

Order Matters: Shuffling Sequence Generation for Video Prediction

Junyan Wang

j.wang81@newcastle.ac.uk

Bingzhang Hu

bingzhang.hu@newcastle.ac.uk

Yang Long

yang.long@newcastle.ac.uk

Yu Guan

yu.guan@newcastle.ac.uk

Open Lab, School of Computing

Newcastle University

Newcastle Upon Tyne, UK

Abstract

Predicting future frames in natural video sequences is a new challenge that is receiving increasing attention in the computer vision community. However, existing models suffer from severe loss of temporal information when the predicted sequence is long. Compared to previous methods focusing on generating more realistic contents, this paper extensively studies the importance of sequential order information for video generation. A novel *Shuffling sEquence gEneration network* (SEE-Net) is proposed that can learn to discriminate between natural and unnatural sequential orders by shuffling the video frames and comparing them to the real video sequences. Systematic experiments on three datasets with both synthetic and real-world videos manifest the effectiveness of shuffling sequence generation for video prediction in our proposed model and demonstrate state-of-the-art performance by both qualitative and quantitative evaluations. The source code is available at <https://github.com/andrewjywang/SEENet>.

1 Introduction

Unsupervised representation learning is one of the most important problem in the computer vision community. Compared to image, video contains more complex spatio-temporal relationship of visual contents and has much wider applications [10, 13, 15]. In order to explicitly investigate the learnt representation, video prediction has become an emerging field that can reflect whether temporal information is extracted effectively. There are recent variations of related work on human activity prediction and recognition [8, 11, 16, 17, 19], motion trajectory forecasting [1, 21], future frame prediction [24, 26] and so on. Also, the application has appeared in robotics [1] and healthcare [30] areas.

In order to predict long-term future frames, the key challenge is to extract the sequential order information from still contexts. Most of state-of-the-art methods [24, 26] exploited advanced generative neural network to directly predict frames based on reconstruction loss that is more sensitive to contextual information. Some recent works [1, 36] attempted to extract spatio-temporal information using motion-sensitive descriptors, *e.g.* optical flow, and



Figure 1: Human can figure out the correct order of shuffled video frames (2-1-3). By doing so, attention is enforced to be paid on the temporal information.

achieved improved results. A common issue in existing works is the severe loss of temporal information, for example, the target becomes blurry and gradually disappears during processive video prediction.

Our work focuses on future frame prediction and this paper is inspired by a fact that the ordinal information among the frames is more important for the humans' perceptions of a video. And such kind of information can be better captured by performing a sorting task. For example, as shown in Fig. 1, to sort the frames, one has to pay his attention to the temporal information. Motivated by the fact above, we propose a *Shuffling sEquence gEneration network* (SEE-Net) that can explicitly enforce the temporal information extraction by shuffling the sequential order of training videos, where a *Shuffle Discriminator* (*SD*) is designed to distinguish the video sequential with natural and shuffled order. As the content information is supposed to be the same between real and shuffled frames, the model is therefore forced to extract the temporal order information explicitly. Extracting temporal information is a very challenging task from raw video frames and optical flow is widely used in temporal information extraction tasks [31, 32, 33], therefore we apply the optical flow network PWCNet [34] to generate optical flow images between adjacent frames. In addition, we evaluate our method on both synthetic dataset (Moving MNIST) and real-world datasets (KTH Actions and MSR Actions). The contributions of this paper are summarized below:

- We propose the SEE-Net for the task of long-term future frame prediction, which can extract both content and motion information by two independent auto-encoder pathways.
- We introduce the shuffle discriminator to explicitly control the extraction of sequential information.
- Extensive results manifest that our model is not only more stable in long-term video frame predictions, but also infers more accurate contents at each frame compared to other methods.

2 Related Work

Content based Approaches The task of video frame prediction has received growing attention in the computer vision community. Early work investigates object motion prediction [35]. Advanced neural network approaches were then applied to directly predict future frames [24, 26, 37, 38]. Mathieu *et al.* [26] proposed a multi-scale auto-encoder network with both gradient difference loss and adversarial loss. PredNet [24] is inspired by the concept of predictive coding from the neuroscience literature. Each layer in the PredNet model produced local predictions and only forward deviations from these predictions to the subsequent network layers. Vondrick *et al.* [37, 38] conducted a deep regression network to

predict future frame representations. Unlike future frame prediction, Babaeizadeh *et al.* [1] and Lee *et al.* [2] addressed the video prediction by stochastic approaches that can predict a different possible future for each sample of its latent variables. A shared drawback of these methods is lack of explicit control of temporal information extraction, and therefore our work disentangles the motion information from video frames to better learn temporal information.

