Video Prediction with Bidirectional Constraint Network

Ruibing Hou[1,2], Hong Chang[1,2], Bingpeng Ma[2], Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]School of Computer and Control Engineering,
University of Chinese Academy of Sciences, Beijing, 100049, China
{ruibing.hou, hong.chang, xilin.chen}@vipl.ict.ac.cn, bpma@ucas.ac.cn

***Abstract*— Future frame prediction in videos is promising avenue for unsupervised video representation learning. However video prediction has the huge solution space since the high-dimensionality and inherent uncertainty of the future video frames. Existing approaches impose weak constraints on the predictions, which results in motion confusion. To alleviate this problem, we propose a novel model named *Bidirectional Constraint Network (BCnet)*. BCnet consists of forward prediction module and backward prediction module. The forward prediction module learns to predict the future sequence from the present sequence, while the backward prediction module learns to invert the task. The closed loop of the two modules allows that the backward prediction module generates informative feedback signals. The feedback signals clamp down the solution space of forward prediction module. Therefore, our approach can effectively alleviate the motion confusion. We further evaluate BCnet by fine-tuning it for a supervised learning problem: human action recognition on the UCF-101 dataset. We show that the representation help improve classification accuracy. Extensive experiments on several challenging public datasets show that our approach significantly outperforms state-of-the-art approaches, which demonstrates the effectiveness and generalization ability of our approach.**

I. INTRODUCTION

Despite the great process of deep learning architectures for supervised learning, unsupervised video representation learning for general visual tasks remains a unsolved yet critical research problem. Recently, video frame prediction (predicting future frames in a video sequence) [1], [2] has surged as a promising direction for unsupervised learning of video data.

There have been a lot of works on the field of video frame prediction [3], [4], [1], [2], [5], [6]. They often incorporate the ideas from the generative models, which use Long Short-Term Memory (LSTM) [3], [4], [1], [2], Variational Autoencoders [5], or PixelCNN [6] as backbones. However, because of the inherent uncertainty and the high-dimensionality of the future frames, there is huge solution space for these models.

The existing approaches typically apply image pixel loss [1] or adversarial loss [2] on the prediction of future sequence, which are very weak constraints. So these methods do not have enough capacity to effectively constrain the solution space of prediction. Therefore, they often produce non-interpretable outputs, which mainly embodies as *motion confusion*. One type of confusion is about the location of motion. Figure 1 (c) shows a walking sequence predicted by [1]. There are two left legs entangling with each other, due to the confusion with the location of the motion of the left leg. The other motion confusion comes from the distortion of motion category. As shown in Figure 1 (d) predicted by [2], the motion of the sequence distorts from walking to running. In fact, the pace and pose variety of running are much larger than those of walking.

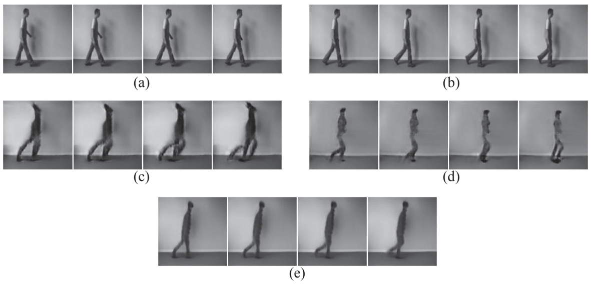


Fig. 1. Video prediction for *walking* sequence in KTH by different approaches. (a) the input sequence for prediction. (b) the ground-truth. (c) the sequence generated by [7] with motion location confusion. (d) the sequence generated by [2] with distortion of motion category. (e) the sequence generated by BCnet with alleviated motion confusion.

