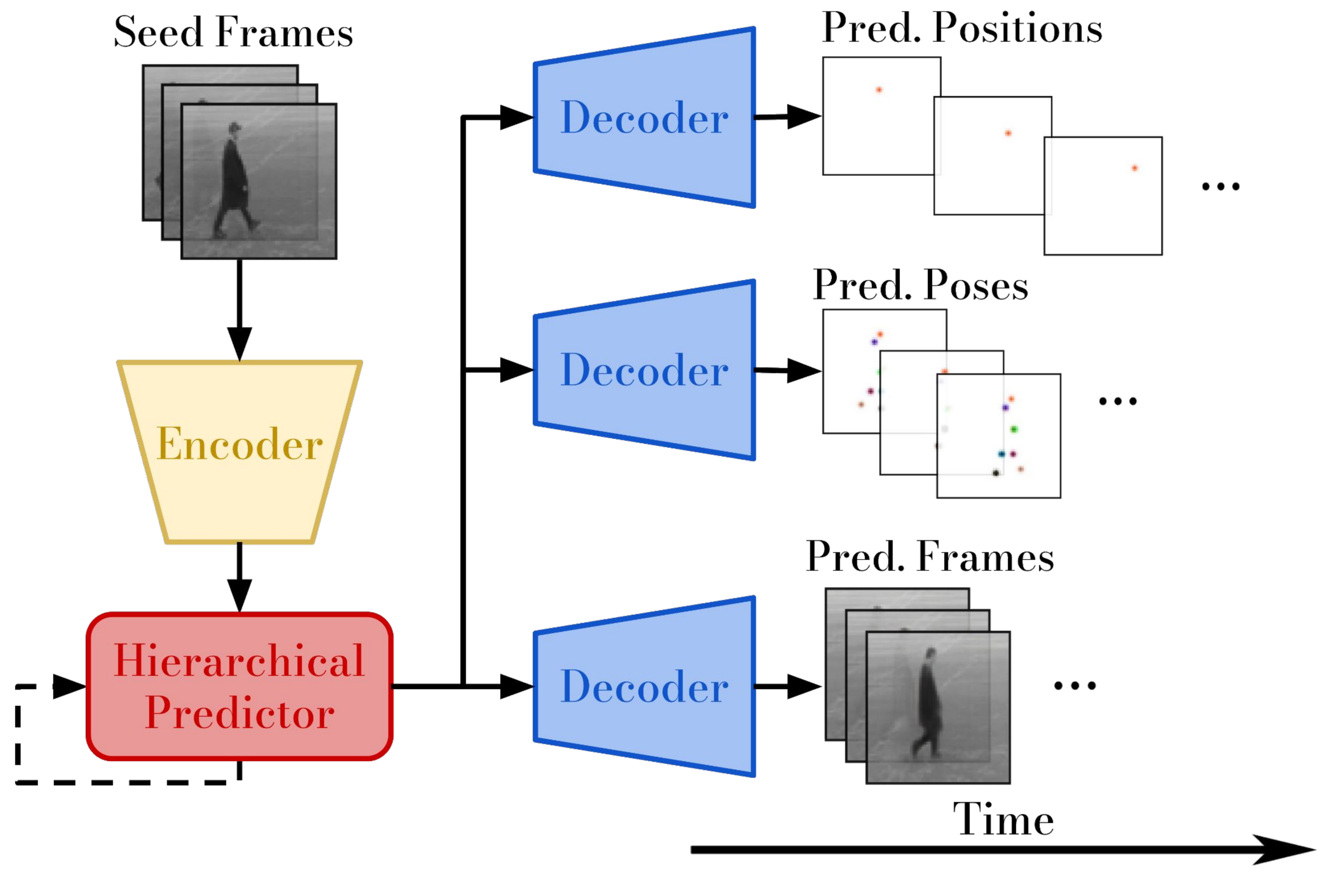
MSPred: Video Prediction at Multiple Spatio-Temporal Scales with Hierarchical Recurrent Networks