Content-motion Disentanglement Many recent works use content-motion disentanglement in many ways [3, 4, 5, 6, 7, 8]. For example, Shi *et al.* [9] proposed a convolutional network which offered a method to obtain time-series information between images and Srivastava *et al.* [10] demonstrated that Long Short-Term Memory was able to capture pixel-level dynamics by conducting a sequence-to-sequence model. Direct prediction usually produces blurred images because the reconstruction-based loss function encourages an averaged output. Instead of directly generating pixels, disentangling video into motion and content representation has been widely adopted in recent work. Early disentangled representation [11] is originally applied to scene understanding [12] and physics modelling [13]. MCNet [14] disentangles motion from content using image differences and applies the last frame as a content exemplar for video prediction. DrNet [15] disentangles the representation into content and pose information that is penalised by a discrimination loss with encoding semantic information. Other than disentangled approaches, optical flow is the most common approach to extract motion of objects explicitly. For instance, Simonyan *et al.* [16] proposed a two-stream convolutional network for action recognition in videos. Recent works [17, 18, 19] focus on using convolutional neural network to generate optical flow images and then extended to future frames generation [20, 21, 22]. However, content-motion disentanglement is not sufficient to learn distinct motion information. Therefore, our work also learns ordinal information among the frames.

Shuffle based Self-supervised Learning Several works utilise shuffle based self-supervised learning methods on videos, which do not require external annotations [23, 24, 25]. In [23], based on ordinal supervision provided by visual tracking, Wang and Gupta designed a Siamese-triplet network with a ranking loss function to learn the visual representations. Misra *et al.* [26] proposed a self-supervised approach using the convolutional neural network (CNN) for a sequential verification task, where the correct and incorrect order frames are formed into positive and negative samples respectively to train their model. Lee *et al.* [27] presented a self-supervised representation learning approach using temporal shuffled video frames without semantic labels and trained a convolutional neural network to sort the shuffled sequences and output the corrects. In this work, we apply the shuffle based self-supervised learning method on optical flow images to extract the ordinal information from the motion of objects, surfaces, and edges.

3 Problem Statement

To formulate the future frame prediction problem we declare some notations in advance. We use bold lowercase letters \mathbf{x} and \mathbf{m} to denote one video frame and one optical flow frame, respectively. In particular, the bold lowercase letter \mathbf{h} is denoted as the extracted latent feature. We employ the capital letter E to represent the encoding network and the capital letter G to represent the generating network. Besides, we use the curlicue letter \mathcal{L} to represent the loss function.

Formally, given a sequence of t frames from video $\mathbb{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t\}$, we aim to build

a model for predicting the next k frames $\{\hat{\mathbf{x}}_{t+1}, \hat{\mathbf{x}}_{t+2}, \dots, \hat{\mathbf{x}}_{t+k}\}$ by learning the video frame representations. The network disentangles the information from input \mathbf{x} into time-varying (motion) information \mathbf{h}^m and time-invariant (content) information \mathbf{h}^c by the motion encoder E_m and content encoder E_c respectively due to the spatial-temporal disentanglement strategy. However, many existing methods only consider the temporal information at the representation level and they [10, 11] can not achieve reasonable results on certain scenarios. Inspired by the shuffle based self-supervised learning, the proposed SEE-Net shuffles sequence generation orders and the shuffle discriminator SD converges only if effective temporal information is extracted.

4 SEE-Net

In this section we introduce the proposed **SEE-Net** and Fig.2 illustrates the overall pipeline of the proposed SEE-Net. There are mainly three components in SEE-Net, which includes a motion encoder E_m with decoder G_m , a content encoder E_c with decoder G_c as well as a frame generator G . In the following, we detail each of them in turn. To explicitly extract the motion information, instead of feeding the original frames directly into E_m , we input the optical flow images $\mathbb{M} = \{\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_{t-1}\}$ which are extracted between adjacent frames by the advanced optical flow network PWCNet [34]. Therefore the motion information can be obtained as $\mathbf{h}^m = E_m(\mathbf{m})$ and the content information can be obtained as $\mathbf{h}^c = E_c(\mathbf{x})$. The E_m and E_c are trained in a fashion auto-encoder together with their corresponding decoder G_m and G_c .