To effectively constrain the solution space, we propose a novel architecture named *Bidirectional Constraint Network* (BCnet). BCnet uses backward information to regularize the prediction of the future sequence. Intuitively, from a non-interpretable future sequence, the backward prediction module will generate the present sequence that deviates even further from the ground-truth. Therefore, through constraining the present sequence generated by backward prediction to be as close as possible to the ground-truth, our model obtains much smaller solution space. Compared to the previous models, the sequences that make up the solution space are more reasonable from the perspective of appearance and motion. In addition, we design a novel objective function that comprises image pixel loss and video adversarial loss to implement the bidirectional constraint. The objective function is used to constrain the appearance and motion of the prediction to be realistic. As Figure 1 (e) shows, our approach has no motion confusion and preserves the human shape more accurately. The main contributions of this paper are summarized below:

- We propose an effective framework for video prediction,

978-1-7281-0089-0/19/$31.00 ©2019 IEEE

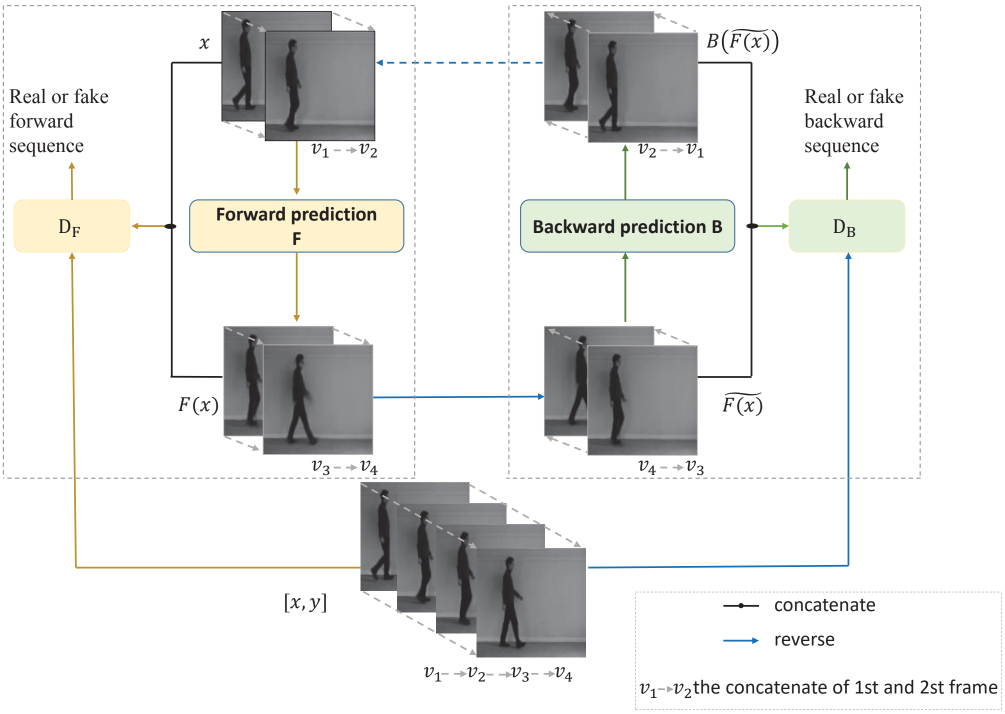


Fig. 2. The architecture of BCnet. It contains two video prediction modules F and B and two adversarial discriminators D_F and D_B, where the modules F and B use the same architecture.

which makes use of the bidirectional information to constrain the video sequence generation.

- We design a novel objective function to implement the bidirectional constraint that comprises image pixel loss and video adversarial loss.
- We conduct extensive experiments to verify the effectiveness and generalization ability of our model. The experiment results demonstrate that our approach significantly outperforms state-of-the-art approaches.

II. Related Work

Video Prediction. Recently, a number of approaches have been proposed for video prediction. A line of research [1], [8], [9], [2], [10], [11], [12] focuses on developing advanced networks to directly generate pixel values. [1] applied a sequence-to-sequence model to video prediction and showed that LSTM is able to capture pixel dynamics. [8] employed an adversarial loss in a multi-scale network. [9] presented a deep predictive coding network where each layer outputs a layer-specific prediction. [2] decomposed the motion and content to model the dynamics in the video. [10] contains a two-stream generation architecture which deals with high frequency video content and low frequency video content respectively. [11] aggregated contextual information for each pixel in all possible directions at each processing layer to cover the available context. [12] propose a new type of recurrent auto-encoder with state sharing between encoder and decoder. However, these approaches often generate video frames with motion confusion, since they usually impose weak constraints on the prediction which cannot effectively constrain the huge solution space.