Autonomous Intelligent Systems, University of Bonn, Germany

Angel Villar-Corrales, Ani Karapetyan, Andreas Boltres, and Sven Behnke

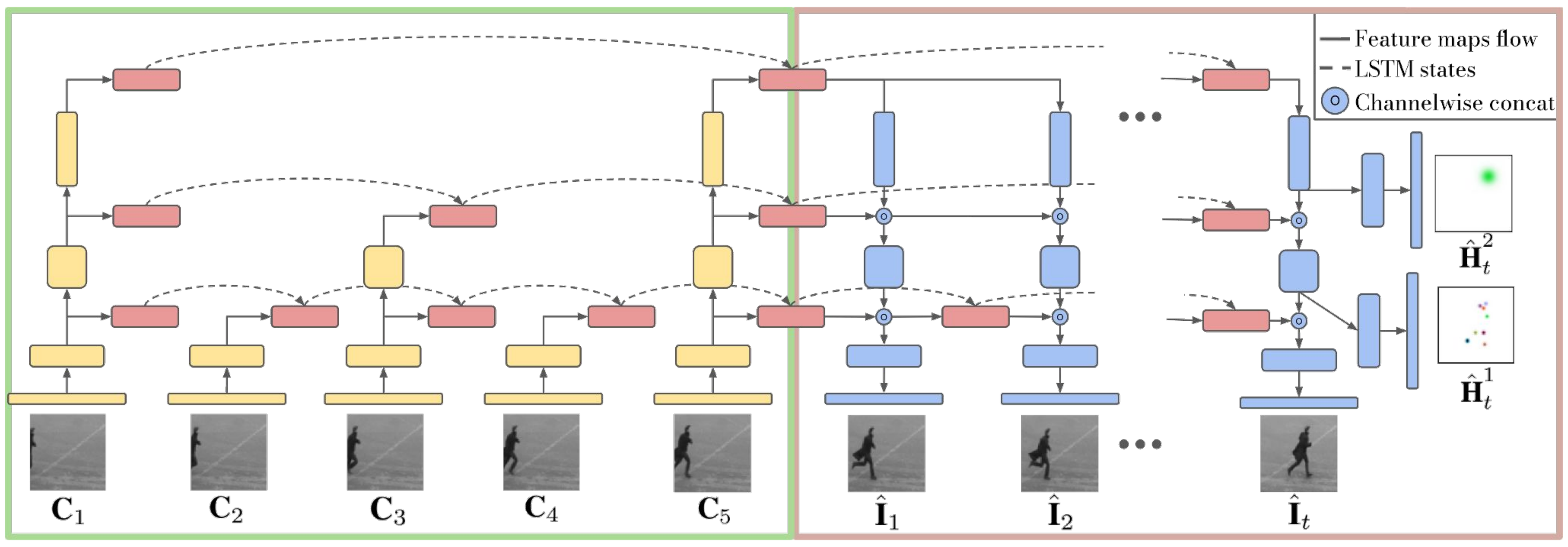
Problem

- **Video Prediction:** Given N seed video frames, generate plausible M subsequent frames.
- Useful in autonomous systems for:
  - Anticipative behavior planning
  - Enabling Human-Robot interaction and collaboration
- **Challenges:**
  - Precise details cannot be foreseen long into the future
  - Frames are often not the most useful representation, leading to blurry predictions
  - ➢ Existing models often not useful for autonomous systems' applications
- **Our approach:** Multi-scale prediction (**MSPred**)
  - Forecasting details (i.e. subsequent video frames) for short time horizons
  - Predicting abstract representations (e.g. poses or semantics) long into the future using coarse temporal resolutions



Proposed Model

- **Convolutional Encoder:** Maps frames to feature maps of increasingly coarser spatial resolution.
- **Predictor:** Three recurrent modules operating at different temporal resolutions:
  - Lowest level processes all inputs and models fast changing details.
  - Higher levels operate with coarser temporal resolutions and model more abstract features.
- **Multi-Scale Decoder:**
  - Three different decoder heads operating at different spatio-temporal resolutions.
  - Each head makes predictions of distinct level of abstraction, e.g., frames, poses and positions.
  - Each head uses the most recent feature maps from current and above hierarchy levels.



- MSPred operates as follows:
  - **Seed stage (left):** Encoding seed frames and feeding features to the recurrent modules.
  - **Prediction stage (right):** Autoregressively forecasting future representations and making predictions of different abstraction level. Images are predicted at every time-step, whereas higher-level representations are predicted with coarser temporal resolutions.

