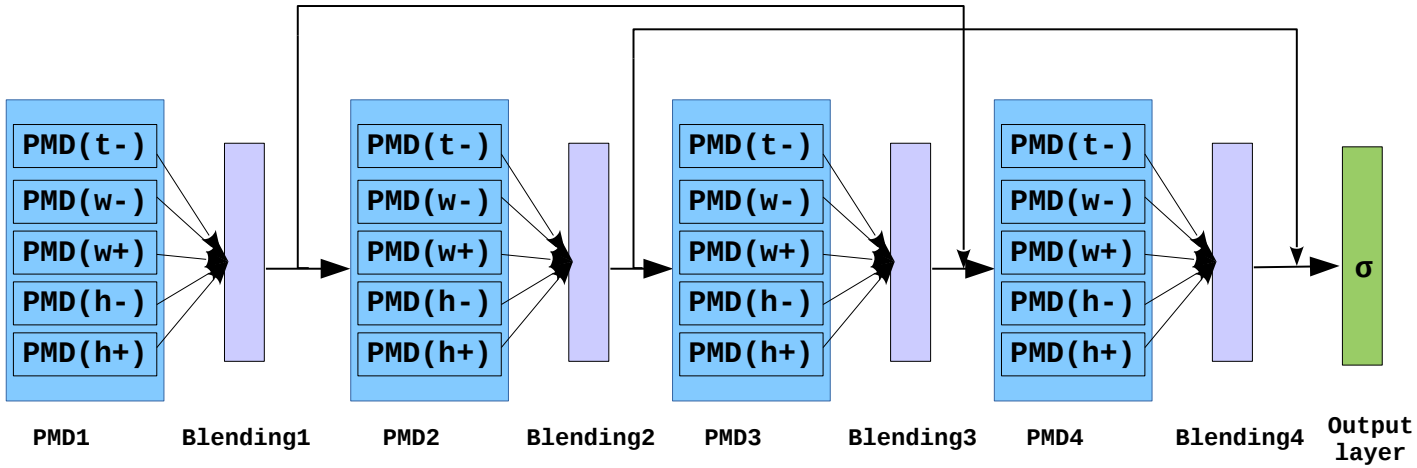
- Parallel MD-LSTM (PMD) Block: s^d, $d \in D = \{h\text{-}, h\text{+}, w\text{-}, w\text{+}, t\text{-}\}$
- Context Blending Block
  - Uniform Blending
  - Weighted Blending
- **Architecture: A stack of 4 PMD and Context Blending Blocks + 2 skip connections**

Fully Context-Aware Video Prediction: ContextVP



- Parallel MD-LSTM (PMD) Block: s^d, $d \in D = \{h\text{-}, h\text{+}, w\text{-}, w\text{+}, t\text{-}\}$
- Context Blending Block
  - Uniform Blending
  - Weighted Blending
- Architecture: A stack of 4 PMD and Context Blending Blocks
  + 2 skip connections
- **Regularization via Directional Weight Sharing (DWS)**: (h-, h+), (w-, w+)
- **Loss: $L1$ + Gradient Difference Loss (GDL)** [Mathieu15]

Fully Context-Aware Video Prediction: ContextVP



- Parallel MD-LSTM (PMD) Block: s^d, $d \in D = \{h\text{-}, h\text{+}, w\text{-}, w\text{+}, t\text{-}\}$
- Context Blending Block
  - Uniform Blending
  - Weighted Blending
- Architecture: A stack of 4 PMD and Context Blending Blocks + 2 skip connections
- Regularization via Directional Weight Sharing (DWS): (h-, h+), (w-, w+)
- Loss: $L1$ + Gradient Difference Loss (GDL) [Mathieu15]

Ablation Study

task: next-frame prediction, dataset: Human 3.6M, input: 10 frames

name	# layers	blending type	DWS	# parameters	PSNR	SSIM
ContextVP1	1	uniform (U)	N	0.7M	38.1	0.990
ContextVP3	3	uniform (U)	N	1.6M	41.2	0.992
ContextVP4-U-big	4	uniform (U)	N	14.0M	42.3	0.994
ContextVP4-W-big	4	weighted (W)	N	14.2M	44.8	0.996
ContextVP4-WD-small	4	weighted (W)	Y	2.0M	45.0	0.996
ContextVP4-WD-big	4	weighted (W)	Y	8.6M	45.2	0.996
PredNet [Lotter16]	-	-	-	6.9M	38.9	-

