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

Angel Villar-Corrales†[1]
villar@ais.uni-bonn.de

Ani Karapetyan†[1]
s6ankara@uni-bonn.de

Andreas Boltres[2, 3]
andreas.boltres@partner.kit.edu

Sven Behnke[1]
behnke@cs.uni-bonn.de

[1] Autonomous Intelligent Systems
University of Bonn
Endenicher Allee 19 A,
53115 Bonn, Germany

[2] SAP SE - Walldorf
Dietmar-Hopp-Allee 16
69190 Walldorf, Germany

[3] Autonomous Learning Robots
Karlsruhe Institute of Technology (KIT)
Adenauerring 4
76131 Karlsruhe, Germany

Abstract

Autonomous systems not only need to understand their current environment, but should also be able to predict future actions conditioned on past states, for instance based on captured camera frames. However, existing models mainly focus on forecasting future video frames for short time-horizons, hence being of limited use for long-term action planning. We propose *Multi-Scale Hierarchical Prediction* (MSPred), a novel video prediction model able to simultaneously forecast future possible outcomes of different levels of granularity at different spatio-temporal scales. By combining spatial and temporal downsampling, MSPred efficiently predicts abstract representations such as human poses or locations over long time horizons, while still maintaining a competitive performance for video frame prediction. In our experiments, we demonstrate that MSPred accurately predicts future video frames as well as high-level representations (e.g. keypoints or semantics) on bin-picking and action recognition datasets, while consistently outperforming popular approaches for future frame prediction. Furthermore, we ablate different modules and design choices in MSPred, experimentally validating that combining features of different spatial and temporal granularity leads to a superior performance. Code and models to reproduce our experiments can be found in https://github.com/AIS-Bonn/MSPred.

1 Introduction

For effective human-robot collaboration, autonomous systems, such as domestic robots, need not only to perceive and understand their surroundings, but should also be able to estimate the intentions of nearby agents and make predictions about their actions and behavior. Depending on the desired prediction time-horizon, the level of abstraction of the predicted

† denotes equal contribution.

representations might differ. On the one hand, when forecasting the immediate future, predictions of high level of detail, such as subsequent video frames, are desirable. On the other hand, for longer time horizons it is no longer possible to foresee exact details, hence it can be advantageous to predict more abstract representations like human poses or scene semantics.

Recently, several deep-learning-based methods [4, 6, 22, 28, 36] have been proposed for video prediction. These methods predict future frames in an autoregressive manner, conditioned on observed or generated images, often achieving realistic predictions for short time horizons or deterministic datasets.

Despite these successes, existing models are explicitly designed to predict future video frames, either in a self-supervised manner or in a supervised setting, thus lacking the flexibility to simultaneously make predictions of different abstraction level. Furthermore, these methods operate in an autoregressive manner in the image space, thus suffering from the increasing difficulty of predicting image-level details for longer time-horizons.

To overcome these issues, we propose *Multi-Scale Hierarchical Prediction* (MSPred), a convolutional and recurrent neural network designed to simultaneously predict future possible outcomes of different levels of abstraction with different spatial and temporal granularity.

To better model the scene dynamics and allow for better temporal modeling, MSPred utilizes a hierarchical predictor module, which applies both spatial and temporal downsampling in order to allow MSPred to extract features of different levels of abstraction that change at different temporal resolutions. The hierarchical predictor module is composed of multiple long short-term memory [13] (LSTM) cells operating at different periods (T), i.e., processing every T-th input frame. LSTMs operating at a higher frequency specialize on modeling low-level fast-changing features, whereas LSTMs with a lower frequency capture more abstract representations that change at slower rates. Using this coarser temporal resolution allows us to predict abstract features far into the future, while requiring only few recurrent iterations.

In summary, our contributions are: **(1)** We propose *MSPred*, a hierarchical video prediction model able to simultaneously predict future possible outcomes of distinct levels of abstraction at different spatio-temporal granularity, conditioned on past video frames. **(2)** MSPred outperforms popular video prediction models on perceptual metrics on three diverse datasets, while achieving more realistic predictions. **(3)** MSPred predicts plausible high-level representations (e.g. semantics or poses) long into the future using a coarse temporal resolution, while maintaining a competitive performance for video frame prediction.

2   Related Work

2.1   Future Frame Video Prediction

Video prediction is the task of forecasting future video frames conditioned on past frames. This task gained popularity due to its relevance for generating anticipative behavior. For a comprehensive review of deep-learning-based video prediction, we refer to Oprea *et al*. [22].

