JOINTLY PREDICTING FUTURE SEQUENCE AND STEERING ANGLES FOR DYNAMIC DRIVING SCENES

Li Du, Zhicheng Zhao, Fei Su

Beijing University of Posts and Telecommunications
School of Information and
Communication Engineering
Beijing, PRC, 100876

Leiquan Wang, Chaohong An[†]

China University of Petroleum
College of computer and
Communication Engineering
Qingdao, PRC, 266580

ABSTRACT

Generative Adversarial Network (GAN) has attracted rising attention for video future sequence prediction in driving scenes. However, the images generated by GAN often miss the target for lack of any constraints for its generated target. In this paper, an encoder-decoder based multi-task video prediction network - SegVAE is proposed by simultaneously accomplishing the predictions (generations) of both future sequence and steering angles for egocentric driving videos at pixel-level. Specifically, the encoder is constructed based on Varitional Auto-Encoder (VAE) to learn the complex latent distribution of real driving scenes. The decoder is exploited with a multi-task manner to jointly predict the future sequence and steering angles of dynamic driving scenes, where an enhanced generation mechanism is also proposed. Varitional Auto-Encoder (VAE) and Long Short Term Memory Networks (LSTM) are introduced to optimize the learning of SegVAE. The experimental results on public KITTI and NVIDIA driving datasets indicate that the proposed SegVAE can effectively mimic humans prediction mechanism, and outperform standard VAE and CNN-based generative adversarial network.

Index Terms— future sequence prediction, Varitional Auto-Encoder (VAE), multi-task learning, dynamic driving scenes

1. INTRODUCTION

Computers have surpassed human in computer vision tasks, such as face verification and image classification. However, there are still major problems to envision how the current scene might change in the following time. To overcome this challenge, many researchers devote to design advanced artificial neural networks to mimic the human brain, hoping to generate videos that simulate the future scene. Computer vision based autonomous driving has become a promising research topic recently. Predicting the future state of driving agent according to its current and former states is of great need in some real applications, such as route planning and abnormal alarm, etc. The early research works focused on simple predictable motions on relative small image patches ([1], [2]) and motions in real videos [3]. Due to the difficulty of solving aperture problem [4], the former mentioned patch-wise method is not suitable in dealing with motion prediction for high resolution video. Consequently, the current video prediction task moves to complete a full frame prediction.

GAN [5] has been used in full frame prediction by generating a new frame under dynamic driving conditions, for its remarkable success on many computer vision tasks. The style of generated image looks fairly real. However, GAN often leads to the lose of actual object in the generated image. The main reason is that there is no constraints on the generated target. In conventional driving conditions, the objects are vehicles, trees, buildings et al., which results in a hard differential image style for dataset. In addition, the input of GAN is an arbitrary noise, it is difficult to use specific features to produce target objects. VAE [6] is a suitable method which can overcome the former mentioned two shortcomings by its unique reparameterization trick. VAE adds a prior distribution constraint on the encoding network to obtain the latent representation, which is then passed into the decoder to produce the target image. The mechanism of VAE encodes source information into a higher level representation, is critical to generate actual objects for future sequence prediction in complex dynamic driving scenes. However, how to preserve the encoded scene information for the decoding phase is worth thinking about. Inspired by SegNet [7], we combine the architecture of SegNet and VAE for effective full frame prediction in driving scene, which is called SegVAE.

*This work is supported by National Natural Science Foundation of China (61532018), the Fundamental Research Funds for the Central Universities (17CX02041A), and Graduate Innovation and Entrepreneurship Project of Beijing University of Posts and Telecommunications (Project No:2018-YC-A147).

[†]Inner Mongolia University, Finance Department, No.235, College Road, Hohhot, Inner Mongolia Autonomous Region, PRC, 010021.