Evaluation

Quantitative Evaluation

1. Comparison with Existing Models

- MSPred outperforms SOTA models on three diverse datasets

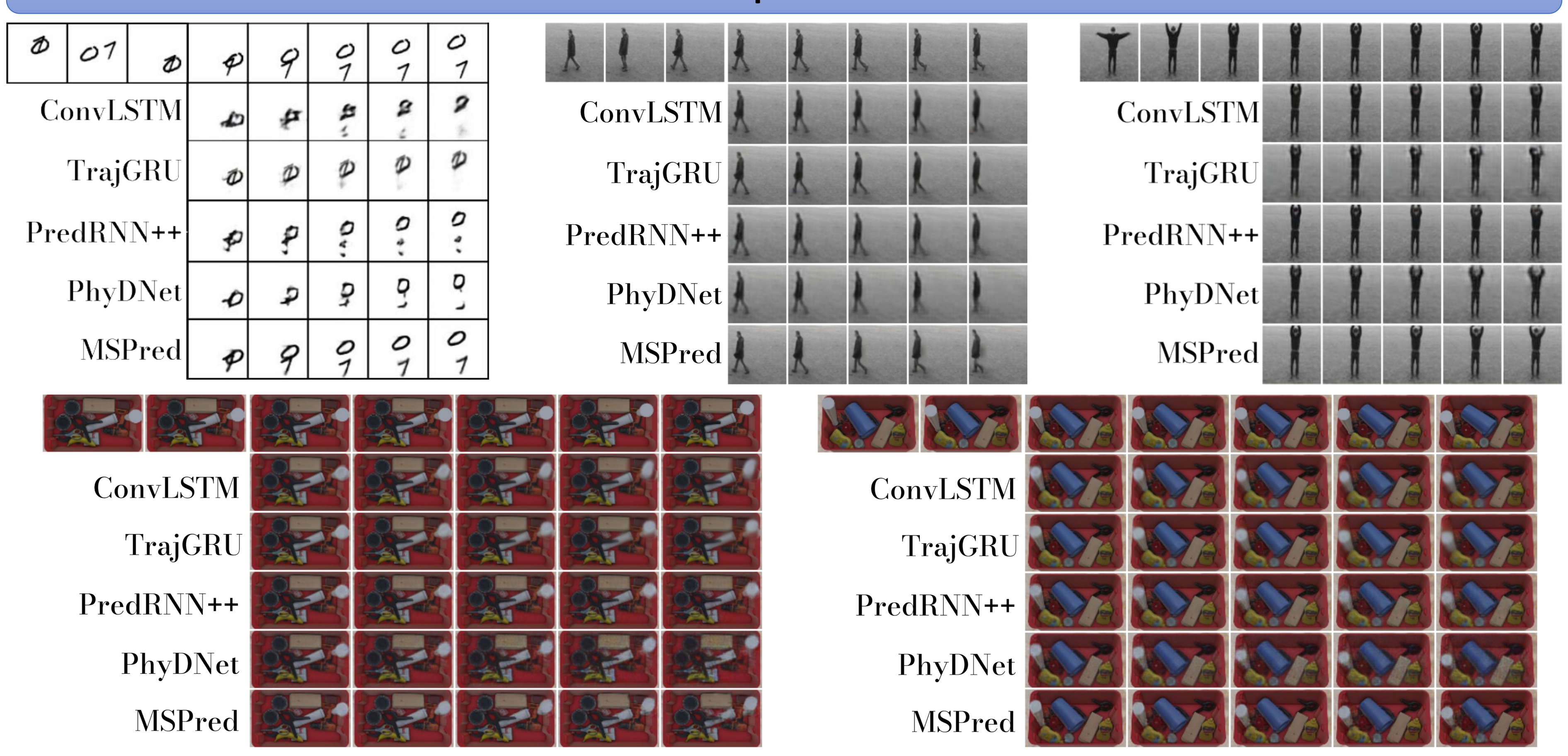
	Moving MNIST			KTH-Actions			SynpickVP		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
ConvLSTM [34]	17.22	0.833	0.144	29.93	0.957	0.048	27.98	0.907	0.059
TrajGRU [35]	20.02	0.895	0.075	**30.02**	**0.958**	0.039	28.10	**0.908**	0.041
SVG-Det [6]	20.31	0.900	0.114	26.64	0.927	0.068	26.92	0.879	0.068
SVG-LP [6]	20.36	0.907	0.115	27.60	0.932	0.063	27.38	0.886	0.066
PredRNN++ [42]	20.20	0.911	0.055	29.51	0.941	0.068	27.50	0.894	0.053
PhyDNet [11]	20.43	0.915	0.054	28.01	0.913	0.125	26.84	0.877	0.053
MSPred NoSup	25.94	**0.970**	**0.030**	28.65	0.929	0.034	**28.92**	0.902	0.031
MSPred (ours)	**25.99**	**0.970**	**0.030**	28.93	0.930	**0.032**	28.61	0.903	**0.030**