Several approaches have been proposed to perform video prediction, including modeling geometric transformations between consecutive frames [14, 21], learning transformations in the frequency domain [8, 9], or using optical flow [19, 24]. The most popular approach to video prediction, also followed by our proposed method, is the use of recurrent networks in combination with convolutional autoencoders in order to extract features from the seed frames and projecting them into the future [1, 11, 30, 31, 34, 37, 38]. These approaches were later extended by integrating variational inference into the models, thus allowing the neural networks to model the underlying uncertainty of the data. Stochastic prediction models [2,

4, 5] use an inference network to approximate the true posterior distribution over a set of latent variables, which model the stochastic properties of the data. Some methods, including MSPred, organize the network into a hierarchical structure [4, 26, 39], allowing the models to use features of different abstraction levels in order to predict future video frames.

The work that is conceptually most similar to ours is CW-VAE [28], in which – similarly to our MSPred model – a hierarchy of recurrent modules ticking at different clock rates is used to predict higher-level features using coarse time resolutions. Despite the similarities, the problems addressed by CW-VAE and MSPred are inherently different: CW-VAE combines features of distinct temporal resolutions to predict detailed frames far into the future. In contrast, we do not aim to forecast realistic frames long into the future, but instead to simultaneously predict detailed frames for short time horizons, as well as higher-level representations longer into the future using coarser temporal resolutions, which suffices for coarse-to-fine behavior planning in the now.

2.2 High-Level Feature Prediction

Another line of work performs video prediction using an intermediate high-level representation, instead of directly predicting future frames in the pixel space. These models first extract some high-level representation from the seed frames, either using pretrained models or human annotations, and project them into the future. Then, the model combines the predicted structured representations and the seed frames in order to forecast the future video frames. This approach simplifies the task of prediction, often leading to long-term accurate predictions, and has been proven successful for different high-level representations, including human poses [10, 35], semantic segmentation maps [23], or instance segmentations [20].

Similarly, MSPred also predicts high-level representations. However, instead of simply using them as intermediate features for future frame prediction, MSPred exploits the high-level of abstraction of these representations in order to forecast them long into the future using RNNs operating with coarse temporal resolutions.

2.3 Multi-Scale Recurrent Networks

Since the introduction of recurrent neural networks [7, 13] (RNNs), several approaches have been proposed to extend recurrent models into temporal hierarchies. Hihi and Bengio [12] propose different architectures utilizing several RNNs operating at different time scales in order to learn long-term dependencies on simple sequential tasks. *Clockwork RNNs* [17] split a recurrent network into parallel recurrent sub-modules that process their inputs at a different temporal granularity, hence allowing the model to learn complex dependencies between temporally distant inputs. Similarly, HM-RNN [5] proposes a multi-scale recurrent model with different modules operating at distinct time periods. However, the specific values for these rates are not fixed, but learned via an adaptive mechanism.

Like the previous methods, MSPred uses recurrent models operating at distinct time-scales to capture representations at different temporal resolutions. However, whereas previous methods processed low-dimensional sequences, we employ convolutional LSTMs [30] in combination with convolutional autoencoders to forecast high-dimensional video sequences.

3 Method

Video prediction is defined as the task of predicting subsequent video frames $\hat{\mathcal{I}} = \hat{\mathbf{I}}_1, \hat{\mathbf{I}}_2, ..., \hat{\mathbf{I}}_N$ conditioned on C past context frames $\mathcal{C} = \mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_C$. In this work, we extend the task of
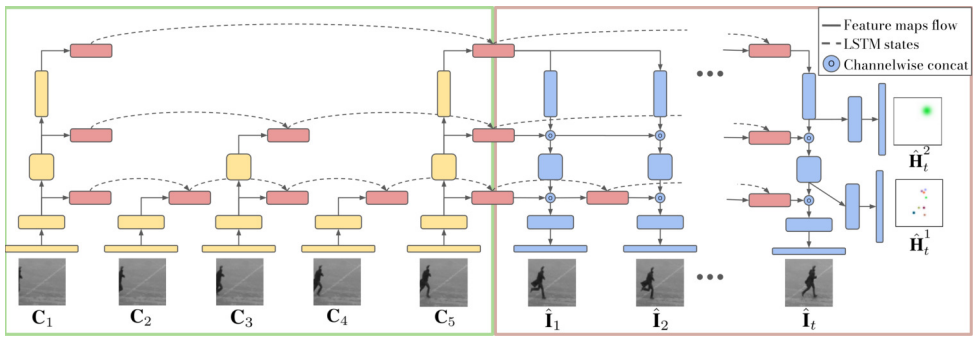
Figure 1: MSPred model. **Left**: Seed video frames are encoded with a **convolutional encoder**. **LSTM modules** ticking at different periods forecast features at future time steps. Slower LSTMs model abstract representations, whereas the fastest LSTM processes low-level details. **Right**: Predicted feature maps are fed through a **decoder**, which processes and fuses information from different LSTM modules. Finally, three distinct **decoder heads** predict outputs of different abstraction level. Note that we do not show the stochastic modules, and only display high-level decoder heads at the last time-step to unclutter the visualization.

