Pseudo Ground Truth Segmentation Mask to Improve Video Prediction Quality

Mu Chien Hsu
Department of Computer Science and Engineering,
Waseda University
Tokyo, Japan
michael.hsu@asagi.waseda.jp

Jui Chun Shyur
Compal Electronics
Taipei, Taiwan
Richard.hsu@compal.com

Hiroshi Watanabe
Department of Communications and Computer Engineering
Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

Abstract—**Video prediction to foresee future events is an extremely difficult job since it involves spatial feature extraction and temporal sequential analysis. We identified that the semantic information is actually crucial to prediction, and proposed using "pseudo ground truth" segmentation masks which are generated automatically in real time and add them to the input layers as extra information to predict future frames. Experiments conducted on our self-defined network demonstrated drastically higher quality predictions are achieved when compared with other state-of-the-art direct video prediction models**.

Keywords—video prediction, segmentation

I. Introduction

Video prediction tries to predict future events from given sequential past frames, while it is notably difficult since it requires both spatial analysis on extracting meaningful semantic information, such as pedestrians on a crossing road; and temporal analysis on that semantic information, such as if pedestrians will cross the road. In this paper we show the prediction quality can be greatly improved by adding an extra mask of "pseudo ground truth segmentation", in short, the "*mask*" as an additional input channel. The *masks* are *pseudo* in that they are automatically generated in *real time* by YOLACT [1] instead of human labeled ground truth. The extra *masks* can be predicted as well, which extends the possibility of other video prediction applications in real time for consumer electronics.

Two models are used for comparison purpose. Our direct video prediction model [2] is used as the baseline model. On top of that baseline model, we form the augmented model by adding the mask channels. Results are compared using PSNR, SSIM, and VGG cosine similarities. We also include qualitative result to mitigate defects from those evaluation metrics.

Experimental results show that i) Prediction quality is greatly improved, ii) Predicted masks are more accurate than using YOLACT on the predicted frames. In the rest of this paper, we describe our method in section 2, experiment and results in section 3, and conclusion in section 4.

II. Proposed Method

The baseline model combines U-net [3] for special analysis and convolutional LSTM [4] as residual connection in each scale for temporal analysis, while the augmented model has extra input and output mask prediction channels. Details of the model will be introduced in [2].

With the extra mask predictions, we designed a new activation function (biased relu) to replace the previous sigmoid to show its superiority; we also defined a new loss function to include the mask prediction loss.

Mask Output Activation

Two different activation functions are used on the output layer: sigmoid and our self-defined biased relu function as follows:

$$y = f(x) = \begin{cases} 0 \ , & x < -4 \\ \frac{x+4}{8} \ , & -4 \leq x \leq 4 \\ 1 \ , & x > 4 \end{cases} \tag{1}$$

Loss Functions

We combine L1 smooth loss and Binary Cross Entropy loss for predicted frames: y and predicted segmentation mask: n, respectively, as our final loss. We show our loss function for frame step t as follow:

$$\text{loss} = \text{L1}(y_t , x_t) + BCE(n_t, m_t) \tag{2}$$

Where *x* and *m* denote ground truth for predicted frames and pseudo ground truth segmentation mask respectively.

III. Experiment

A. Dataset

We use the KTH dataset [5] for training, and we resize the videos from 160×120 to 128×128. To avoid significant frame overlap, each one frame out of two frames is extracted. For pseudo ground truth segmentation mask, we use the real time segmentation model: YOLACT. Dataset are split as: 80% for training, 20% for validation and testing. All results shown use test sets.

B. Evaluation metrics

For quantitative evaluation, we use traditional evaluation metrics, PSNR and SSIM to evaluate the resulting predicted frames. However, recent researches [6, 7, 8] have pointed out that per pixel loss functions do not reflect human's perceptual judgement as precise as neuron-network based evaluation metrics, suggesting that we include VGG cosine similarity index as well. In addition to quantitative evaluation, qualitative results are also shown (Figure 2.)

