Photo-Realistic Video Prediction on Natural Videos of Largely Changing Frames

Osamu Shouno[1][0000−0001−7385−3280]

Honda Research Institute Japan Co., Ltd., Wako, Saitama, Japan

Abstract. Recent advances in deep learning have significantly improved performance of video prediction. However, state-of-the-art methods still suffer from blurriness and distortions in their future predictions, especially when there are large motions between frames. To address these issues, we propose a deep residual network with the hierarchical architecture where each layer makes a prediction of future state at different spatial resolution, and these predictions of different layers are merged via top-down connections to generate future frames. We trained our model with adversarial and perceptual loss functions, and evaluated it on a natural video dataset captured by car-mounted cameras. Our model quantitatively outperforms state-of-the-art baselines in future frame prediction on video sequences of both largely and slightly changing frames. Furthermore, our model generates future frames with finer details and textures that are perceptually more realistic than the baselines, especially under fast camera motions.

Keywords: Video prediction, perceptual loss, GAN

1 Introduction

The capability of an artificial agent to forecast how a visual environment can evolve in the future can be applied to robotics, autonomous driving, healthcare, and action recognition. Video prediction is the task of predicting future frames given past video frames. A model is trained from unlabeled videos and learns representations for object and scene structures and transformations of their appearances. However, unsupervised learning of visual prediction on natural videos is challenging because of the diversity of objects and backgrounds in a scene, various sizes of object appearance, occlusions, disocclusions, camera movements, and other dynamic scene changes between frames.

Deep learning technologies have enabled a video prediction model to learn from large-scale, unlabeled real-world videos and improved its performance of future prediction [25,28,4,20,32,19,18,30,2,15,5]. One of the major unsolved problems in video prediction is blurry predictions. Two factors have been addressed as the primary causes of this artifact. Firstly, simple pixel-wise loss functions based on the mean squared error (MSE) or mean absolute error (MAE), which are often used for objective functions of video prediction models, are not capable of capturing long-range correlations among pixels which are characteristic

to natural images [29,23,12], nor accounting for uncertain futures [20,31,33]. Other objective functions, such as adversarial loss and perceptual loss functions, are shown to be effective at reducing blurry artifacts in the predicted frames [20,32,15,5], but not as effective as it is in other image generation tasks such as the super-resolution [11,16]. Secondly, network architectures can be inappropriate for video prediction. Byeon et al. [2] discussed the blind spot problem in which a video prediction model is not capable of accessing to entire available past information for predicting each pixel, increasing model uncertainty. Empirically, removing the blind spots improved performance of video prediction [2] .

Fig. 1. Qualitative comparison of predicted frames between the state-of-the-art model, ContextVP [2], and our model (GAN-VGG). ContextVP and our models are trained for next-frame prediction on the KITTI dataset, and tested to recursively predict future frames on the Caltech Pedestrian dataset. As an input for the prediction, ContextVP and our model take 10 and 8 frames, respectively. Our model reproduces finer details and textures compared to ContextVP as denoted by bounding boxes. Results are best viewed in color with zoom

However, current state-of-the-art video prediction models still suffer from blurriness and distorted object shapes in their predictions, especially when there are large camera movements [2,5] as illustrated in Fig. 1. Additionally, their next-frame predictions often lack fine details and textures and are overly smooth. Even if such prediction errors seem to be slight, they can quickly accumulate when recursively predicting further future frames due to the discrepancy between training and prediction, leading to large distortions of objects.

In this work, we propose a framework for video prediction that takes advantage of the adversarial and perceptual loss functions to generate accurate and photo-realistic future frames. In order to fully utilize the potential of these two loss functions, a neural network model for video prediction should be sufficiently powerful and expressive to be capable of modeling a variety of structures of objects, and various types of spatio-temporal transformations of their appearances not only at a low-frequency level but also at a high-frequency level. Inspired by the architecture of the PredNet [19], we propose a deep residual network with hierarchical architecture for video prediction where each layer models object and scene structures and their spatio-temporal transformations for making a prediction of their future state at different spatial resolution. These predictions of different layers are sequentially merged via top-down connections to generate future frames. Our model is trained end-to-end with the adversarial loss function computed using the generative adversarial network framework (GAN) [8,24] and perceptual loss functions based on differences between learned visual feature representations of pre-trained deep CNNs [6,1,11,16]. The main contribution of this paper is the development of a new method for predicting accurate and perceptually realistic future frames of natural videos even if there are large motions between frames. Examples of perceptually realistic future frames are shown in Fig. 1.

2 Related Work

Recent work on unsupervised learning of video prediction has tackled next-frame and longer-term prediction in synthetic and real-world videos. Srivastava et al. [28] proposed a recurrent neural network model featuring LSTM for predicting future frames. Oh et al. [22] proposed encoder-decoder neural network models for predicting future frames conditioned on actions. These models suffered from blurry predictions due to the pixel-wise MSE loss functions that inherently generate blurry results.