video prediction to predict not only the future frames, but also higher-level representations $(\hat{\mathcal{H}}^1, \hat{\mathcal{H}}^2)$, such as human poses or object locations, conditioned on the same seed frames.

In this section we present MSPred, our proposed model for simultaneous prediction of representations of different levels of abstraction at multiple spatio-temporal scales, as illustrated in Figure 1. Its key component is a hierarchical predictor module (Sec. 3.2), which forecasts features of different granularity. These features are extracted from seed frames using a convolutional encoder (Sec. 3.1), and decoded into future frames or higher-level representations using convolutional decoders (Sec. 3.3). Additionally, we discuss the stochastic components of MSPred (Sec. 3.4), as well as training and implementation details (Sec. 3.5).

3.1 Encoder

MSPred uses a 2D convolutional encoder to process the seed frames. This module consists of four convolutional blocks, which extract increasingly abstract features of coarser spatial resolution. After the second and third blocks, skip connections bridge from the encoder to the decoder through the two lowest-level RNNs respectively, providing features of different levels of abstraction to the decoder. Supplying low-level representations of high-spatial resolution prevents the loss of information in the bottleneck layers, whereas higher-level features provide the decoder with abstract semantic information. The combination of these features allows our model to achieve feasible future predictions for increasingly complex structures.

3.2 Multi-Scale Prediction

Information in images is often processed in a hierarchical manner using features of different spatial resolution and level of abstraction. Similarly, the flow of information in videos can be represented in a temporal hierarchy. Higher-level features model the slowly changing information that is shared across many frames, whereas lower-level representations model faster-changing information. To account for this temporal hierarchy, MSPred uses a predictor module composed of three recurrent neural networks operating at different temporal resolu-

tions, i.e., processing input frames with a period of one, T_1 and T_2 respectively ($1 < T_1 < T_2$). As depicted in Figure 1, the recurrent module at a particular level receives feature maps from the respective stage of the encoder. The lowest-level RNN, which processes all inputs, receives low-level feature maps of high spatial resolution, which correspond to fine-grained representations that quickly change between consecutive frames. The second RNN receives feature maps of coarser spatial resolution, containing more slowly-changing representations, thus operating at a slower clock-rate of T_1. Finally, the highest-level RNN receives abstract feature maps of even coarser spatial resolution, which contain high-level features that are shared across many video frames, hence processing just one input every T_2 time steps.

The use of this hierarchy of RNNs allows our model to disentangle the temporal information into three different flows, each modeling features varying at distinct time-scales. Furthermore, the temporal abstraction in higher levels allows MSPred to forecast high-level features far into the future using just a small number of iterations, hence mitigating the error accumulation characteristic of autoregressive models.

3.3 Decoder

The decoder architecture corresponds to a mirrored version of the convolutional encoder, and is composed of four decoding stages. After the first and second one, features from the skip connections are fused with the decoded feature maps via channel-wise concatenation. As shown in Figure 1, MSPred uses three separate decoder heads in correspondence to its three levels of processing in order to predict representations of different level of abstraction. Forecasting detailed video frames requires high-level knowledge (i.e. semantics or dynamics) as well as low-level information (i.e. texture and color). Therefore, the lowest-level decoder uses the most recent predicted features from all levels of the hierarchy to predict the subsequent video frames. As higher levels operate at coarser time scales, predicted feature maps from higher levels are re-used until a new feature map is generated. The mid-level decoder produces more abstract representations (e.g. poses or semantic segmentation) every T_1 time steps by processing the most recent feature maps of its own level and the level above. Finally, the highest-level decoder generates abstract representations, such as person positions, every T_2 time steps using only the predicted features from the highest level.