We compare our results obtained from baseline model to other state-of-the-art direct video prediction models using different techniques such as generative adversarial networks (GANs) [9], variational auto encoder (VAE) [10], and network uses both GANs and VAE [6].

C. Result

During training, the input and predict are both a sequence of 5 frames. While testing, the input sequence is still of 5 frames, but predict sequence is extended to 15 frames.

Figure 1 shows quantitative results for predicted frames between (1) training without any segmentation mask and (2) training using pseudo ground truth segmentation mask. In figure 1, we can see that our baseline model performs much

978-1-7281-9802-6/20/$31.00 ©2020 IEEE

better in all three evaluation metrics while associate with pseudo ground truth. Moreover, our baseline model, no matter associate with pseudo ground truth or not, outperforms other state-of-the-art direct video prediction method in VGG cosine similarity.

Figure 2 shows the qualitative results for predicted frames by our baseline model for both use pseudo ground truth as extra information and direct video prediction.

Figure 3 shows predicted frames and their predicted *masks* produced by our baseline model. In figure 3, results done by per-frame segmentation model fails to separate shadow and human in some cases, but the segmentation mask produced by our model avoid this problem since it considers temporal features while predicting.

In figure 1, 2, and 3 "biased relu" denotes video prediction with *mask* and use biased relu as activation function; "sigmoid" denotes video prediction with *mask* and use sigmoid as activation function.

In our experiment, we discover by comparing results from models trained *with mask* for around 600,000 iterations and model trained *without mask* around 1,200,000 iterations (16 videos per iteration), that the model trained with *mask* achieves higher quality results. Which implies that by using *mask,* we can obtain higher quality results in half amount of time.

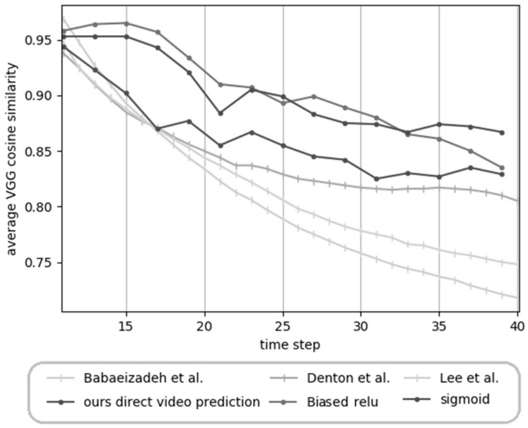

Fig. 1. Quantitative Results for predicted frames, we show the average evaluation result of our model that gives the best VGG cosine similarity evaluation index.

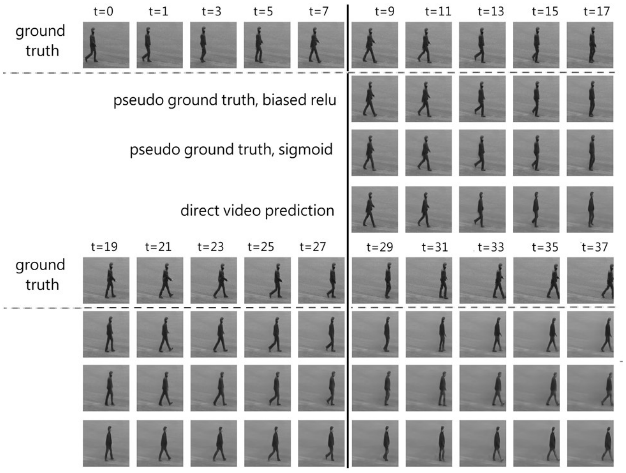

Fig. 2. Qualitative Results for Predicted Frames. From top to bottom, each row shows Ground truth, biased relu model, sigmoid model, ours direct video prediction. The first five frames (t=0 to t=7) in the ground truth row are input frames.