Another line of approaches [13], [14], [15], [16], [17], [18], [19] resorts to predicting motion field for utilizing the existing pixel values in previous frames. [13] describe a spatio-temporal video autoencoder with a nested differential short-term memory module that features a modified Spatial Transformer Network Layer [20] for improved motion estimation and frame prediction. [14] explores a hybrid multi-task framework to jointly optimize optical flow estimation and frame prediction. [15] dynamically generated convolutional filters, which are applied to the previous frames for generating the future ones. [16], [17] estimated optical flow between frames and then extrapolated along the optical flow direction. [18] generated the future by transforming pixels in the past with adversarial learning. [19] predicted future flows in multiple time steps and then generated the pixel-level appearance of future frames based on predicted flows. These approaches usually have a smaller solution space because the motion trajectories are much more tractable and lower dimensional than pixel appearances. However, they are failed in long-term video prediction since they cannot handle objects entering from outside the visual field.

Our model directly hallucinates the pixels of unseen future sequence because we aim to generate longer-term video prediction. In contrast to above methods, we introduce an additional backward prediction module to regularize the future prediction, which effectively alleviates motion confusion though constraining the huge pixel-wise solution space.

Bidirectional Method. There is a long history to use transitivity to regularize structured data. In visual tracking, enforcing simple forward-backward consistency has been a common technique for decades [21]. In the language domain,

[22] improves machine translation via back translation and reconstruction. More recently, the bidirectional model has been used in unpaired image-to-image translation [23], [24] and video super-resolution [25]. In this work, we introduce a bidirectional mapping to constrain the solution space for video prediction. Our work is different from above methods as it has totally different objective and model architecture.

III. Bidirectional Constraint Network

In this section, we first describe the proposed BCnet. Then, we discuss how BCnet implements bidirectional constraint. Finally, we introduce the full objective to optimize BCnet.

A. Network architecture

The architecture of BCnet is shown in Figure 2. Formally, BCnet takes a present video sequence ($x = [v_1, ..., v_n]$) as the input to predict the future sequence ($y = [v_{n+1}, ..., v_{n+n}]$), where v_i is the i-th frame of the video sequence v. BCnet includes two modules: *forward spatial-temporal prediction* model F and *backward spatial-temporal prediction* model B. The module F: $x \rightarrow y$ maps the present sequence to the future sequence and its output is $\hat{y} = F(x)$. The module B: $\widetilde{y} \rightarrow \widetilde{x}$ maps the future sequence to the present sequence, where $\widetilde{x}$ denotes the inverse sequence of x, i.e., $\widetilde{x} = [v_n, ..., v_1]$. In addition, we introduce two video adversarial discriminators D_F and D_B. The discriminator D_F aims to distinguish the real forward sequence of $[x, y]$ and the fake forward sequence $[x, F(x)]$. The discriminator D_B aims to distinguish the real backward sequence of $[\widetilde{y}, \widetilde{x}]$ and the fake backward sequence $[\widetilde{F(x)}, B(\widetilde{F(x)})]$.

B. Forward constraint

In BCnet, the module F can predict the future sequences from present sequences. To constrain the appearance and motion of the predictions to be realistic, image pixel loss and video adversarial loss are grouped to forward constraint.

Image Pixel Loss. In BCnet, image pixel loss is used to match the appearance of the predictions, and defined as:

$$L_I(\hat{z}, z) = \|\hat{z} - z\|_2^2. \tag{1}$$

where L_I is the Euclidean distance between the predicted sequence $\hat{z}$ and target z with respect to individual frames. Image pixel loss guides the model to match pixel appearances directly, which makes the predictions close to the targets.