3.4 Stochastic Components

Inspired by Denton and Fergus [6], we enhance our MSPred model with a stochastic component to account for the uncertainty of real-world data and to achieve more diverse predictions. During training, MSPred uses the current target frame $\mathbf{I}_t$, as well as all previous frames $\mathbf{I}_{1:t-1}$ to compute a posterior distribution $q_\phi(\mathbf{z}_t | \mathbf{I}_{1:t})$ and sample a latent variable $\mathbf{z}_t$, which encodes the dynamics of the sequence. By constraining the posterior to be close to a prior distribution $p_\psi(\mathbf{z}_t | \mathbf{I}_{1:t-1})$, we enforce the model to encode the dynamics of the sequence, instead of simply condensing information from the target frame. At inference time, MSPred combines previous frames $\mathbf{I}_{1:t-1}$ and latent vectors $\mathbf{z}_{1:t}$ to predict future frames and representations.

In practice, we use RNNs to estimate the mean and covariance of Gaussian prior and posterior distributions, and use the reparameterization trick [16] to sample latent vectors. Due to the hierarchical structure of MSPred, we employ separate recurrent prior p_ψ and posterior q_ϕ modules for each level. The latent vectors from each level are concatenated with the corresponding encoded features, and then fed to the respective predictor module [4, 53]. For a more detailed description of stochastic components in video prediction, we refer to [6].

3.5   Training and Implementation Details

Given a sequence of seed images, MSPred encodes these frames and feeds the embedded features to the corresponding recurrent modules (Figure 1 left). During the prediction stage (Figure 1 right), the model forecasts future representations in an autoregressive manner in the feature space, i.e., the output of a recurrent module is used as input in the subsequent time step. The forecasted features are fed to the corresponding decoder stage in order to decode future frames and high-level representations. Images are predicted at every time-step, whereas higher-level representations are predicted at the same clock-rate as their corresponding higher-level recurrent module, i.e., once every T_1 and T_2 time-steps, respectively.

For a fair comparison with baseline methods, for the Moving MNIST dataset our encoder and decoder follow DCGAN-like [27] discriminator and generator architectures, respectively; whereas for other datasets we use VGG16-like [32] modules. Each level of our hierarchical predictor uses four ConvLSTM [30] cells. The stochastic components are implemented using, for each level in the hierarchy, a single ConvLSTM cell operating at the same frequency as the corresponding predictor module. We provide further implementation details in the supplementary material.

Similar to [6, 36], we include skip connections from the last observed context frame to the decoder for all prediction time-steps. The features from the skip connections are added to the outputs of the corresponding ConvLSTM. The role of these skip connections is to directly provide the decoder with features of the background and static objects, hence allowing the predictor to focus on modeling pixel-level dynamics that change throughout the sequence.

Given C seed frames, we train MSPred to make five predictions ($N = 5$) on each level, hence predicting the five subsequent video frames, and forecasting higher-level representations for five prediction steps with a temporal resolution of T_1 and T_2, respectively. Our model is trained using the Adam optimizer [15] and the following loss function:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(||\mathbf{I}_i - \hat{\mathbf{I}}_i||^2 + \lambda_1 \mathcal{L}_{H^1}(\mathbf{H}^1_{T_1 i}, \hat{\mathbf{H}}^1_{T_1 i}) + \lambda_2 \mathcal{L}_{H^2}(\mathbf{H}^2_{T_2 i}, \hat{\mathbf{H}}^2_{T_2 i}) + \beta D_{KL}(q_\phi, p_\psi) \right),$$

where $|| \cdot ||$ is the ℓ_2 norm, $\mathbf{I}_t, \mathbf{H}^1_t, \mathbf{H}^2_t$ correspond to the ground truth frames and higher-level targets at time-step t, $\hat{\mathbf{I}}_t, \hat{\mathbf{H}}^1_t, \hat{\mathbf{H}}^2_t$ are the corresponding predictions of each level at time-step t, and λ_1, λ_2 and β are coefficients to weight the different loss terms. $\mathcal{L}_{H^1}$ and $\mathcal{L}_{H^2}$ correspond the loss functions used on the higher-level decoder heads. We employ the mean-squared error when predicting poses or positions, and pixel-wise cross-entropy when forecasting semantic maps. Finally, D_{KL} corresponds to the KL-divergence error between the posterior q_ϕ and prior p_ψ distributions estimated by the stochastic component, averaged over the three levels in the hierarchy.