2. Ablation Study

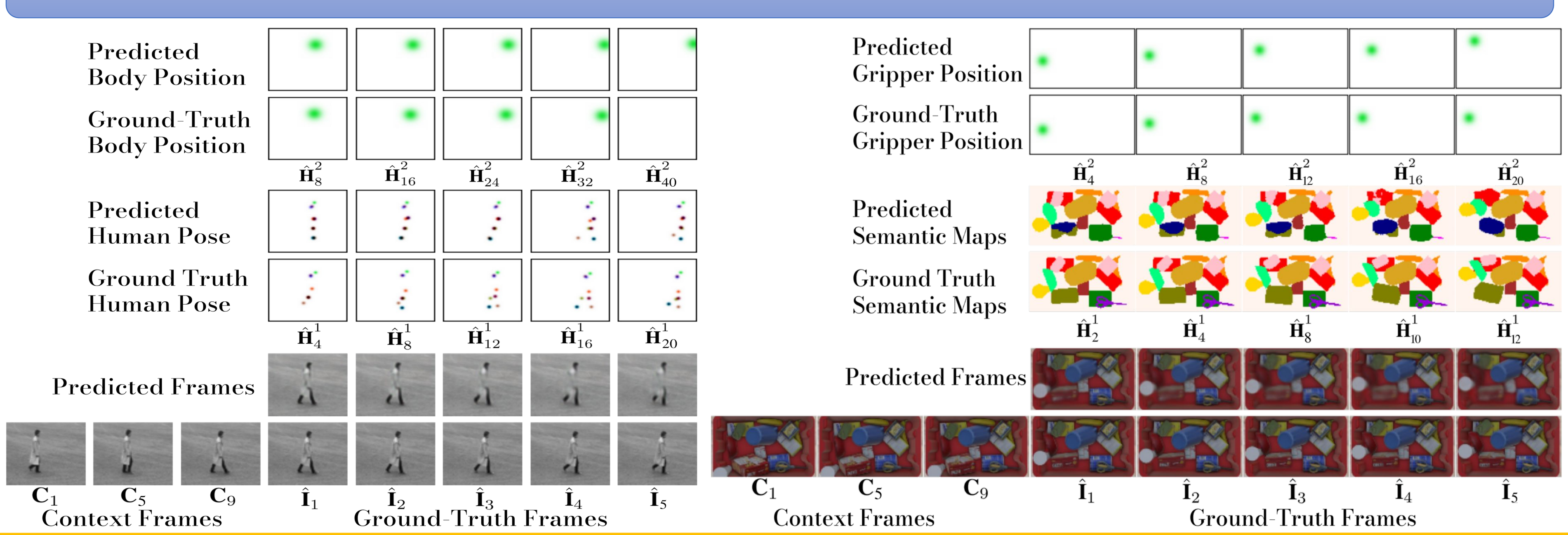
- Temporal and spatial hierarchy lead to best results
- Hierarchical supervision not a key factor for MSPred success

	MSPred Modules				Video Prediction Results			
	RNN	Spatial	Temporal	Hierarch. Supervision	MSE↓	PSNR↑	SSIM↑	LPIPS↓
1	Conv.	✓	✓	✓	**41.52**	**25.99**	**0.970**	**0.030**
2	Conv.	✓	✓	-	42.47	25.94	**0.970**	**0.030**
3	Linear	✓	✓	✓	208.71	17.95	0.827	0.202
4	Conv.	-	✓	✓	73.47	22.81	0.950	0.057
5	Conv.	✓	-	✓	92.45	20.81	0.921	0.093
6	Conv.	-	-	-	112.18	20.97	0.912	0.097

Comparison



Multi-Scale Predictions



Acknowledgement

This work was funded by grant BE 2556/16-2 (Research Unit FOR 2535 Anticipating Human Behavior) of the German Research Foundation (DFG).