Video Adversarial Loss. In BCnet, video adversarial loss is used to constrain the motion of the predictions to be realistic, and defined as:

$$L_A(D, \hat{z}, z) = \log D(z) + \log(1 - D(\hat{z})), \tag{2}$$

where D is the video discriminator and aims to distinguish between the sequence $\hat{z}$ predicted by F and the real sequence z, while F tries to generate sequence that look similar to z. Therefore, D aims to maximize this objective against F that tries to minimize it.

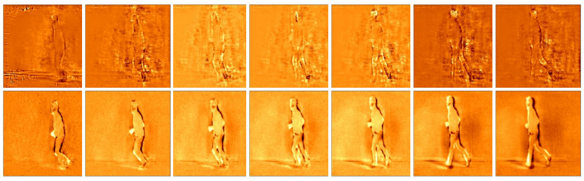


Fig. 3. The heat maps of the gradient images and the difference images. Top: the gradient images. Bottom: the difference images.

Based on image pixel loss and video adversarial loss, the forward constraint can be expressed as:

$$\begin{aligned} L_F(F, D_F) = \mathbb{E}_{[x,y] \sim p_{data}([x,y])}[L_I(F(x), y) \\ + \alpha L_A(D_F, [x, F(x)], [x, y])] \end{aligned} \tag{3}$$

where α is the hyper-parameter to control the relative effect of video adversarial loss during optimization. L_F can provide the forward constraint on F and D_F.

C. Backward Constraint

The forward constraint can constrain the appearance and motion of the predictions of the module F. However, because of the inherent uncertainty and high-dimensionality of the video sequence, the solution space of F is still huge, which leads to motion confusion. As the severe motion confusion often occurs in the last a few frames of the predictions, most predicted pixels have the similar appearance with the targets and the discriminator D_F can hardly distinguish the predictions from the targets. In other words, only applying the forward constraint on the predictions of F cannot effectively alleviate this phenomenon.

Therefore, in this paper, we introduce an additional backward constraint to effectively restrict the solution space and alleviate the motion confusion. Inspired by the reversibility of most physical processes, we argue that the input sequence x should be equivalent to the reversed sequence $\widetilde{x}$ after successively going through the modules F and B. So, in BCnet, we construct the objective for the backward module B and the corresponding discriminator D_B:

$$\begin{aligned} L_B(F, B, D_B) = \mathbb{E}_{[x,y] \sim p_{data}([x,y])}[L_I(B(\widetilde{F(x)}), \widetilde{x}) \\ + \alpha L_A(D_B, [\widetilde{F(x)}, B(\widetilde{F(x)})], [\widetilde{y}, \widetilde{x}])] \end{aligned} \tag{4}$$

Note that the module F, which is our target video predictor, is also involved in the backward constraint.

The backward constraint can alleviate motion confusion by restricting the huge solution space of the module F. When the motion confusion occurs in the forward prediction process, the module B will amplify the bias and produce the sequences that deviate a lot from the targets during the backward prediction process. Consequently, a large penalty is put on the predictions, which makes the module F to be optimized in the direction of reducing motion confusion. In this way, F obtains much smaller solution space that is consisted of more reasonable sequences.