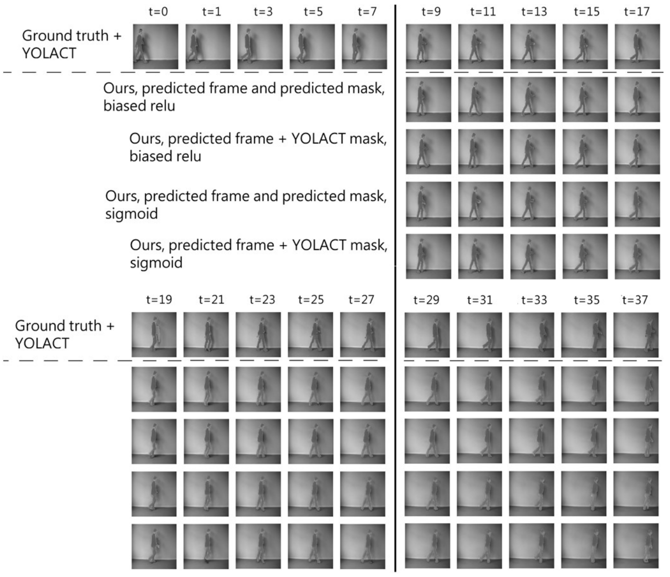
t=0
t=1
t=3
t=5
t=7
t=9
t=11
t=13
t=15
t=17
Ground truth + YOLACT
Ours, predicted frame and predicted mask, biased relu
Ours, predicted frame + YOLACT mask, biased relu
Ours, predicted frame and predicted mask, sigmoid
Ours, predicted frame + YOLACT mask, sigmoid
t=19
t=21
t=23
t=25
t=27
t=29
t=31
t=33
t=35
t=37
Ground truth + YOLACT

Fig. 3. Predicted frame and segmentation mask result and comparison. First row shows pseudo ground truth labelled by YOLACT and the first five (t=0 to t=7) frames are the input, second row show predicted frames by our direct video prediction model, third row shows predicted frame and mask, forth row shows YOLACT mask on predicted frame produced by our "biased relu" model, and the fifth and sixth row show result in same protocal as second and third row but using our "sigomid" model.

IV. Conclution

In this paper, we showed the importance of introducing semantic information in the form of *masks* to video prediction problems in that the quality of video prediction is improved significantly. We also showed that *pseudo* ground truth generated by AI can achieve quite good results without human's ground truth labelling, this makes possible the video prediction process be in real time and more practical for consumer electronic products. Last but not the least, our method can also predict the masks as semantic information accurately which is useful to help foresee future events in a better way.

References

[1] Daniel Bolya, Chong Zhou, Fanyi Xiao, Yong Jae Lee, "YOLACT: Real-Time Instance Segmentation," In ICCV 2019 pp. 9156-9165.

[2] Mu Chien Hsu, Jui Chun Shyur, Hiroshi Watanabe, "Multiscale Spatial-Temporal Feature Extraction Network for Video Prediction and Segmentation" Submitted to NeuIPS.

[3] Ronneberger O., Fischer P., Brox T., "U-Net: Convolutional Networks for Biomedical Image Segmentation" In. MICCAI 2015, vol 9351.

[4] Xingjian SHI, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, Wang-chun WOO, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," In NeurIPS, 2015.

[5] Schuldt, C., Laptev, I., Caputo, B., "Recognizing human actions: a local SVM approach," Proc. ICPR'04, Cambridge, UK.

[6] Alex X. Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, Sergey Levine, "Stochastic Adversarial Video Prediction," arXiv:1804.01523, 2018.

[7] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, Oliver Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," In CVPR, 2018, pp. 586-595.

[8] Justin Johnson, Alexandre Alahi, Li Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," In ECCV, 2016.

[9] Denton, Emily L. and Rob Fergus, "Stochastic Video Generation with a Learned Prior," In ICML, 2018

[10] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy Campbell, Sergey Levine, "Stochastic Variational Video Prediction," In ICLR, 2018

978-1-7281-9802-6/20/$31.00 ©2020 IEEE