Later models have challenged to solve blurry predictions with different approaches. One approach focuses on loss functions, especially on adversarial loss functions computed from a GAN framework [8]. Mathieu et al. [20] proposed a GAN-based framework with multi-scale convolutional network for predicting sharper future frames. Vondrick et al. [32] defined a GAN-based model featuring differential transformers for generating future frames. Liang et al. [17] showed that shared internal representations for future-frame prediction and optical flow estimation improved future frame prediction. Kwon and Park [15] proposed a

GAN-based solution featuring retrospective cyclic constraints. Another approach considers effective usage of optical flows. Liu et al. [18] proposed a model that explicitly represents optical flow of pixels for generating future frames with warping function. Their results are sharp, however, tend to errors where flow predictions are incorrect or ill-posed. Gao et al. [5] augmented the flow-based solution with the inpainting module that hallucinates pixels for erroneous regions and with the perceptual loss functions. Other approaches focus on network architectures. Lotter et al. [19] proposed the PredNet architecture for unsupervised learning of videos based on the predictive coding, and demonstrated on challenging natural videos recorded from car-mounted cameras. Byeon et al. [2] pointed out the blind spot problem in network architecture that prevents effective usage of given past information for predicting each pixel. These approaches have improved performance of video prediction, however, still suffer from blurry, inaccurate predictions that lack fine details and textures.

Our approach considers both loss functions and network architecture for photo-realistic video prediction. Prior works on the single image super resolution [11,16] show that the adversarial and perceptual loss functions are effective for generating photo-realistic images with fine details and textures. They adopted a deep residual neural network with an encoder-decoder architecture in which most of residual blocks are used for processing at a lower spatial resolution. However, video prediction needs to handle a more diverse range of image transformations than the single image super-resolution. The works mentioned above [15,5] used frameworks very similar to the super-resolution [11,16], but their results are not qualitatively comparable to those of the super-resolution. We therefore focus on enhancing network capacity of a video prediction model by introducing a hierarchical architecture for parallel multi-scale spatio-temporal processing to balance between network capacity and constraints imposed by these loss functions.

Our network architecture is a solution for the blind spot problem similar to the multi-scale convolutional network [20], but different in three points. Firstly, instead of conventional image precessing methods [20], our model uses residual neural networks for down- and up-sampling which can be optimized through the back-propagation. Secondly, we adopt a single adversarial loss for a finally generated frame instead of scale-wise adversarial losses. Finally, we use three-dimensional convolutions for spatio-temporal processing instead of two-dimensional ones. Another strong solution for the blind spot problem is exhaustive scanning of a video frame sequence along not only temporal but also all spatial directions by recursive application of convolutional LSTMs [2]. This solution can be optimized through the back-propagation-through-time which consumes a large amount of memory proportional to the length of sequence, hampering increasing the model complexity due to the limited amount of computational resources of commercially available devices.

3 Proposed Method

Our goal is to learn a video prediction model which can predict future frames that are sharp and perceptually realistic given a sequence of past frames from natural videos, even when there are large scene changes between frames.

3.1 Objective function

Our model adopts a framework of a generative adversarial network (GAN) that consists of a generator network G, and discriminator network D. Here, G is a video prediction model that learns to map a given input video sequence of t frames $\mathbf{x_{0:t}} = \{x_0, \dots x_{t-1}\}$ to the predicted future frames $\hat{\mathbf{x}}_{\mathbf{t:T}}$, while D is trained to distinguish the predicted frames $\hat{\mathbf{x}}_{\mathbf{t:T}}$ from real videos $\mathbf{x_{t:T}}$. The GAN optimization with the hinge version of the standard adversarial loss functions [21] involves the following objectives:

$$G^* = \arg\min_G \lambda_1 \mathcal{L}^G_{adv} + \lambda_2 \mathcal{L}_{MAE} + \lambda_3 \mathcal{L}_{VGG} \tag{1}$$

$$D^* = \arg\min_D \mathcal{L}^D_{adv} \tag{2}$$

The objective function for the generator is a weighted combination of an adversarial loss function and two types of reconstruction loss functions. The hinge-version adversarial loss functions are defined as:

$$\mathcal{L}^G_{adv} = -\mathbb{E}_{\mathbf{x_{0:t}}, \mathbf{z}} \left[D(G(\mathbf{x_{0:t}}, \mathbf{z}), \mathbf{x_{0:t}}) \right] \tag{3}$$

$$\begin{aligned} \mathcal{L}^D_{adv} = {} & \mathbb{E}_{\mathbf{x_{0:t}}, \mathbf{x_{t:T}}} \left[\max(0, 1 - D(\mathbf{x_{t:T}}, \mathbf{x_{0:t}})) \right] \\ & + \mathbb{E}_{\mathbf{x_{0:t}}, \mathbf{z}} \left[\max\left(0, 1 + D\left(G(\mathbf{x_{0:t}}, \mathbf{z}), \mathbf{x_{0:t}}\right)\right) \right] \end{aligned} \tag{4}$$

where $\mathbf{z}$ is a noise vector.