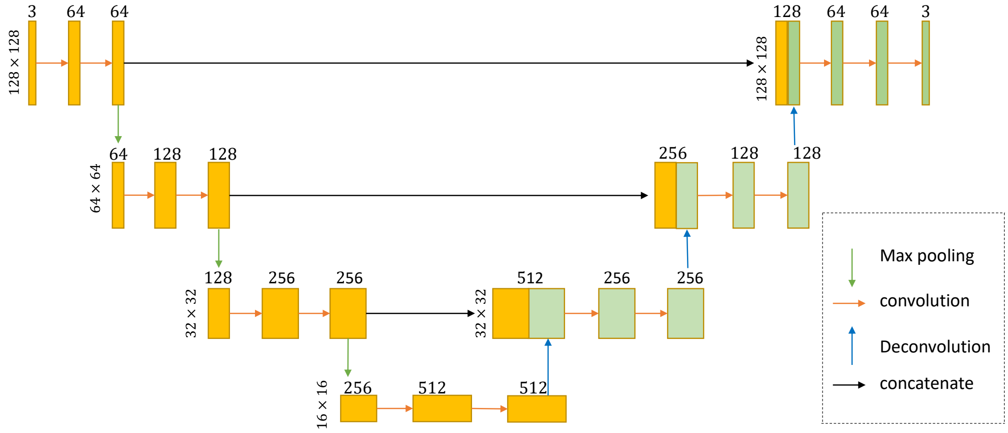


Fig. 4. The architecture of video prediction modules F and B.

D. A Illustrative Example of Backward Constraint

To further demonstrate the effectiveness of backward constraint, we introduce an illustrative example. Let g be the gradient of the backward loss $L_B(F, B, D_B)$ with respect to the prediction of the module F, i.e., $g = \nabla_{\hat{y}} L_B(F, B, D_B)$, and d be the difference between the prediction $\hat{y}$ and the target y, i.e., $d = \hat{y} - y$. Intuitively, if g is consistent with d, g will guide the optimization of F in the direction of making $\hat{y}$ close to y. Figure 3 visualizes g and d by heat maps when a person is running in a static background. We can clearly observe that when some pixel values in d are large, the gradient of the backward loss at those pixels are large accordingly.

E. Full Objective of BCnet

Based on bidirectional constraint, BCnet's full objective can be formulated as:

$$L(F, B, D_F, D_B) = L_F(F, D_F) + L_B(F, B, D_B). \quad (5)$$

We use the pre-trained B to provide the informative signals for F. In order to prevent the parameters of B from being interfered by the generated results of F, we fix the parameters of B. Therefore, we obtain the forward prediction module and two discriminators by optimizing the above objective:

$$F^*, D_F^*, D_B^* = \arg\min_{F} \max_{D_F, D_B} L(F, B, D_F, D_B). \quad (6)$$

In our model, different losses involved in the full objective play the different roles. In the forward prediction process, both the image pixel loss and video adversarial loss play critical role in achieving the high-quality results. Moreover, the backward constraint can alleviate the motion confusion and significantly improve the performance. We verify the roles of different losses by the ablation study in Section VI-B. In addition, we show that the backward constraint can improve the generalization ability of BCnet in Section VI-C and VI-D.

IV. BCnet-Double

In order to further enhance the quality of predictions, we propose BCnet-Inverse based on BCnet. BCnet-Inverse has a similar architecture with BCnet. As shown in Figure 2, the BCnet-Inverse network constrains the input sequence $\widetilde{y}$ to be equivalent to the original sequence y after looping through the modules B and F. In addition, we use the discriminator D_F and D_B to constrain the motion of the predictions generated by the B to be realistic.

Further, BCnet and BCnet-Inverse make up BCnet-Double (shorthand as BCnet-D), where BCnet is used to optimize the module F while BCnet-Inverse to optimize the module B. As shown in Section III-E, B provides the informative signals to assist the optimization of F in BCnet. Intuitively, the more informative the feedback signals provided by B, the more conducive to optimize F. So we use BCnet-Inverse to optimize B further. We alternately optimize BCnet and BCnet-Inverse in the training of BCnet-D. F and B can generate the informative feedback signals to each other for the better video prediction. We compare BCnet-D and BCnet in Section VI-B.

V. Implementation

A. Network configuration

Our model is constructed with identical network architecture for the modules F and B based on Convolutional LSTM Network [7]. As shown in Figure 4, the module is configured with the equal number of downsampling and upsampling layers. In addition, we configure the generator with skip connections between mirrored downsampling and upsampling layers, making it a U-shaped net. Such a design enables the low-level information to be shared between input and output, which is beneficial for prediction since the adjacent frames usually have the similar pixel appearances. In addition, the LSTM is used at the bottleneck layer. For the discriminator networks, we use a five-layer spatial-temporal convolutional network [26] with kernels $3 \times 3 \times 3$ followed by a fully-connected layer. By this way, the discriminator can capture the motion between frames.