Subsequently, with the motion information of the previous t frames, we employ a long short-term memory (LSTM) network f^{lstm} to predict the motion information of the next k frames. For short term video prediction, we assume there is no significant background change in the next k frames, and use the content feature \mathbf{h}_t^c together with the predicted motion feature designed as follows.

$$\begin{aligned} \mathbf{h}_{t+i-1}^m &= f^{lstm}(\mathbf{h}_i^m, \mathbf{h}_{i+1}^m, \dots, \mathbf{h}_{i+k-2}^m), \\ \hat{\mathbf{x}}_{t+i} &= G([\mathbf{h}_t^c, \mathbf{h}_{t+i-1}^m]), \end{aligned} \quad (1)$$

where $i = 1, \dots, k$ and $[\cdot, \cdot]$ denotes the vector concatenate operation.

4.1 Content Consistency

To extract time-invariant content information, the contrastive loss is applied to optimise the content encoder. We use subscripts to denote the indices of different video clips, therefore the odd and even frames in i^{th} video can be denoted as \mathbf{x}_{2n-1}^i and \mathbf{x}_{2n}^i , where $n = 1, \dots, \lfloor \frac{t}{2} \rfloor$. It is intuitive that the content information, such as the background and the objects within the same video clip are supposed to be consistent while the content information in different video clips may be various. Thus we have

$$\begin{aligned} \mathcal{L}_{consistency} &= \sum_{i,j} y(\mathcal{D}_{ij})^2 + (1-y)\max(\delta - \mathcal{D}_{ij}, 0)^2, \\ \text{where } \mathcal{D}_{ij} &= \frac{1}{n} \sum_n \|\mathcal{E}_c(\mathbf{x}_{2n-1}^i) - \mathcal{E}_c(\mathbf{x}_{2n}^i)\|_2. \end{aligned} \quad (2)$$

\mathcal{D}_{ij} is used to denote the frame-wise l_2 distances between content information in the i^{th} and j^{th} video clips. We set y as 1 if $i = j$ and 0 otherwise. Therefore minimising $\mathcal{L}_{consistency}$ can

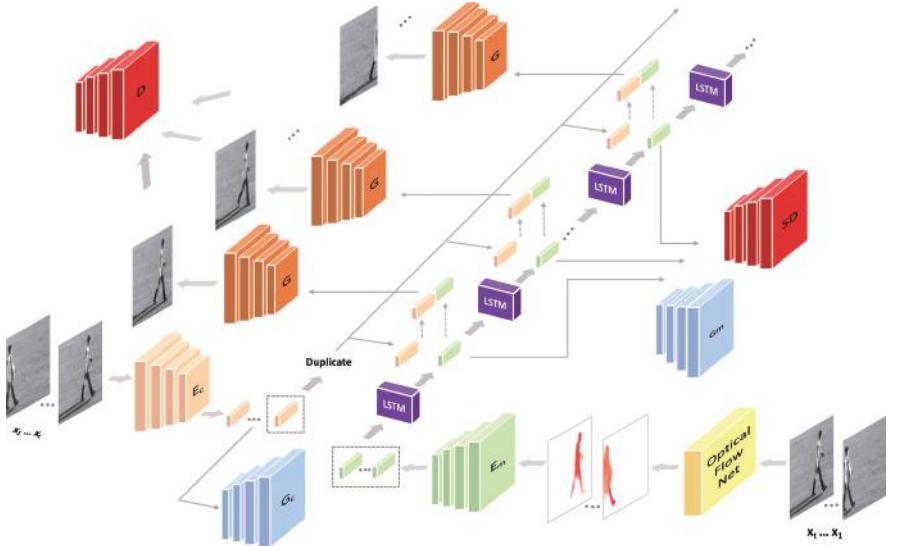


Figure 2: The proposed video prediction framework

ensure E_c to extract similar content information for all frames in the same video clip while at least a δ difference for those from different clips.

4.2 Shuffle Discriminator

To explicitly extract motion information, we propose a novel shuffle discriminator (SD), which takes a sequence of predicted motion information from f^{lstm} as input and determines if they are in the correct order. For example, we manually construct two sequences $S_{predicted}$, which is organised in the right order; and $S_{shuffled}$, in which the sequential order information is shuffled. The SD consists of a bidirectional LSTM as the first layer and a fully-connected layer.