Number of layers

Ablation Study

task: next-frame prediction, dataset: Human 3.6M, input: 10 frames

name	# layers	blending type	DWS	# parameters	PSNR	SSIM
ContextVP1	1	uniform (U)	N	0.7M	38.1	0.990
ContextVP3	3	uniform (U)	N	1.6M	41.2	0.992
ContextVP4-U-big	4	uniform (U)	N	14.0M	42.3	0.994
ContextVP4-W-big	4	weighted (W)	N	14.2M	44.8	0.996
ContextVP4-WD-small	4	weighted (W)	Y	2.0M	45.0	0.996
ContextVP4-WD-big	4	weighted (W)	Y	8.6M	45.2	0.996
PredNet [Lotter16]	-	-	-	6.9M	38.9	-

Uniform Blending (U) / Weighted Blending (W)

Ablation Study

task: next-frame prediction, dataset: Human 3.6M, input: 10 frames

name	# layers	blending type	DWS	# parameters	PSNR	SSIM
ContextVP1	1	uniform (U)	N	0.7M	38.1	0.990
ContextVP3	3	uniform (U)	N	1.6M	41.2	0.992
ContextVP4-U-big	4	uniform (U)	N	14.0M	42.3	0.994
ContextVP4-W-big	4	weighted (W)	N	14.2M	44.8	0.996
ContextVP4-WD-small	4	weighted (W)	Y	2.0M	45.0	0.996
ContextVP4-WD-big	4	weighted (W)	Y	8.6M	45.2	0.996
PredNet [Lotter16]	-	-	-	6.9M	38.9	-

Directional Weight Sharing

Yes / No

Ablation Study

task: next-frame prediction, dataset: Human 3.6M, input: 10 frames

name	# layers	blending type	DWS	# parameters	PSNR	SSIM
ContextVP1	1	uniform (U)	N	0.7M	38.1	0.990
ContextVP3	3	uniform (U)	N	1.6M	41.2	0.992
ContextVP4-U-big	4	uniform (U)	N	14.0M	42.3	0.994
ContextVP4-W-big	4	weighted (W)	N	14.2M	44.8	0.996
ContextVP4-WD-small	4	weighted (W)	Y	2.0M	45.0	0.996
ContextVP4-WD-big	4	weighted (W)	Y	8.6M	45.2	0.996
PredNet [Lotter16]	-	-	-	6.9M	38.9	-