The reconstruction loss functions include the pixel-wise, mean absolute error loss (**MAE loss**) function defined as follows:

$$\mathcal{L}_{MAE} = \|\hat{\mathbf{x}}_{\mathbf{t:T}} - \mathbf{x_{t:T}}\|_1^1 \tag{5}$$

Along with the MSE, this is one of the most widely used loss functions for future video prediction on which the state-of-the-art methods rely [2,19]. Although solutions of MAE/MSE optimization problems achieved particularly high scores in the traditional metrics such as the peak signal-to-noise ratio (PSNR) and Structural Similarity Index Measure (SSIM), they often result in blurry images which lack high-frequency details and thus *perceptually* distant from the target frames for human observers.

Recent works show that instead of a pixel-wise loss function, a perceptual loss function based on differences between learned visual feature representations of pre-trained deep CNNs is very effective for generating high-quality images with fine details [6,1,11,16]. We therefore adopt the second reconstruction function, a perceptual loss function based on cosine distances between visual feature

representations of pre-trained CNNs which is found to correlate well with human perception [35]. Let ϕ^l be the visual feature representations extracted from the last convolutional layer of the l-th block of VGG-16 [27]. We define the perceptual loss function, **VGG loss**, as follows:

$$\mathcal{L}_{VGG} = \sum_{l} \frac{1}{H_l W_l} \sum_{h,w} \left\| \frac{\phi^l_{h,w}(\hat{x})}{\|\phi^l_{h,w}(\hat{x})\|^1_2} - \frac{\phi^l_{h,w}(x)}{\|\phi^l_{h,w}(x)\|^1_2} \right\|^2_2 \tag{6}$$

where H_l and W_l denote the dimensions of the l-th feature representations.

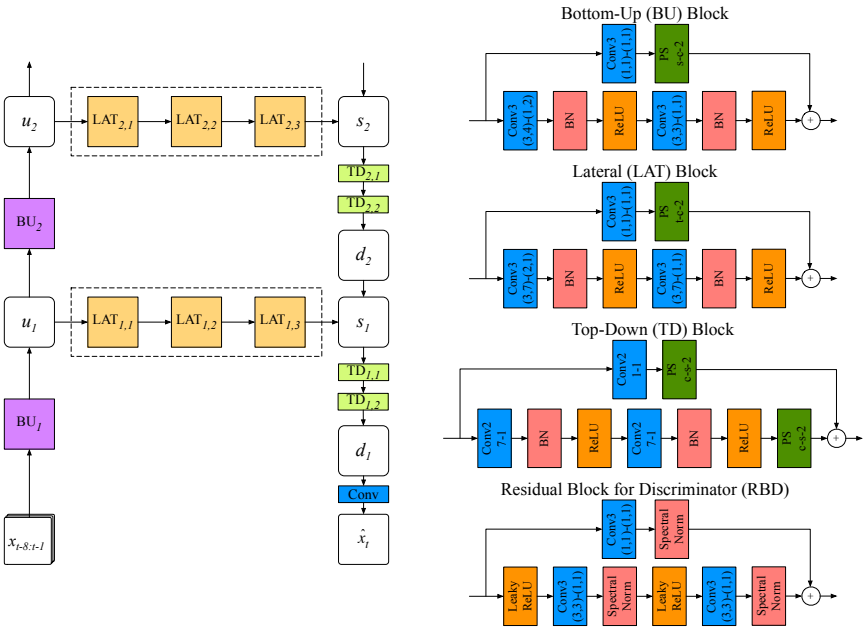

Fig. 2. Network architecture of the generator network and its building blocks. Left: Illustration of information flow within two layers of the generator. Each layer consists of the bottom-up block (BU), a series of the lateral blocks (LAT) and the top-down blocks (TD). Right: Designs of the building blocks. Conv3 (k_t, k_s)-(s_t, s_s): k_t and k_s (s_t and s_s) indicate kernel sizes (strides) of temporal and spatial dimensions, respectively. Conv2-k-s: k and s indicate kernel size and stride of a spatial convolution. BN, and Spectral Norm indicate the batch normalization [9] and spectral normalization [21], respectively. PS $d0$-$d1$-r indicates the pixel shuffler [26] that upscales an input array along its $d1$ dimension from interspersed $d0$ layer by subpixel array sampling with the upscaling factor r. Consider that the shape of an input array is (c, t, h, w) where c, t, h, and w indicate channel, time, height and width dimensions. Similar to $d0$ and $d1$, we use c, t, or s to indicate channel, time, and space (h and w) dimensions of an input array, respectively