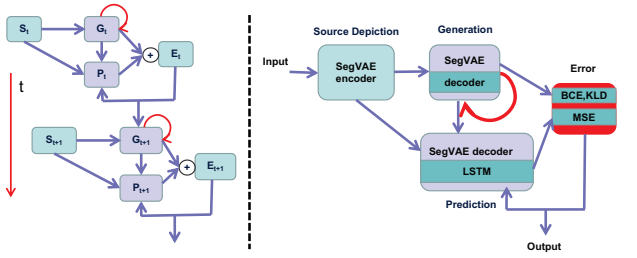

Fig. 1. **Flowchart of SegVAE with Enhanced Generation Mechanism.** Left: Information flow between two blocks within time steps. Right: Operation module of each block, including four basic units: source depiction unit S_t , generation unit G_t, prediction unit P_t and error unit E_t.

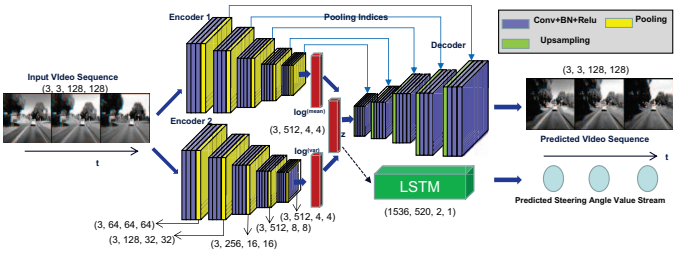

Fig. 2. **The Architecture of SegVAE Prediction Network.**

In driving prediction task, many other related tasks, such as steering angles prediction, can also be accomplished based on the learning process. Recent works show that learning correlated tasks simultaneously can boost the performance of individual tasks [8]. In this paper, we formulate the driving prediction task as an end-to-end multi-task prediction problem. We aim to generate the future driving states of the target agent by predicting both the future video sequences and their corresponding steering angle value streams of ego-centric video.

By analyzing the research experience above, we visualize the possibility of VAE and SegNet architecture for pixel-level future prediction task. The future prediction task is a multi-task learning approach, containing two tasks: future sequence prediction and steering angles prediction. The contributions of this paper main includes two folds: 1) A multi-task learning approach is proposed by jointly predicting future sequence and steering angles for dynamic driving scenes. 2) An effective driving video prediction (generation) network with enhanced generation mechanism - SegVAE is proposed by combining the architecture of VAE and SegNet.

2. METHOD

The details of the proposed end-to-end multi-task prediction progress for ego-centric driving videos based on Seg-VAE model with an enhanced generation mechanism will be described in the following part.

2.1. Problem Definition

Let $X_t = (x_{t-n+1}, x_{t-n+2}, ..., x_t)$ denotes a video sequence with n frames, $A_t = (a_{t-n+1}, a_{t-n+2}, ..., a_t)$ stands for their corresponding vehicle steering angle values. x_t, a_t are the t-th frame and its steering angle value. Therefore, the multi-task prediction problem can be defined by the following two functions: future frame generation function G: $R^{128\times128\times3} \rightarrow R^{128\times128\times3}$ that generates $x_{t+1} = G(x_t)$, and future steering angel prediction function P: $R^{128\times128\times3} \times R \rightarrow R$ that predicts $a_{t+1} = P(x_t, a_t)$.

2.2. The Overall Architecture of SegVAE

Videos captured in general driving conditions always contain more complex dynamic spatiotemporal information than that captured in monitored scenes. The two functions defined for the two tasks are hard to converge with high-dimension deep learning method. Therefore, an effective learning architecture is of great need to complete the multi-task progress. The detail of our proposed SegVAE is shown in Fig.1.

The left part shows the information transportation progress between two time steps, while the right part is the specific prediction progress of each learning block. Each block is constructed by specific deep learning function units with local recurrence at each stage. Briefly, each one consists of four units: a source depiction unit S_t, a generation unit G_t, a prediction unit P_t, and an error unit E_t. The difference between true future sequence and its predicted one is represented as error, E_t, which including two populations, sequence (BCE+KLD) and angle (MSE) values. It will then be passed forward to prediction unit as input of the next generation unit G_{t+1}. The input of generation unit G_{t+1} including the copies of error and the source depiction units of prior step.