The SD is supposed to predict 1 for $S_{predicted}$ for its correct order whilst 0 for $S_{shuffled}$. Therefore a *shuffle loss* can be defined as:

$$\mathcal{L}_{shuffle} = -\log(SD(S_{predicted})) - \log(1 - SD(S_{shuffled})). \quad (3)$$

The shuffle loss forces the predicted motion information $\mathbf{h}_t^m, \mathbf{h}_{t+1}^m, \dots, \mathbf{h}_{t+k-1}^m$ to be distinct with each other. Otherwise the SD cannot distinguish the correct ordered sequences from shuffled ones. The f^{lstm} is constrained to learn temporal change between frames thus generate reasonable motion information for upcoming frames. As the correct and shuffled sequences can be constructed without extra labelling, it can be regarded as self-supervised learning.

4.3 Adversarial Objective for Generator

Generative adversarial networks (GANs) have yielded promising results in many areas, especially in image generation. Such success is mainly benefited from the adversarial training

idea in GANs where the generator is trained competing with the discriminator. In our work, to generate realistic frames and train the generator G , we employ the adversarial loss as:

$$\min_G \max_D \mathcal{L}_{adversarial} = -\log(D(\mathbf{x}_{t+i})) - \log(1 - D(G([\mathbf{h}_t^c, \mathbf{h}_{t+i-1}^m]))) \quad (4)$$

To further enhance the quality of the generated frames, inspired by [3], we employ the l_1 loss to minimise the difference between the generated frame and ground truth.

4.4 Optimisation and Training

Combining the Eq. 2 to 4, the overall objective can be written as:

$$\mathcal{L} = \mathcal{L}_{content} + \mathcal{L}_{motion} + \mathcal{L}_{generate}, \quad (5)$$

where

$$\begin{aligned} \mathcal{L}_{content} &= \lambda_1 \mathcal{L}_{consistency} + \lambda_2 \|\mathbf{x}_j - G_c(E_c(\mathbf{x}_j))\|, \text{ for } t \geq j \geq 1, \\ \mathcal{L}_{motion} &= \lambda_3 \mathcal{L}_{shuffle} + \lambda_4 \|\mathbf{m}_{t+i-1} - G_m(E_m(\mathbf{m}_{t+i-1}))\|, \\ \mathcal{L}_{generate} &= \alpha \mathcal{L}_{adversarial} + \beta \|\mathbf{x}_{t+i} - G([\mathbf{h}_t^c, \mathbf{h}_{t+i-1}^m])\|_1. \end{aligned} \quad (6)$$

The λ_{1-4} , α and β are hyperparameters that control the occupation of each loss. As a framework with multiple loss is difficult to train, in practice, we first optimise the $\mathcal{L}_{content}$ and \mathcal{L}_{motion} respectively until model converges. Then we fixed the weights in E_m and E_c to train the generator G and discriminator D . After G and D reaching their optimal, we combine each component and fine-tune the whole network.

5 Experiments

Considering video prediction is an early stage problem with various settings, for a fair comparison, we mainly compared our approach with two representative state-of-the-art methods MCNet [36] and DrNet [4]. This paper provides a thorough evaluation of the proposed SEE-Net on both a synthetic dataset (Moving MNIST [32]) and two simple real-world datasets (KTH Actions, MSR Actions) for many existing work reports they [4, 36] cannot achieve reasonable results on certain scenarios. Both qualitative and quantitative evaluation metrics were adopted to better understand the advantages of our model.

Model Configuration Content encoder E_c , motion encoder E_m , content decoder G_c , motion decoder G_m , and the generator G consist of 4 convolutional layers and two fully connected layers. Each convolutional layer is followed by instance normalization [35] and leaky ReLU activation. Both of the feature decoders are mirrored structures of the encoder except for the final sigmoid activation functions. Our model configuration is consistent for all of the three datasets. The sizes of hidden content feature \mathbf{h}^c and motion feature \mathbf{h}^m are both 128. We employed the ADAM optimiser [35] and set $\lambda_1 = 1$, $\lambda_3 = 1$ and $\alpha = 1$ for all models. Both the LSTM network for optical flow feature prediction and the Bi-direction LSTM for shuffle discriminator consist of two layers with 64 hidden units. Besides, all optical flow images were generated by pre-trained PWCNet [34] model.

	t = 1	t = 2	t = 3	t = 4	t = 5	t = 6	t = 7	t = 8	t = 9	t = 10
Input										
	t = 11	t = 12	t = 13	t = 14	t = 15	t = 16	t = 17	t = 18	t = 19	t = 20
Ground Truth										
DrNet										
SEE-Net (Ours)										

Figure 3: Qualitative Comparison to state-of-the-art methods on the Moving MNIST dataset. Using the first ten frames as input, the goal aims to predict the following 10 frames up to the end of the video. MCNet result is missing due to all the output images are black.