3.2 Network architecture

The generator network consists of a series of repeating stacked modules which have spatial receptive fields of increasing size and communicate via bottom-up and top-down connections. Briefly, each module of the network consists of three basic parts: a bottom-up block (BU), a sequence of lateral blocks (LAT) and top-down blocks (TD). These network blocks are sequences of a 3-dimensional/2-dimensional convolutional layer, batch normalization [9], ReLU activation function and/or pixel shuffler [26] with a skip connection. The bottom-up block of the l-th module, BU_l, receives a bottom-up input u_{l-1} from the lower module and computes u_l, a spatially-downscaled representation of the input, and then passes it to the higher module.

$$u_l = \mathrm{BU}_l(u_{l-1}) \tag{7}$$

The bottom-up input to the lowest module, u_0, is a sequence of t video frames $\mathbf{x_{0:t}} \in \mathbb{R}^{3\times t\times h\times w}$. We use a fixed value 8 for t in this paper. u_l is also passed forward through a series of lateral blocks ($\mathrm{LAT}_{l,1}$, $\mathrm{LAT}_{l,2}$, $\mathrm{LAT}_{l,3}$) to predict s_l, a next-frame state at the spatial scale of the module, along with top-down input, d_{l+1}, from the higher module.

$$s_l = [\mathrm{LAT}_{l,3}(\mathrm{LAT}_{l,2}(\mathrm{LAT}_{l,1}(u_l))), d_{l+1}] \tag{8}$$

Then s_l is spatially up-sampled through the top-down blocks ($\mathrm{TD}_{l,1}, \mathrm{TD}_{l,2}$) to generate a top-down output to the lower module.

$$d_l = \mathrm{TD}_{l,2}(\mathrm{TD}_{l,1}(s_l)) \tag{9}$$

The top-down output from the bottom module (l=1), d_1, is further passed through a convolutional layer and sigmoid function to predict the next-frame image. We use four modules whose number of channels are (64, 128, 256, 512) for the generator network. Random noises are injected at the top-down blocks of upper two modules via dropout before applying ReLU activation function as [10].

The discriminator network consists of a series of 5 repeating blocks of the residual blocks for the discriminator (RBD) followed by spatial average pooling. Number of RBDs and their channels are (1, 2, 2, 2, 2) and (64, 128, 512, 1024, 2048), respectively. An input to the discriminator network is a predicted next frame $\mathbf{\hat{x}_t}$ or its ground truth $\mathbf{x_t}$ concatenated to the corresponding input frame sequence to the generator $\mathbf{x_{0:t}} \in \mathbb{R}^{3\times t\times h\times w}$. The last spatial average pooling is followed by temporal average pooling, and the discriminator finally tries to classify if each $N \times M$ patch in a frame is *real* or *predicted* [10].

4 Experiments

4.1 Dataset

We evaluate the proposed approach on car-mounted camera video prediction. We use two famous datasets, the KITTI dataset [7] and the Caltech Pedestrian dataset [3], which were taken from vehicles moving around urban areas

in Germany and Los Angels, respectively. They consist of diverse and complex visual motions of a variety of real-world objects (vehicles, pedestrians, bicycles, roads, buildings, trees, etc.) at different scales due to a combination of objects' own movements and camera movements. Compared to other popular, real-wold datasets, such as Human3.6M and UCF-101 datasets, which have static background, these datasets are characterized by their broad range of background motions, from a static background when an ego-vehicle stops at a red light to large background motions when an ego-vehicle turns right or left at a crossing.

The model is trained on the KITTI dataset and tested on the Caltech Pedestrian test dataset as proposed by [19]. Every ten input frames from the training subdivision [19] of the KITTI dataset are sampled for training (about 41K frames as total). Note that our models use the last 8 frames as input from the sampled 10-frame-length sequence. Frames from both datasets are center-cropped and down-sampled to 128×160 pixels. In the case of the Caltech Pedestrian dataset whose sampling rate is 3 times higher than the KITTI dataset, we beforehand temporally down-sample all sequences of the Caltech Pedestrian dataset to match their sampling rate to that of the KITTI dataset. All pixel values are normalized to [0,1].