4   Experiments

We perform an ablation study investigating several MSPred modules and design choices in Section 4.2. In Section 4.3 we compare MSPred with existing video prediction models on three diverse datasets. Finally, in Section 4.4 we present results of multi-scale prediction. Animations and further qualitative results are available in the project website[1].

---

[1] https://sites.google.com/view/mspred/home

Table 1: Ablation study investigating several MSPred modules and design choices, i.e. different RNNs, temporal and spatial hierarchy, and the effect of supervision on the higher levels of the hierarchy. Best result is highlighted in boldface, second best is underlined.

		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.	✓	✓		-	<u>42.47</u>	<u>25.94</u>	**0.970**	**0.030**
3	Linear	✓	✓		✓	208.71	17.95	0.827	0.202
4	Conv.	-	✓		✓	73.47	22.81	<u>0.950</u>	<u>0.057</u>
5	Conv.	✓	-		✓	92.45	20.81	0.921	0.093
6	Conv.	-	-		-	112.18	20.97	0.912	0.097

4.1 Datasets

We evaluate MSPred for different prediction tasks on three video datasets of different levels of complexity, namely Moving MNIST [34], KTH-Actions [29], and SynpickVP. In all cases, we average the results across five predicted frames or high-level representations. Further details about datasets and evaluation metrics are provided in the supplementary material.

Moving MNIST is a standard video prediction dataset containing sequences of two random digits from the MNIST dataset [18] moving with constant speed in a 64×64 grid, and bouncing off the image boundaries. We train our models on random sequences generated on the fly, and evaluate on a fix test set containing 10,000 sequences.

KTH-Actions is a dataset consisting of real videos of humans performing one out of six possible actions, e.g. jogging or waving. The dataset includes 600 videos of 25 different humans performing the actions in various indoor and outdoor environments.

SynpickVP is a video prediction dataset containing sequences of bin-picking scenarios, in which a suction-cap gripper robot moves objects in a cluttered box. We generate the dataset by selecting sequences from the recently proposed SynPick [25] dataset. This is a challenging benchmark, in which the model needs to predict the motion of the robotic gripper, as well as the displaced objects, while representing a complex and cluttered background.

4.2 Ablation Study

We train modified variants of MSPred to investigate how different components and design choices affect the video prediction performance of our model. Namely, we evaluate the relevance of temporal and spatial hierarchy, the type of RNN used, and the effect of multi-level supervision. For our ablation study, we focus on the Moving MNIST dataset. The results are reported in Table 1.

First, the hierarchical MSPred model (Rows 1 and 2) outperforms all other variants, whereas removing the spatial (Row 4), temporal (Row 5), or all (Row 6) hierarchical structure from the predictor leads to loss of performance, thus demonstrating that providing features of different spatio-temporal granularity improves the prediction performance of the model. Second, MSPred (Row 1) performs comparably to an MSPred variant trained only for image-level prediction (Row 2), thus indicating that adding hierarchical supervision to the loss function is not a key factor for the success of our model for future frame prediction. Finally, when replacing the ConvLSTM predictor with a linear LSTM (Row 3), the performance is significantly decreased, leading to the worst results among all compared models.

Table 2: Quantitative comparison between video prediction models. MSPred outperforms all other methods on Moving MNIST, and achieves the best perceptual results (LPIPS) on KTH-Actions and SynpickVP. Best result is highlighted in boldface, second best is underlined.

	Moving MNIST			**KTH-Actions**			**SynpickVP**		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
ConvLSTM [30]	17.22	0.833	0.144	<u>29.93</u>	<u>0.957</u>	0.048	27.98	<u>0.907</u>	0.059
TrajGRU [31]	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++ [38]	20.20	0.911	0.055	29.51	0.941	0.068	27.50	0.894	0.053
PhyDNet [11]	<u>20.43</u>	<u>0.915</u>	<u>0.054</u>	28.01	0.913	0.125	26.84	0.877	0.053
MSPred NoSup	<u>25.94</u>	**0.970**	**0.030**	28.65	0.929	<u>0.034</u>	**28.92**	0.902	<u>0.031</u>
MSPred (ours)	**25.99**	**0.970**	**0.030**	28.93	0.930	**0.032**	<u>28.61</u>	0.903	**0.030**

4.3   Comparison to Existing Methods

We compare MSPred for the task of future frame prediction with several existing video prediction methods, including ConvLSTM [30], TrajectoryGRU [31], two variants of SVG [6], PredRNN++ [38], and PhyDNet [11]. Additionally, we include an MSPred variant, denoted as *MSPred NoSup*, trained only for future frame prediction, i.e., without any additional supervision. For a fair comparison, all models use the same encoder and decoder architectures. The evaluation results are reported in Table 2.