2.3. Detail of SegVAE with Enhanced Generation Mechanism

Inspired by the motivations in [9] and [10], our SegVAE with enhanced generation mechanism constructed by repeating stacked blocks to complete a continuous prediction progress for ego-centric driving videos, which is depicted in Fig2. The two encoders and one decoder are all constructed by fully convolutional layers. Encoder 1 maps the source input video sequence to high level latent variable $log(mean)$, while encoder 2 maps the same source information to $log(var)$. Then we add a constraint to force the two generated latent vectors roughly follow an unit gaussian distribution and output a constrained variable z. The decoder upsamples z using the transferred pool indices form encoder 1 to produce a sparse feature map(s), then performs convolution with a trainable filter filter bank to densify the feature map [7] to target prediction (reconstruction) sequence. Meanwhile, a 2-layer LSTM [11] regresses z into a 1D value stream, that corresponding to the target video sequence. The source depiction

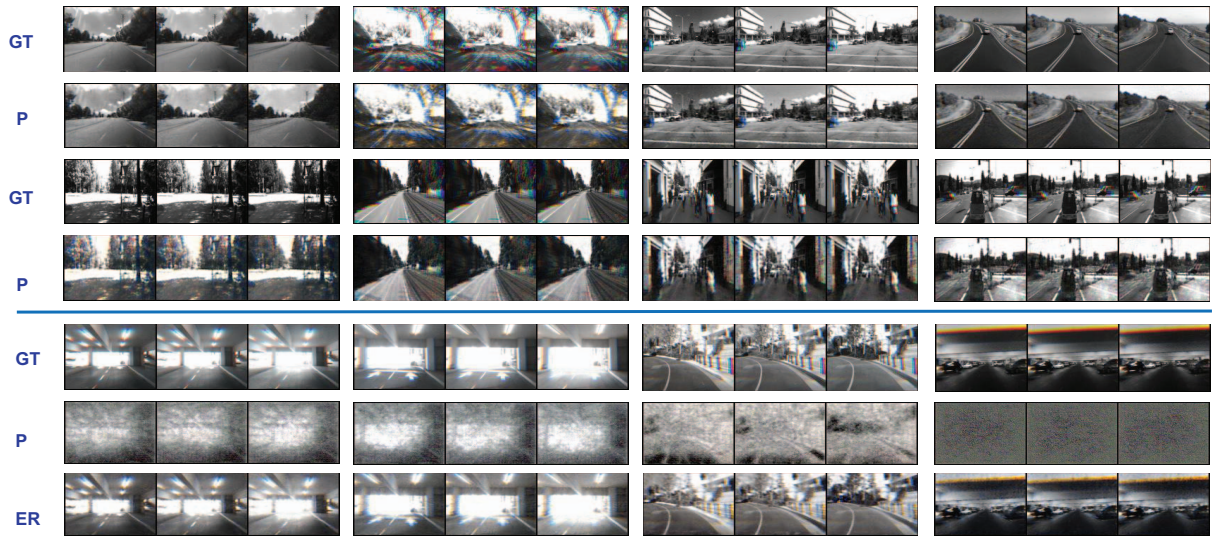

Fig. 3. **Prediction Results with resolution of** 128×128 **when** α **= 0.3.** P denotes the predicted results of our method without initial process from enhanced generation. GT are the normalized outputs from pre-processing progress in model training progress. ER shows the improved bad cases from enhanced generation of basic SegVAE. It is obvious that our model can learns to generate sequences captured under serious light changing conditions, and can produces images contain complex street characteristics. Meanwhile, the bad cases in the last three rows emphasize the aid-generation ability of our prediction mechanism in SegVAE for images captured in dark places.

unit S_t is the encoder part in our proposed SegVAE, which aims to map the source video sequences to high level latent hidden variables $mean$ and $stand\ deviations$. The generation unit G_t is the corresponding decoder in SegVAE, which works to map the low resolution encoder features to full input resolution features in a pixel-wise image generation method. Specifically, an enhanced generation mechanism is performed on G_t by transferring the max-pooling indices of encoder to the decoder.