4.2 Training details

We trained all our networks on a NVIDIA Tesla V100 GPU with 32GB onboard memory with the KITTI dataset. Input sequences were horizontally flipped with the probability of 0.5 for data augmentation. For optimization, we used Adam [13] with $\beta_1 = 0.0$, $\beta_2 = 0.9$ and a batch size of 8. In order to avoid unstable optimization and suboptimal solutions, we trained the model end-to-end in three phases. In the first phase, the discriminator was disabled and the generator was trained to predict the next frame given 8 past frames with the MAE and VGG losses with a learning rate $\alpha = 10^{-4}$, $\lambda_2 = 1000$ and $\lambda_3 = 400$. Then in the second phase, the discriminator was trained with a learning rate $\alpha = 10^{-4}$ while the parameters of the generator was fixed. Finally in the third phase, the actual GAN was trained with 8 updates of the discriminator per one update of the generator and with $\lambda_1 = 1$, $\lambda_2 = 1000$, $\lambda_3 = 400$, learning rates 2×10^{-5} and 1×10^{-4} for the generator and discriminator, respectively. For a baseline model, the ContextVP, more specifically ContextVP4-WD-big, was trained for the next-frame prediction given an input of 10 past frames on two NVIDIA Tesla V100 GPUs with a batch size of 8 and the same optimization parameters with [2], with the additional flip-flop data augmentation as above mentioned.

4.3 Metrics

We adopt Structural Similarity Index Measure (SSIM) [34] to measure the prediction accuracy. However, it is well-known that this metric correlates poorly with human perceptual judgement of visual similarity, and tends to prefer blurriness to naturalness. Therefore, we also include the Learned Perceptual Image Patch Similarity (LPIPS) [35], a metric based on the linearly weighted cosine

distance of visual features of the pre-trained CNNs, which has been shown to correlate better with human perception than SSIM. More specifically, LPIPS is calculated by using the AlexNet [14] as the pre-trained CNN architecture with the **lin** configuration.

Table 1. Evaluation of next frame prediction on the Caltech Pedestrian dataset. All models are trained for next-frame prediction on the KITTI dataset. As an input for the prediction, all the model except for ours take 10 frames, and ours use 8 frames. Higher values of SSIM, lower values of LPIPS indicate better results. (†) This score is from [17]. (+) This score is computed by us using their trained network. (*) This score is calculated by us using our implementation of their best model.

Method	SSIM	LPIPS (×100)
Copy-Last-Frame	0.775	5.23
BeyondMSE [20]†	0.881	—
PredNet[19]+	0.906	7.83
DualMotionGAN[17]	0.899	—
ContextVP[2]*	**0.924**	5.14
DPG[5]	0.923	5.04
Ours(GAN-VGG)	0.916	3.61
Ours(G-VGG)	0.917	**3.52**
Ours(GAN-MAE)	0.923	4.09
Ours(G-MAE)	0.923	4.30

4.4 Next-Frame Prediction

We compare our approach with the baseline models, ContextVP [2] and PredNet [19], as well as the BeyondMSE [20], DualMotionGAN [17], and DPG [5]. We also include the Copy-Last-Frame that simply copies the last input frame. As shown in Table 1, our GAN-based model (GAN-VGG) outperforms the state-of-the-art on LPIPS score, while achieving slightly lower SSIM score than ContextVP [2] and DPG [5]. Samples of results of next-frame prediction from the test set are shown in Fig. 1 and Fig. 3. Our model suffers less from blurriness and distortions, and produces sharper predictions with finer details than our baselines, confirming the previous findings that LPIPS correlates better with human perception than SSIM [35].

We further investigate the relationships between the accuracy of next-frame prediction of models and the degree of motions between frames of sample sequences. We use the LPIPS scores of the Copy-Last-Frame to measure the motions of sample sequences. Fig. 4 shows that our model is better than the baseline model over a broad range of motions. Although major components of the Caltech Pedestrian dataset show relatively slight changes between frames as discussed in [5] (Table 1, Fig. 4 *bottom*), the dataset also contains challenging