We observe that both MSPred variants consistently perform among the best models on all datasets. Furthermore, MSPred NoSup shows comparable video prediction performance to MSPred trained to simultaneously make predictions of different levels of granularity. This indicates that adding hierarchical supervision to the loss function (e.g. poses or segmentation) allows MSPred to predict higher-level representations over long time-horizons, but it is not a key factor for the success of our model for video prediction.

Moving MNIST: In general, due to the simplicity of the dataset, all models achieve overall good prediction scores and accurate future frame predictions. However, both MSPred variants achieve exceptionally sharp and precise reconstructions, outperforming all other models by a large margin.

KTH-Actions: MSPred produces low PSNR and SSIM scores, indicating higher pixel differences with respect to the target frames. However, MSPred achieves the best LPIPS result, demonstrating a high perceptual similarity to the target frames. Figure 2 contains a qualitative comparison on three KTH-Actions sequences. Baseline methods with the highest PSNR results, i.e. ConvLSTM and TrajGRU, blur the predictions around the arms and legs of the person, whereas our MSPred model achieves sharper predictions.

SynpickVP: MSPred outperforms existing video prediction models on this challenging dataset, achieving the highest PSNR result, and the best LPIPS perceptual score, indicating more realistic predictions than the baselines. Figure 3 depicts a qualitative comparison on the SynpickVP dataset. Baseline video prediction methods tend to blur the suction cap gripper as well as the objects moved by it, whereas MSPred achieves more realistic predictions.
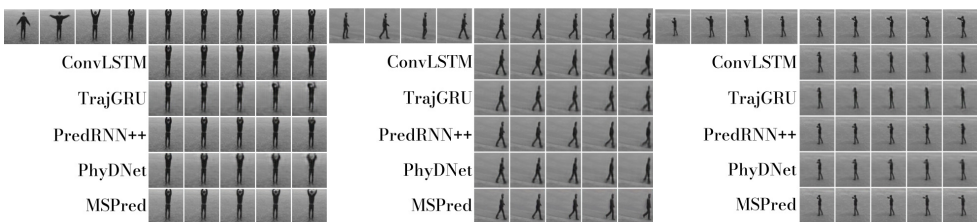
Figure 2: Qualitative results on the KTH-Actions dataset. Top row shows ground truth frames. We display four seed frames and five predictions for three test-set sequences. MSPred achieves the sharpest and most accurate predictions among the compared methods.
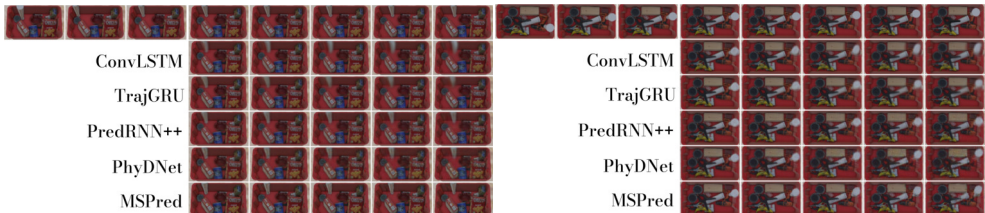

Figure 3: Qualitative results on the SynpickVP dataset. Top row shows ground truth frames. We display three seed frames and five predictions for two test-set sequences. Whereas all baseline methods blur the robot gripper in the predicted frames, MSPred achieves sharper predictions,

4.4 Multi-Scale Prediction

Unlike existing video prediction models, MSPred predicts higher-level representations with coarse temporal resolutions in addition to video frames. For KTH-Actions, MSPred predicts human keypoints on its intermediate level, and a center-point on its highest level. Keypoint annotations are obtained using OpenPose [3]. On SynpickVP, MSPred predicts semantic segmentation and the location of the gripper on the mid- and high-levels, respectively. We average the segmentation results in three different groups: *gripper*, *static*, and *background*.

For benchmarking the performance of MSPred, we train a baseline model *SVG'*, based on a modified SVG-LP [6] model, which predicts high-level representations conditioned on input frames. Further details about SVG' are provided in the supplementary material. Furthermore, we include a CopyLast baseline that copies the ground-truth representations of the last seed frame.