5.1 Results on Synthetic dataset

One of the popular video prediction benchmarks is the Moving MNIST dataset that contains 10000 sequences, each of which has 20 frames in length showing two digits moving in the frame size of 64×64 . It has been widely used in recent video prediction works [4, 28, 33]. We follow the same experimental setting that use the first ten frames to predict the last ten in the video. For training, we set learning rate as 1×10^{-5} , $\lambda_2 = 0.01$, $\lambda_4 = 0.01$ and $\beta = 1 \times 10^{-5}$.

Our major comparison is illustrated in Fig.3. It is noticeable that the results of MCNet are missing due to completely failed outputs with no contents generated. This result is consistent with existing reports that contain only quantitative MSE evaluation results. We attribute such failure to their temporal information is simply captured by frame difference. The disentangle representation in DrNet can force to preserve both the spatial and temporal information. In contrast, our explicit control over the generated sequence order can benefit the predicted digital numbers to be separated by the estimated movement trend from the sequential orders. It can be seen that the centre of each digital number aligns well with the ground truth.

5.2 Results on Realistic Datasets

KTH action dataset [28] and MSR action dataset [27] are both used for evaluating video prediction methods. Actions in the traditional KTH dataset are simpler (walking, jogging, running, boxing, handwaving, handclapping) with more solid background therefore are suitable for predicting longer sequences, *i.e.* using first 10 frames to predict the rest 20. Following [36], we applied person 1-16 for training and person 17-25 for test. The size of the input video was resized to 128×128 . We set learning rate as 1×10^{-5} , $\lambda_2 = 1$, $\lambda_4 = 1$ and $\beta = 0.001$ for training. MSR action dataset, in comparison, is closer to realistic scenarios with more cluttered background. We adopted a similar setting as that of KTH dataset and apply person 2 and person 6 for test. Following Mathieu *et al.* [26], the number of input frames was also 10 and the goal was to predict the rest 10 future frames.

In Fig.4, we can see that MCNet performs better than DrNet. On KTH dataset, DrNet often leads to loss of temporal information or distortion of the target. Although the content