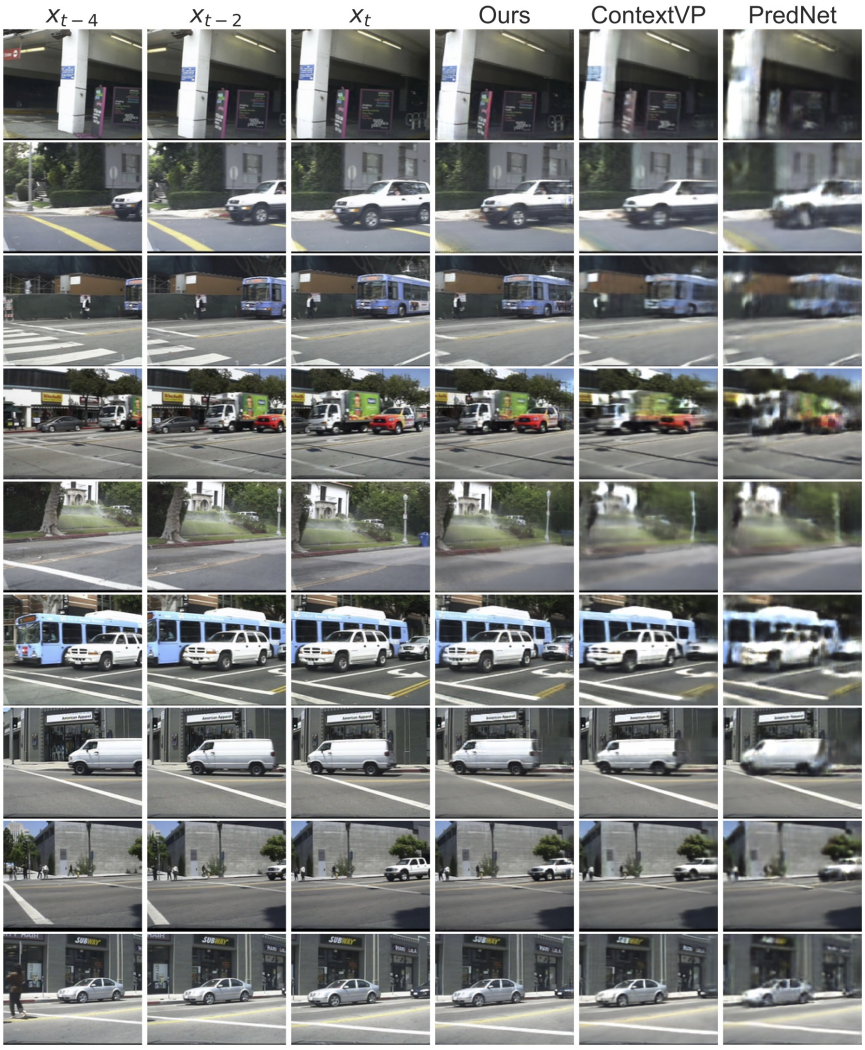

Fig. 3. Qualitative comparisons of predicted next frames from the Caltech Pedestrian dataset among our model (GAN-VGG), and the state-of-the-art models (ContextVP, PredNet). All models are trained for next-frame prediction on the KITTI dataset, and tested on the Caltech Pedestrian dataset. As an input for the prediction, ContextVP and PredNet take 10 frames, and ours use 8 frames. Note that x_t indicates the Ground Truth. Our model generates much sharper results with finer details and textures, and less distorted appearance of object compared to other models. Results are best viewed in color with zoom

sample sequences with large motions between frames. You can find such challenging samples in Fig. 1 and Fig. 3, whose motions range between 0.16 and 0.4. It is the range of motions between frames in which our model outperforms the baseline by increasing margins.

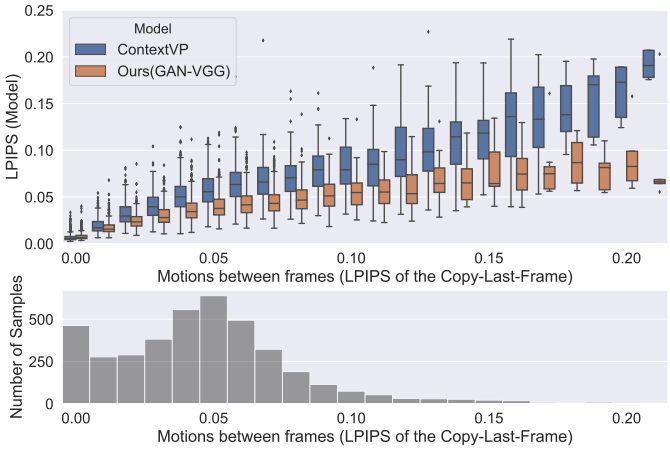

Fig. 4. Breakdown of model performance in next-frame prediction on the Caltech Pedestrian dataset based on motions between frames of sample sequences (top), and distribution of motions in the Caltech Pedestrian dataset (bottom). We adopt the LPIPS scores of the Copy-Last-Frame to measure motions in sample sequences. Note that the bottom-axis is truncated at 0.22 because we observed at most a single sample per bin above this value. Results are best viewed in color

Moreover, we investigate the effect of different choices of loss functions for the GAN-based networks. Specifically, we investigate the GAN-based networks with the standard MAE loss but not with VGG loss (GAN-MAE). We also evaluate the performance of the generator network optimized for the two losses without the adversarial loss function $\mathcal{L}^G_{adv}$, $\mathcal{L}^G_{MAE} = \lambda_2 \mathcal{L}_{MAE}$ (G-MAE) and $\mathcal{L}^G_{VGG} = \lambda_2 \mathcal{L}_{MAE} + \lambda_3 \mathcal{L}_{VGG}$ (G-VGG). Quantitative results are summarized in Table 1, and visual examples are provided in Fig. 5. The MAE loss by itself provides solutions with the highest SSIM scores which are comparable to, and with better LPIPS scores than those of the state-of-the-art models, indicating the advantage of our design of network architecture over baseline models. However, results of the solutions are perceptually rather blurry and smooth than results with the VGG loss, even combined with the adversarial loss which further improves the LPIPS score. Our model variants which include both VGG and MAE loss functions (GAN-VGG, G-VGG) outperform other model variants without VGG loss (GAN-MAE, G-MAE) and the baseline models on the LPIPS score. G-VGG achieves the new state-of-the-art on the Caltech Pedestrian dataset in terms of LPIPS, slightly outperforming GAN-VGG. Visual results of G-VGG