The prediction unit P_t predicts the future frames and their corresponding steering angle values with the same decoder architecture in the generation unit. It simultaneously generates future frame information with decoder and predicts the target steering angle values through the following two-layer LSTM. The parameters of P_t in decoder are initialized by pre-trained parameters in G_t, to pass history memory information to the future prediction phase. Equations (1)-(4) show the calculation progresses of all the function units.

$$S_t = VAE(SegNet(encoder(FCN(X_t)))) \tag{1}$$

$$G_{t+1} = VAE(SegNet(decoder(FCN([S_t, E_t, P_t])))) \tag{2}$$

$$P_{t+1} = \begin{cases} VAE(SegNet(decoder(FCN[S_t, G_t]))) \\ LSTM(VAE(SegNet(encoder(FCN(X_t))))) \end{cases} \tag{3}$$

$$E_t = [Relu(\hat{X}_t - X_t); Relu(\hat{A}_t - A_t)] \tag{4}$$

We train the generation/prediction unit with a mean squared error (MSE) loss $L_{p'}$, to measures the generation accuracy between source sequence and its reconstructed/predicted one.

$$L_{p'}(X_t, X_t^{'}) = \frac{1}{N}\sum_{i=0}^{N-1}(x_i - x_i^{'})^2 \tag{5}$$

where N denotes the length of an input frame sequence, X_t is the source sequence and $X_t^{'}$ is the generated/predicted target.

As it's difficult to calculate X's distribution, we introduce two encoders to separately generate two parameters which can describe the distribution of each dimension in the latent space. We experimentally assume that the prior $P(X)$ follows a normal distribution. The decoder then generates the latent vector z by sampling from the defined distribution and constructs a reconstruction of the original input X. In the sampling progress, we leverage the "reparameterization trick" in [12], which contributes to our parameter optimization progress. At the same time, we combine the KL divergence loss with the prior distribution $P(X)$, to ensure the latent distribution z holding the same distribution as them. Meanwhile, the generation/prediction loss $L_{p'}$ shows the distribution correlation between z and X. The newly combined prediction formulation is attributed to L_p.

$$L_p = L_{p'}(X, X^{'}) + \alpha\sum_j KL(q_j(z|x)||N(log(mean), log(var))) \tag{6}$$

where $log(mean)$ and $log(var)$ are the logarithm values of mean and standard deviation, which are output by the two

encoders respectively. KL is KL-divergence, q is the variational distribution and z are samples following the variational distribution. In this work, we assume z and $X^{'}$ are both independent, α is fixed to -0.5. We empirically define the latent $q(z)$ produced by encoder following a prior Gaussian distribution. $X^{'}$ is produced by decoder.

Meanwhile, the steering angle prediction task takes the same measure method as frame generation and be attributed to a 1D regression progress, which is denoted as a L_s loss optimization problem.

$$L_s(a, a^{'}) = \frac{1}{N}\sum_{i=0}^{N-1}(a_i - a_i^{'})^2 \tag{7}$$

Therefore, the final prediction process attributed to minimizing the multi-task loss L_{final}, which is the sum of L_p and L_s.

$$L_{final} = L_p + \lambda * L_s \tag{8}$$

The objective calculation progress is supervised by Stochastic Gradient Descent (SGD) [13] algorithm with Back-propagation(BP) to optimize L_{final}.

3. EXPERIMENTS

We test our proposed method on public released ego-centric video sequences in KITTI [14] and NVIDIA driving ([15], [16]) datasets, which are both captured by front-mounted cameras in driving cars. For KITTI dataset, we randomly select video sequences with frame rate of 30 frames per second (FPS) from "City", "Residential" and "Road" categories, to form a 40K training set and a 5K testing set. For NVIDIA dataset, as it was continuously recorded at the same frame rate as KITTI with 45567 frames, we separately construct training and testing datasets with continue 31786 and 13782 frames. All the data frames are center-cropped, resized to 128×128 pixels and normalized with empirical mean μ and standard deviations σ values. To verify the robustness of AutoPre for spatial and temporal dynamics, we conduct a 3-frame prediction experiment when there isn't overlap between two continuous inputs. We complete our network on a GTX 1080ti GPU with pytorch deep learning architecture. Each model is trained within 50 epochs.