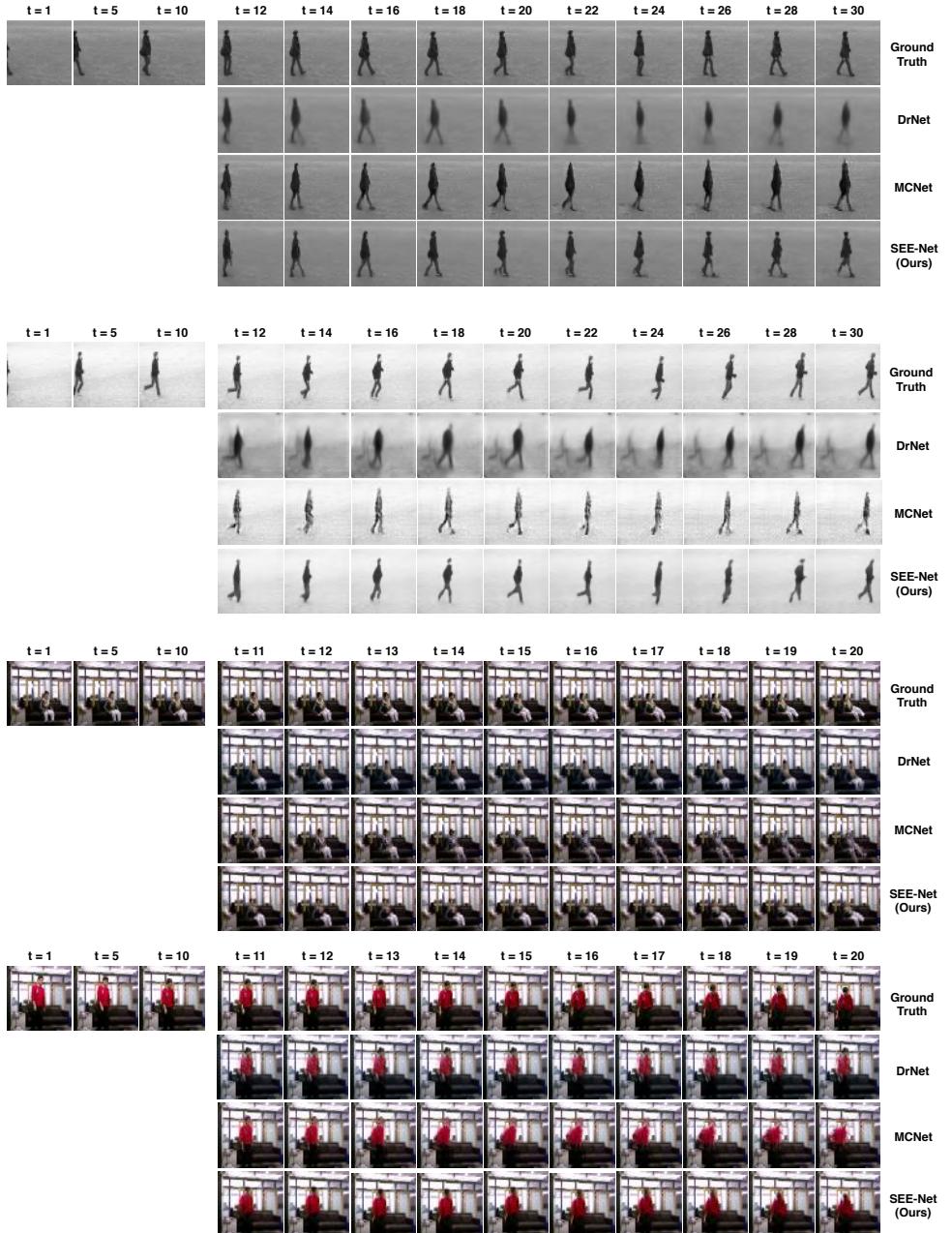


Figure 4: Qualitative comparison to state-of-the-art methods on KTH and MSR datasets. Due to the paper length, we only visualise input frame 1, 5 and 10 and predicted results on both datasets.

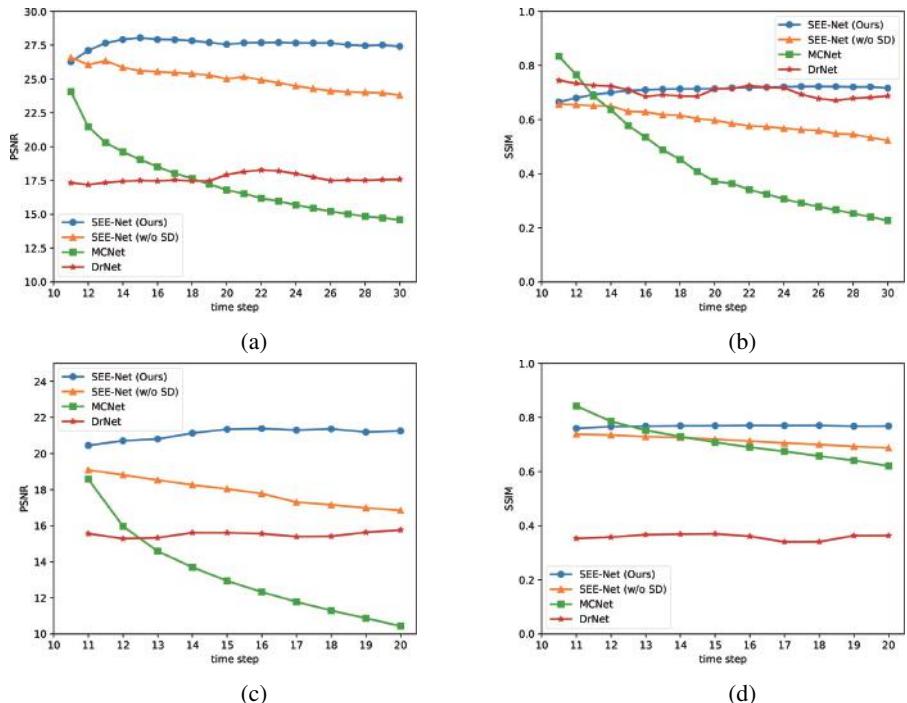


Figure 5: Quantitative results of PSNR and SSIM on KTH (a and b) and MSR (c and d) datasets. Compared with MCNet, DrNet, and our model without shuffling sequence.