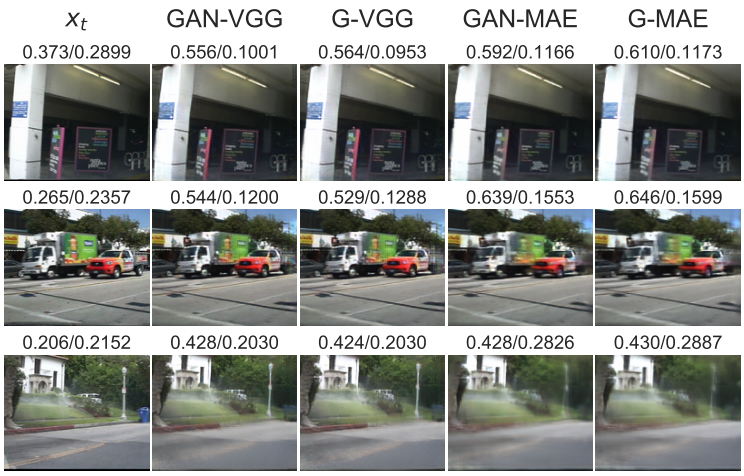

Fig. 5. Qualitative comparisons of predicted next frames of our model variants with different choices of loss functions. We report SSIM/LPIPS for the example image. Note that SSIM/LPIPS of x_t (ground truth) is of the Copy-Last-Frame. Results are best viewed in color with zoom

are sharper and have finer details and textures than results of the model variants without VGG loss, however, are perceptually indistinguishable from results of GAN-VGG. These results indicate that the VGG loss significantly contributes for improving next-frame predictions while the adversarial loss is helpful when without the VGG loss.

4.5 Multi-Step Prediction

Fig. 6 compares multi-step prediction results of our models with the baseline models, ContextVP and PredNet. All models are trained for next-frame prediction on the KITTI dataset, and tested to recursively predict 10 future frames on the Caltech Pedestrian dataset. As an input for the prediction, baseline models and our models take 10 and 8 frames, respectively. Like the results of next-frame prediction, our models with the combination of the VGG and MAE loss functions (GAN-VGG, G-VGG) are better than the baseline models and our model variants without the VGG loss function (GAN-MAE, G-MAE). In contrast to the next-frame prediction, GAN-VGG is better than G-VGG when predicting further future frames. The distributions of LPIPS scores at the 9-th future step are bimodal (Fig. 6, right). These two peaks correspond to samples of very small motions (< 0.01) and larger motions, respectively. Our models are better than the baselines with samples of larger motions, but not with samples of very small or no motions as in the case of next-frame prediction (leftmost bin in Fig. 4). Quantitative comparisons with samples of average and large motions are shown in Fig. 7 and Fig. 1, respectively. Although appearances of objects in the

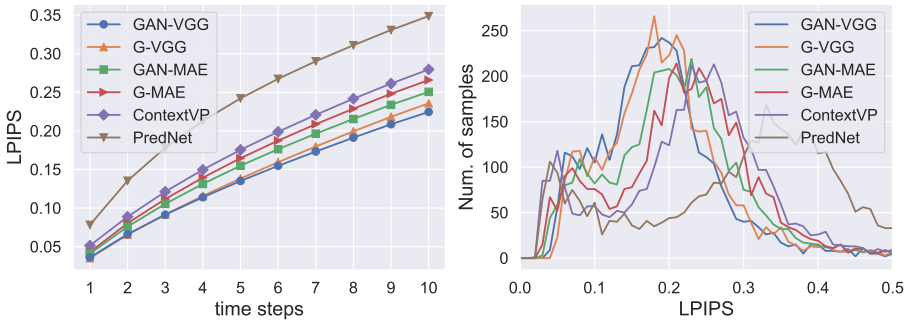

Fig. 6. Comparison of multi-step prediction on the Caltech Pedestrian dataset (left) and distributions of LPIPS scores at the 9-th future step (right): our models (GAN-VGG, G-VGG, GAN-MAE, G-MAE), and baseline models (ContextVP, PredNet). Baseline models and our models are trained for next-frame prediction on the KITTI dataset, and tested to recursively predict 10 future frames on the Caltech Pedestrian dataset. Results are best viewed in color

predicted frames are getting blurred and distorted with recursive predictions, GAN-VGG suffers less from these artifacts.