Table 3 reports a quantitative comparison for pose and semantic segmentation forecasting, respectively. MSPred outperforms the two baselines for human pose forecasting on the KTH-Actions dataset. On the SynpickVP dataset, MSPred precisely predicts the segmentation of the robot gripper. However, due to the complex and mostly static scenes, the CopyLast baseline outperforms SVG' and MSPred when forecasting the segmentation of static objects and background. This is due to the fact that CopyLast has access to the last ground-truth representations, thus having access to the perfect segmentation of static objects and background; whereas MSPred and SVG' predict the high-level representations condition on seed images, without access to ground truth segmentation maps.

Figure 4 depicts examples of multi-scale prediction on the KTH-Actions and SynpickVP datasets. The lowest-level decoder achieves detailed subsequent frame predictions for a short time horizon, whereas higher level decoders accurately predict abstract representations (e.g. poses or locations) up to 40 frames into the future using coarser temporal resolutions.
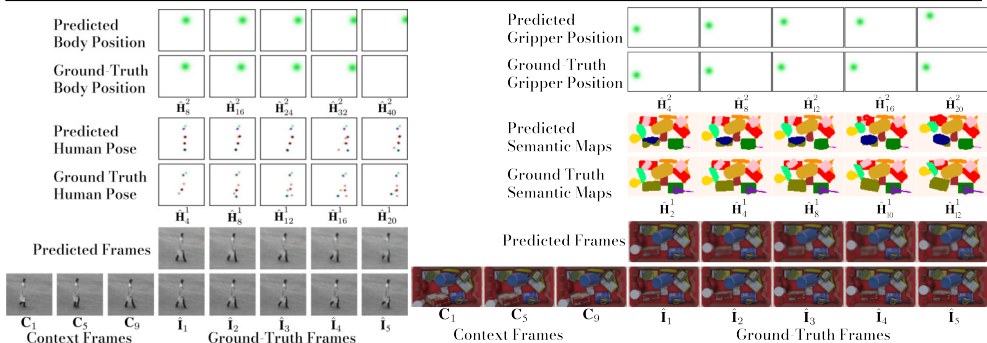
Figure 4: Multi-scale predictions on the KTH-Action (left) and SynpickVP (right) datasets. We display three seed frames and five targets and predictions for each decoder head. MSPred forecasts frames on short time horizons while also predicting higher-level representations longer into the future using a coarse temporal resolution.

Table 3: Comparison between MSPred and two baselines on pose prediction and segmentation forecasting. Best result is highlighted in boldface, second best is underlined.

| | KTH-Actions | | | | | | SynpickVP | | | | |
| | Pose Forecasting | | | | Gripper Seg. | | Static Seg. | | Backgr. Seg. | |
	MPJPE↓	PCK↑	PDJ↑	AP↑	Acc↑	IoU↑	Acc↑	IoU↑	Acc↑	IoU↑
CopyLast	10.50	0.550	0.558	0.620	0.397	0.324	**0.557**	**0.455**	**0.957**	**0.919**
SVG' [6]	<u>4.32</u>	<u>0.810</u>	**0.888**	<u>0.825</u>	0.606	0.443	<u>0.255</u>	<u>0.197</u>	0.956	0.898
MSPred	**3.41**	**0.812**	<u>0.867</u>	**0.833**	**0.651**	**0.471**	0.202	0.151	0.952	<u>0.901</u>

5   Conclusion

We proposed MSPred, a novel video prediction model that extends the effective prediction horizon of related approaches by leveraging hierarchies of recurrent neural networks operating at different spatio-temporal resolutions in order to predict outcomes of varying levels of abstraction with different granularity. At its lowest prediction level, MSPred forecasts subsequent video frames, whereas at higher levels it predicts more abstract representations longer into the future using coarser spatio-temporal resolutions. In our experiments, we show how MSPred outperforms several existing video prediction methods for the task of future frame prediction. Furthermore, we show how the higher level decoder heads can be used to forecast more abstract representations, such as human poses or semantic maps, over longer time horizons. We firmly believe that the hierarchical features from MSPred could be used as representations to improve perception and reasoning capabilities in autonomous agents, and serve as basis for planning anticipative behavior.

Acknowledgments

References

[1] Adil Kaan Akan, Erkut Erdem, Aykut Erdem, and Fatma Güney. SLAMP: Stochastic latent appearance and motion prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[2] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H. Campbell, and Sergey Levine. Stochastic variational video prediction. In *International Conference on Learning Representations (ICLR)*, 2018.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2017.