and motion information is forced preserved by the disentangled representation, the model fails to predict the correct movement trends between frames. MCNet can better match the trend of moving targets. However, it suffers from severe content information loss. The details of person and face are significantly distorted on both of the datasets. In contrast, our SEE-Net first demonstrates accurate trend prediction that can well synchronise with the ground truth movement. Meanwhile, our method can preserve more content details. We ascribe our success to the shuffle discriminator that can not only detect shuffled orders but also examine whether the generated features are realistic.

In order to understand the overall performance on the whole dataset, we followed Mathieu *et al.* [26] and employed Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) as quantitative evaluation metrics in Fig.5, from which we can draw similar conclusions as that from the qualitative comparison. Firstly, DrNet tends to preserve better content information and therefore achieves higher PSNR and SSIM scores compared to that of MCNet in KTH action dataset. In comparison, MCNet is more predictive so the scores are changing with the consecutive prediction. But its temporal information is gradually losing that results in severe detail loss of the generated frames. Our method consistently outperforms all of the compared approaches on both datasets and evaluation metrics.

It is also interesting to note that, in Fig.5, the PSNRs and the SSIMs with respect to the SEE-Net (with or without SD) and the DrNet are more consistent than those of the MCNet. In our work, based on the assumption that a relatively stable background is provided, we fuse the content feature and the motion feature for future frame generation as described in Eq.1. Compared with MCNet, our approach can effectively avoid the accumulated error in

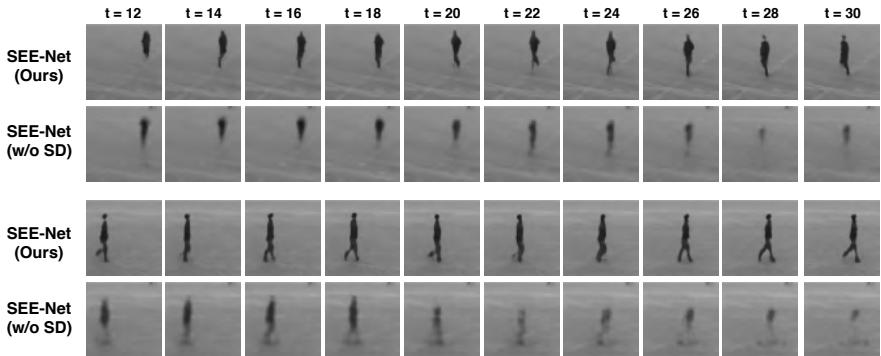


Figure 6: Comparing the predicted results on the KTH dataset based on the proposed SEE-Net and that without *SD*.

the generation process.

Effect of Shuffling Sequence Generation To show the major contribution provided by our proposed shuffling method, we conducted an ablation study that removed the effect of shuffle discriminator by changing its hyper-weight to zero. Clear evidence can be found from the consecutive generated frames in Fig.6, where the predicted target gradually resolves due to loss of temporal information. The powerful temporal and content information extracted by the consistency loss and optical flow network can retain the predictive power. However, the sequential information is not enhanced by discriminating between correct and incorrect frame orders. Therefore, its overall performance is not as good as SEE-Net with shuffle discriminator.

6 Conclusion

This paper investigated shuffling sequence generation for video prediction using a proposed SEE-Net. In order to discriminate natural order from shuffled ones, sequential order information was forced to be extracted. The introduced consistency loss and optical flow network effectively disentangled content and motion information so as to better support the shuffle discriminator. On both synthetic and realistic datasets, SEE-Net consistently achieved improved performance over state-of-the-art approaches. The in-depth analysis showed the effect of shuffling sequence generation in preserving long-term temporal information. In addition to the improved visual effects, SEE-Net demonstrated more accurate target prediction accuracy in both qualitative and quantitative evaluations. One of the important future work directions is to investigate new alternatives to reconstruction loss because it leads to averaged output in a long-term process. Another issue to be resolved is the computational cost that limits the sequence prediction length at training stages.