To achieve a more efficient feature extraction progress, we firstly conduct a sequence reconstruction experiment to recover the input frames with a basic SegVAE. A basic SegVAE is a single-stream architecture with only one encoder and one decoder. Then we initialize the encoder and decoder with the former trained parameters to complete a more efficient training progress. Finally, super parameter λ in the final objective training function is set to 0.3 when it ranges in (0, 1) with a 0.1 stride.

We compare our trained SegVAE and enhanced generation SegVAE (SegVAE-EG) with standard VAE and GAN on two testing datasets separately. MSE [17] is used to measure the prediction accuracy of the final task, and Structural Similarity Index (SSIM) [18] is introduced to measure the quality of the predicted results.

Table 1. Testing Results.

Methods	Datasets	MSE	SSIM(%)
SegVAE-ER	NVIDIA	0.009	76.7
SegVAE		0.012	76.5
VAE		0.350	69.4
GAN		0.056	65.5
SegVAE-ER	KITTI	0.032	70.8
SegVAE		0.041	69.5
VAE		0.480	59.8
GAN		0.087	71.6

Experimental results are shown in Table 1. SegVAE-EG outperforms the other baselines for the designed mechanism effectively transport the useful high-level memory through an efficient and reasonable generation method based on SegNet and VAE. The standard VAE doesn't work well due to its weak spatiotemporal feature extraction ability for continuous video sequence, which is the key to the latent variables generation progress. The generalization of our method still needs to be improved to adapt to different datasets with various data distribution. All the VAE based methods perform well on continuously captured dataset NVIDIA, which holds a narrower distribution range than KITTI. As KITTI is a newly combined dataset with large amounts of short term video sequences captured in diverse places. The traditional GAN gets the worst result when there isn't constraints on its generated targets. It is obvious that our proposed prediction mechanism can be trained to complete the target task with a good result. Figure 3 shows some typical cases of our designed method.

4. CONCLUSION

In this paper, we propose an end-to-end multi-task prediction task relating to ego-centric driving videos. An enhanced generation mechanism with SegVAE is proposed to predict both the future video sequence and its corresponding steering angle value stream from the input continuous video sequences. Experimental results on KITTI and NVIDIA datasets show that our model can simultaneously complete the mentioned multi-task with a reasonable measure output, and outperform the standard VAE and traditional GAN methods.

5. REFERENCES

[1] I. Sutskever and G. Hinton, "Temporal-kernel recurrent neural networks," *Neural Networks*, vol. 23, no. 2, pp. 239–243, 2010.

[2] V. Michalski, R. Memisevic, and K. Konda, "Modeling deep temporal dependencies with recurrent gram-

mar cells," in *Advances in neural information processing systems*, 2014, pp. 1925–1933.

[3] T. Wiegand and B. Girod, *Multi-frame motion-compensated prediction for video transmission*. Springer Science & Business Media, 2012, vol. 636.

[4] K. Nakayama and G. H. Silverman, "The aperture problem–i.perception of nonrigidity and motion direction in translating sinusoidal lines," *Vision research*, vol. 28, no. 6, pp. 739–746, 1988.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[6] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[7] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[8] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[9] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," *arXiv preprint arXiv:1605.08104*, 2016.

[10] R. P. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nature neuroscience*, vol. 2, no. 1, pp. 79–87, 1999.

[11] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," 1999.

[12] T. Xue, J. Wu, K. Bouman, and B. Freeman, "Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 91–99.

[13] L. Bottou, "Stochastic gradient descent tricks," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 421–436.

[14] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[15] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[16] M. Bojarski, P. Yeres, A. Choromanska, K. Choromanski, B. Firner, L. Jackel, and U. Muller, "Explaining how a deep neural network trained with end-to-end learning steers a car," *arXiv preprint arXiv:1704.07911*, 2017.

[17] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE signal processing magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.