Model Size

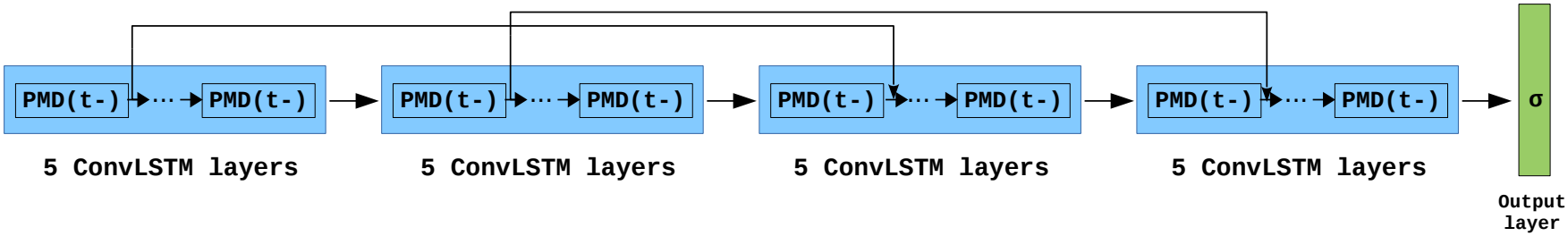
big / small

Ablation Study

task: next-frame prediction, dataset: Human 3.6M, input: 10 frames

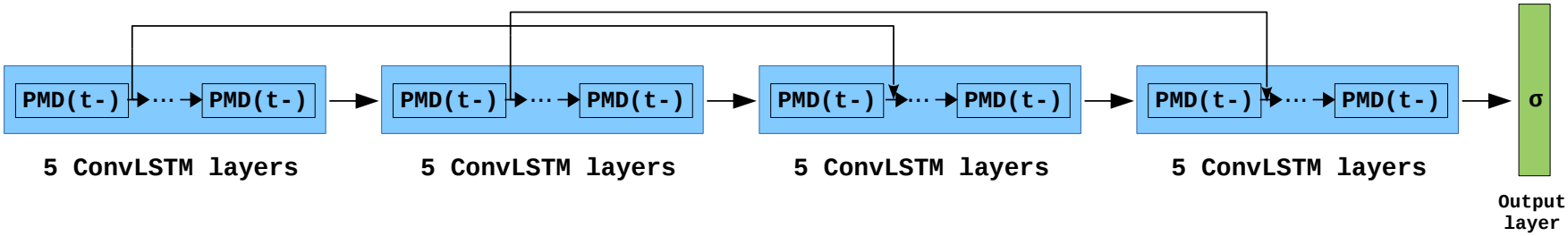
name	# layers	blending type	DWS	# parameters	PSNR	SSIM
ContextVP1	1	uniform (U)	N	0.7M	38.1	0.990
ContextVP3	3	uniform (U)	N	1.6M	41.2	0.992
ContextVP4-U-big	4	uniform (U)	N	14.0M	42.3	0.994
ContextVP4-W-big	4	weighted (W)	N	14.2M	44.8	0.996
ContextVP4-WD-small	**4**	**weighted (W)**	**Y**	**2.0M**	**45.0**	**0.996**
ContextVP4-WD-big	**4**	**weighted (W)**	**Y**	**8.6M**	**45.2**	**0.996**
PredNet [Lotter16]	-	-	-	6.9M	38.9	-

Baseline



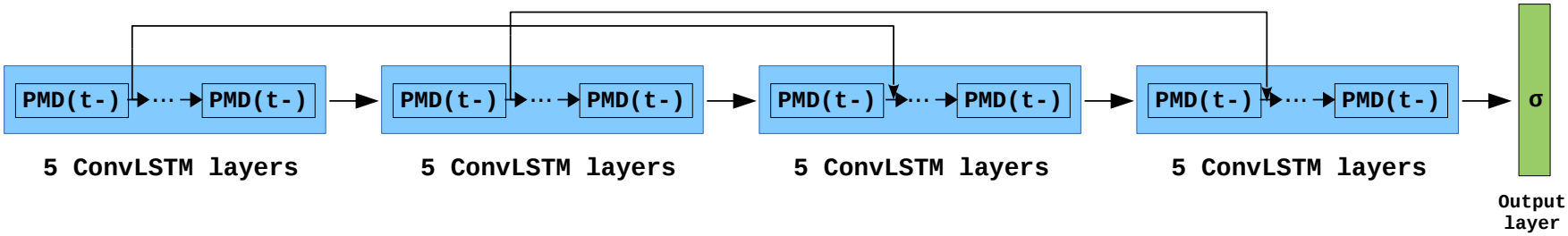
- Architecture: a stack of **20 ConvLSTM layers + 2 skip connections**
- The number of parameters is comparable to our best model

Baseline



- Architecture: a stack of **20 ConvLSTM layers + 2 skip connections**
- The number of parameters is comparable to our best model
- **It outperforms almost all state of the art models!**
  (except Deep Voxel Flow [Liu17] on UCF-101)

Baseline



- Architecture: a stack of **20 ConvLSTM layers + 2 skip connections**
- The number of parameters is comparable to our best model
- **It outperforms almost all state of the art models!**
  (except Deep Voxel Flow [Liu17] on UCF-101)
- Comparison to ContextVP?
  - ContextVP **performs better** than the baseline with the similar number of parameters
  - ContextVP is more suitable for **parallelization**:
    pixel-level and direction-level parallelization are possible

Next-Frame Prediction

Method	MSE ($\times 10^{-3}$)	PSNR	SSIM	#parameters
Copy-Last-Frame	7.95	23.3	0.779	-
BeyondMSE [Mathieu15]	3.26	-	0.881	-
PredNet [Lotter16]	2.42	27.6	0.905	6.9M
Dual Motion GAN [Liang17]	2.41	-	0.899	113M
ConvLSTM20 (baseline)	2.26	28.0	0.913	9.0M
ContextVP4-WD-small	2.11	28.2	0.912	2.0M
ContextVP4-WD-big	**1.94**	**28.7**	**0.921**	8.6M