7 Acknowledgements

Bingzhang Hu and Yu Guan are supported by Engineering and Physical Sciences Research Council (EPSRC) Project CRITiCaL: Combatting cRiminals In The CCloud (EP/M020576/1). Yang Long is supported by Medical Research Council (MRC) Fellowship (MR/S003916/1).

References

- [1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.
- [2] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017.
- [3] Michael B Chang, Tomer Ullman, Antonio Torralba, and Joshua B Tenenbaum. A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*, 2016.
- [4] Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, pages 4414–4423, 2017.
- [5] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.
- [6] SM Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Geoffrey E Hinton, et al. Attend, infer, repeat: Fast scene understanding with generative models. In *Advances in Neural Information Processing Systems*, pages 3225–3233, 2016.
- [7] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in neural information processing systems*, pages 64–72, 2016.
- [8] Ruohan Gao, Dinesh Jayaraman, and Kristen Grauman. Object-centric representation learning from unlabeled videos. In *Asian Conference on Computer Vision*, pages 248–263. Springer, 2016.
- [9] Yu Guan and Thomas Plötz. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):11, 2017.
- [10] Yu Guan, Chang-Tsun Li, and Sruti Das Choudhury. Robust gait recognition from extremely low frame-rate videos. In *2013 International Workshop on Biometrics and Forensics (IWBF)*, pages 1–4. IEEE, 2013.
- [11] Yu Guan, Chang-Tsun Li, and Fabio Roli. On reducing the effect of covariate factors in gait recognition: a classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence*, 37(7):1521–1528, 2014.
- [12] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems*, pages 515–524, 2018.

- [13] BingZhang Hu, Yan Gao, Yu Guan, Yang Long, Nicholas Lane, and Thomas Ploetz. Robust cross-view gait identification with evidence: A discriminant gait gan (diggan) approach on 10000 people. *arXiv preprint arXiv:1811.10493*, 2018.
- [14] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.
- [15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [16] Yu Kong, Zhiqiang Tao, and Yun Fu. Deep sequential context networks for action prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1473–1481, 2017.
- [17] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. A hierarchical representation for future action prediction. In *European Conference on Computer Vision*, pages 689–704. Springer, 2014.
- [18] Ivan Laptev, Barbara Caputo, et al. Recognizing human actions: a local svm approach. In *null*, pages 32–36. IEEE, 2004.
- [19] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018.
- [20] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017.
- [21] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 336–345, 2017.
- [22] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 9–14. IEEE, 2010.
- [23] Yang Long, Li Liu, Ling Shao, Fumin Shen, Guiguang Ding, and Jungong Han. From zero-shot learning to conventional supervised classification: Unseen visual data synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1627–1636, 2017.
- [24] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.
- [25] Shugao Ma, Leonid Sigal, and Stan Sclaroff. Learning activity progression in lstms for activity detection and early detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1942–1950, 2016.

- [26] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.
- [27] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016.
- [28] Marc Oliu, Javier Selva, and Sergio Escalera. Folded recurrent neural networks for future video prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 716–731, 2018.
- [29] Viorica Patrascu, Ankur Handa, and Roberto Cipolla. Spatio-temporal video autoencoder with differentiable memory. *arXiv preprint arXiv:1511.06309*, 2015.
- [30] Chris Paxton, Yotam Barnoy, Kapil Katyal, Raman Arora, and Gregory D Hager. Visual robot task planning. *arXiv preprint arXiv:1804.00062*, 2018.
- [31] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.
- [32] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015.
- [33] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015.
- [34] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018.
- [35] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.
- [36] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033*, 2017.
- [37] Carl Vondrick and Antonio Torralba. Generating the future with adversarial transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1020–1028, 2017.
- [38] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621, 2016.
- [39] Jacob Walker, Abhinav Gupta, and Martial Hebert. Patch to the future: Unsupervised visual prediction. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3302–3309, 2014.

- [40] Jacob Walker, Abhinav Gupta, and Martial Hebert. Dense optical flow prediction from a static image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2443–2451, 2015.
- [41] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015.
- [42] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.
- [43] Tianfan Xue, Jiajun Wu, Katherine Bouman, and Bill Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2016.
- [44] Haofeng Zhang, Yang Long, Yu Guan, and Ling Shao. Triple verification network for generalized zero-shot learning. *IEEE Transactions on Image Processing*, 28(1): 506–517, 2018.
- [45] Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, and Ling Shao. Towards universal representation for unseen action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9436–9445, 2018.