[4] Lluis Castrejon, Nicolas Ballas, and Aaron Courville. Improved conditional VRNNs for video prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7608–7617, 2019.

[5] Junyoung Chung, Sungjin Ahn, and Yoshua Bengio. Hierarchical multiscale recurrent neural networks. *International Conference on Learning Representations (ICLR)*, 2017.

[6] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *International Conference on Machine Learning (ICML)*, pages 1174–1183, 2018.

[7] Jeffrey L Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

[8] Hafez Farazi and Sven Behnke. Motion segmentation using frequency domain transformer networks. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2020.

[9] Hafez Farazi, Jan Nogga, and Sven Behnke. Local frequency domain transformer networks for video prediction. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2021.

[10] Naoya Fushishita, Antonio Tejero-de Pablos, Yusuke Mukuta, and Tatsuya Harada. Long-term human video generation of multiple futures using poses. In *European Conference on Computer Vision (ECCV)*, pages 596–612, 2020.

[11] Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11474–11484, 2020.

[12] Salah Hihi and Yoshua Bengio. Hierarchical recurrent neural networks for long-term dependencies. *Advances in Neural Information Processing Systems (NeurIPS)*, 8, 1995.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[14] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 29, 2016.

[15] Diederik P Kingma and Jimmy Ba. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[16] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[17] Jan Koutnik, Klaus Greff, Faustino Gomez, and Juergen Schmidhuber. A clockwork RNN. In *International Conference on Machine Learning (ICML)*, pages 1863–1871, 2014.

[18] Yann LeCun, Corinna Cortes, and Christopher JC Burges. The MNIST database of handwritten digits. *URL http://yann. lecun. com/exdb/mnist*, 10:34, 1998.

[19] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Flow-grounded spatial-temporal video prediction from still images. In *European Conference on Computer Vision (ECCV)*, pages 600–615, 2018.

[20] Pauline Luc, Camille Couprie, Yann Lecun, and Jakob Verbeek. Predicting future instance segmentation by forecasting convolutional features. In *European Conference on Computer Vision (ECCV)*, pages 584–599, 2018.

[21] Vincent Michalski, Roland Memisevic, and Kishore Konda. Modeling deep temporal dependencies with recurrent grammar cells. *Advances in Neural Information Processing Systems (NeurIPS)*, 27, 2014.

[22] Sergiu Oprea, Pablo Martinez-Gonzalez, Alberto Garcia-Garcia, John Alejandro Castro-Vargas, Sergio Orts-Escolano, Jose Garcia-Rodriguez, and Antonis Argyros. A review on deep learning techniques for video prediction. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(6):2806–2826, 2020.

[23] Junting Pan, Chengyu Wang, Xu Jia, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Video generation from single semantic label map. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3733–3742, 2019.

[24] Viorica Pătrăucean, Ankur Handa, and Roberto Cipolla. Spatio-temporal video autoencoder with differentiable memory. In *International Conference on Learning Representations (ICLR) Workshop*, 2016.

[25] Arul Selvam Periyasamy, Max Schwarz, and Sven Behnke. SynPick: A dataset for dynamic bin picking scene understanding. In *IEEE 17th International Conference on Automation Science and Engineering (CASE)*, pages 488–493, 2021.

[26] Isabeau Prémont-Schwarz, Alexander Ilin, Tele Hao, Antti Rasmus, Rinu Boney, and Harri Valpola. Recurrent ladder networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

[27] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *International Conference on Learning Representations, (ICLR)*, 2016.

[28] Vaibhav Saxena, Jimmy Ba, and Danijar Hafner. Clockwork variational autoencoders. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[29] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *IEEE International Conference on Pattern Recognition (ICPR)*, volume 3, pages 32–36, 2004.

[30] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wangchun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[31] Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations, (ICLR)*, 2015.

[33] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[34] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using LSTMs. In *International Conference on Machine Learning (ICML)*, pages 843–852, 2015.

[35] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In *International Conference on Machine Learning (ICML)*, pages 3560–3569, 2017.

[36] Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V Le, and Honglak Lee. High fidelity video prediction with large stochastic recurrent neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[37] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 879–888, 2017.

[38] Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, Philip Yu, and Mingsheng Long. PredRNN: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1–1, 2022.

[39] Bohan Wu, Suraj Nair, Roberto Martin-Martin, Li Fei-Fei, and Chelsea Finn. Greedy hierarchical variational autoencoders for large-scale video prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2318–2328, 2021.