Train: KITTI dataset, Test: CalTech Pedestrian dataset, 10 input frames

CalTech Pedestrian dataset video frames

Next-Frame Prediction

input t = 8	Ground Truth	(ours) ContextVP	(baseline) ConvLSTM20	[Lotter16] PredNet

Train: KITTI dataset, Test: CalTech Pedestrian dataset, 10 input frames

Next-Frame Prediction

input t = 8	Ground Truth	(ours) ContextVP	(baseline) ConvLSTM20	[Lotter16] PredNet

Train: KITTI dataset, Test: CalTech Pedestrian dataset, 10 input frames

Next-Frame Prediction

input t = 8	Ground Truth	(ours) ContextVP	(baseline) ConvLSTM20	[Lotter16] PredNet

Train: KITTI dataset, Test: CalTech Pedestrian dataset, 10 input frames

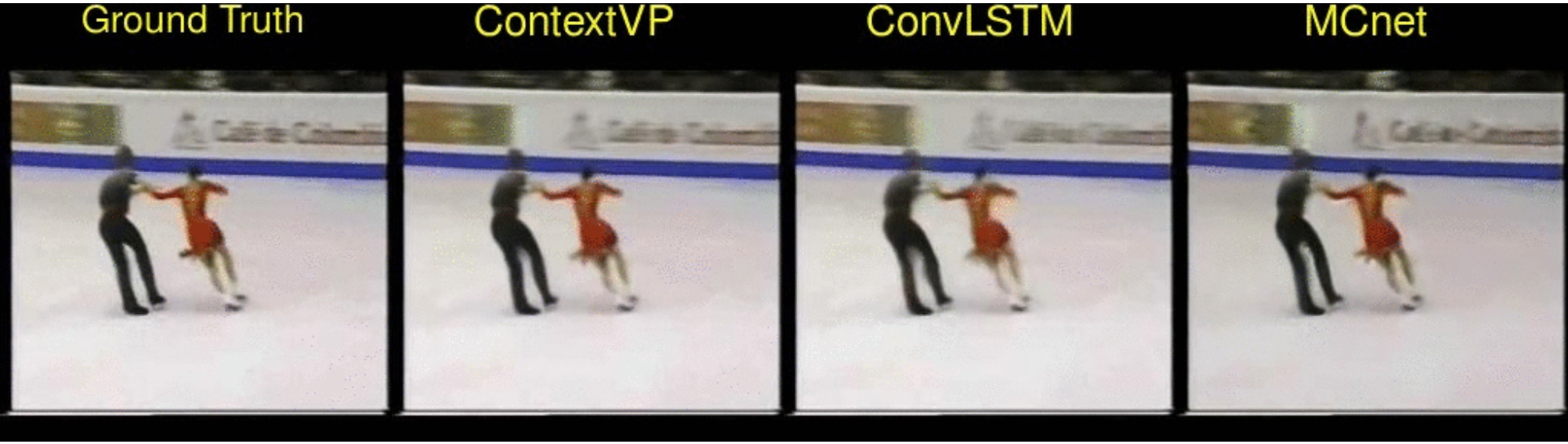
Multi-Frame Prediction

- Trained for next-frame prediction
- Tested to recursively predict 8 future frames

UCF-101 dataset, 4 input frames

Multi-Frame Prediction

- Trained for next-frame prediction
- Tested to recursively predict 8 future frames



UCF-101 dataset, 4 input frames

Conclusion

- ContextVP results in **improved performance** without the use of separation of motion and content, learning optical flow or adversarial training.
- ContextVP enables separation of depth and context. **Increasing depth is not necessary** to cover full context.
- **More computations can be parallelized** compared to very deep models.

Conclusion

What causes blurry predictions?

- **The future may be ambiguous given the past**
- The model does not utilize the all available information from the past

Thank you

wbyeon@nvidia.com

Poster: 1st floor, 4 - 6 PM

4B-02

project page

https://wonmin-byeon.github.io/publication/2018-eccv