5 Conclusions

We have described deep hierarchical residual network models for photo-realistic video prediction (G-VGG and GAN-VGG) that set new state of the arts of next-frame and multi-step predictions, respectively, on the public dataset of car-mounted camera videos when evaluated with LPIPS measure. They generate sharp future frames with finer details and textures that are more perceptually realistic than the baseline models, especially under fast and large camera movements. Ablation studies with various settings of loss functions indicate that the perceptual loss function is very effective for generating photo-realistic predictions and that the adversarial loss function further improves multi-step predictions.

References

1. Bruna, J., Sprechmann, P., LeCun, Y.: Super-resolution with deep convolutional sufficient statistics. In: 4th International Conference on Learning Representations (ICLR) (2016)
2. Byeon, W., Wang, Q., Kumar Srivastava, R., Koumoutsakos, P.: Contextvp: Fully context-aware video prediction. In: The European Conference on Computer Vision (ECCV) (2018)
3. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 304–311. IEEE (2009)

Fig. 7. Qualitative comparison of predicted frames between the state-of-the-art model, ContextVP [2], and our models (GAN-VGG, GAN-MAE). ContextVP and our models are trained for next-frame prediction on the KITTI dataset, and tested to recursively predict 10 future frames on the Caltech Pedestrian dataset. As an input for the prediction, ContextVP and our models take 10 and 8 frames, respectively. We report SSIM/LPIPS for the example image. Results are best viewed in color with zoom

4. Finn, C., Goodfellow, I., Levine, S.: Unsupervised learning for physical interaction through video prediction. In: Advances in neural information processing systems. pp. 64–72 (2016)
5. Gao, H., Xu, H., Cai, Q.Z., Wang, R., Yu, F., Darrell, T.: Disentangling propagation and generation for video prediction. In: Proceedings of the IEEE International Conference on Computer Vision (2019)
6. Gatys, L., Ecker, A.S., Bethge, M.: Texture synthesis using convolutional neural networks. In: Advances in neural information processing systems (NIPS) (2015)
7. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
9. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
10. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
11. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision (2016)
12. Kalchbrenner, N., van den Oord, A., Simonyan, K., Danihelka, I., Vinyals, O., Graves, A., Kavukcuoglu, K.: Video pixel networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1771–1779. JMLR. org (2017)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, (ICLR) (2015)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
15. Kwon, Y.H., Park, M.G.: Predicting future frames using retrospective cycle gan. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1811–1820 (2019)
16. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
17. Liang, X., Lee, L., Dai, W., Xing, E.P.: Dual motion gan for future-flow embedded video prediction. In: The IEEE International Conference on Computer Vision (ICCV) (2017)
18. Liu, Z., Yeh, R.A., Tang, X., Liu, Y., Agarwala, A.: Video frame synthesis using deep voxel flow. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4463–4471 (2017)
19. Lotter, W., Kreiman, G., Cox, D.: Deep predictive coding networks for video prediction and unsupervised learning. In: 5th International Conference on Learning Representations (ICLR) (2017)
20. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. In: 4th International Conference on Learning Representations (ICLR) (2016)
21. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: International Conference on Learning Representations (2018)

22. Oh, J., Guo, X., Lee, H., Lewis, R.L., Singh, S.: Action-conditional video prediction using deep networks in atari games. In: Advances in neural information processing systems. pp. 2863–2871 (2015)
23. Oord, A.v.d., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. arXiv preprint arXiv:1601.06759 (2016)
24. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
25. Ranzato, M., Szlam, A., Bruna, J., Mathieu, M., Collobert, R., Chopra, S.: Video (language) modeling: a baseline for generative models of natural videos. arXiv preprint arXiv:1412.6604 (2014)
26. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, (ICLR) (2015)
28. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using lstms. In: International conference on machine learning. pp. 843–852 (2015)
29. Theis, L., Hosseini, R., Bethge, M.: Mixtures of conditional gaussian scale mixtures applied to multiscale image representations. PloS one **7**(7) (2012)
30. Villegas, R., Yang, J., Zou, Y., Sohn, S., Lin, X., Lee, H.: Learning to generate long-term future via hierarchical prediction. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 3560–3569. JMLR. org (2017)
31. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: Advances in neural information processing systems. pp. 613–621 (2016)
32. Vondrick, C., Torralba, A.: Generating the future with adversarial transformers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1020–1028 (2017)
33. Walker, J., Doersch, C., Gupta, A., Hebert, M.: An uncertain future: Forecasting from static images using variational autoencoders. In: European Conference on Computer Vision. pp. 835–851. Springer (2016)
34. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
35. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)