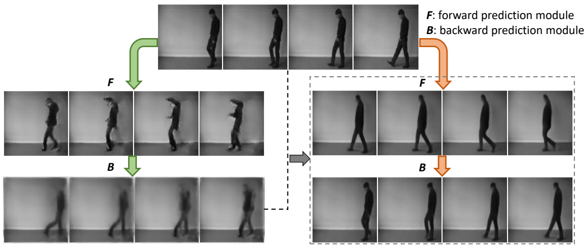


Fig. 5. Visualization of intuitive interpretation of BCnet.

TABLE I

COMPARISONS OF THE AVERAGE PERFORMANCE OF VARIANTS OF OUR APPROACH.

metrics	PSNR			SSIM		
loss	L_I	L_A	$L_I + L_A$	L_I	L_A	$L_I + L_A$
F	24.973	24.124	25.357	0.759	0.710	0.768
BCnet	26.228	24.478	27.300	0.782	0.730	0.824
BCnet-D	26.230	24.501	27.541	0.785	0.731	0.837

B. Training and Inference

The optimization of BCnet (BCnet-Inverse) can be divided into two stages. In the first stage, we pre-train the modules F and B with the image pixel loss. In the second stage, we train the whole network while keeping the parameters in the module B (F) fixed. Following Generative Adversarial Net [27], we apply an alternating gradient update scheme, performing one gradient descent step on the discriminator D_F and D_B, and one step on the predictor F (B). We use the Adam solver [28] with a learning rate of $1e$-4 in the first stage, while a learning rate of $5e$-5 in the second stage. When training BCnet-D, we alternately optimize BCnet and BCnet-Inverse. The modules F and B generate the informative feedback signals to each other, which is more conducive for them to learn from each other.

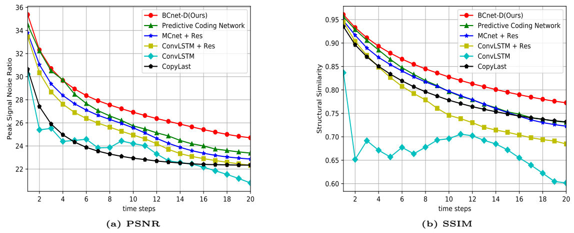
When inference, we only employ the module F, which is used recursively by using the newly generated frame as the input. Notably, although our whole architecture seems sophisticated, the network used for inference is very simple.

VI. EXPERIMENTS

In this section, we first conduct an experiment to validate the intuitive interpretation of BCnet. Then, we present the ablation studies on BCnet. Finally, we verify its effectiveness on the KTH [29] and UCF-101 [30] datasets. Further more, we evaluate the generalization performance of our approach on the Weizmann action [31] and THUMOS-15 [32] datasets, respectively.

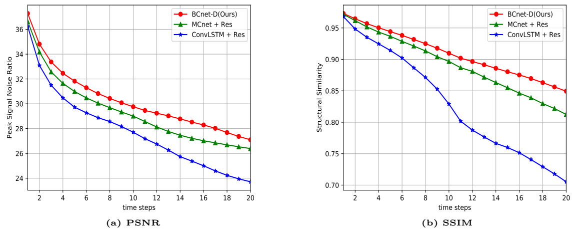
Following [8], [9], [2], [6], we use two common metrics for video prediction analysis: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). PSNR measures the quality of the reconstruction frames and SSIM measures the perceived quality. For both metrics, the higher the values, the better the results.

A. Visualization of intuitive interpretation.



(a) PSNR (b) SSIM

Fig. 6. Frame-wise PSNR and SSIM of different models on KTH.



(a) PSNR (b) SSIM

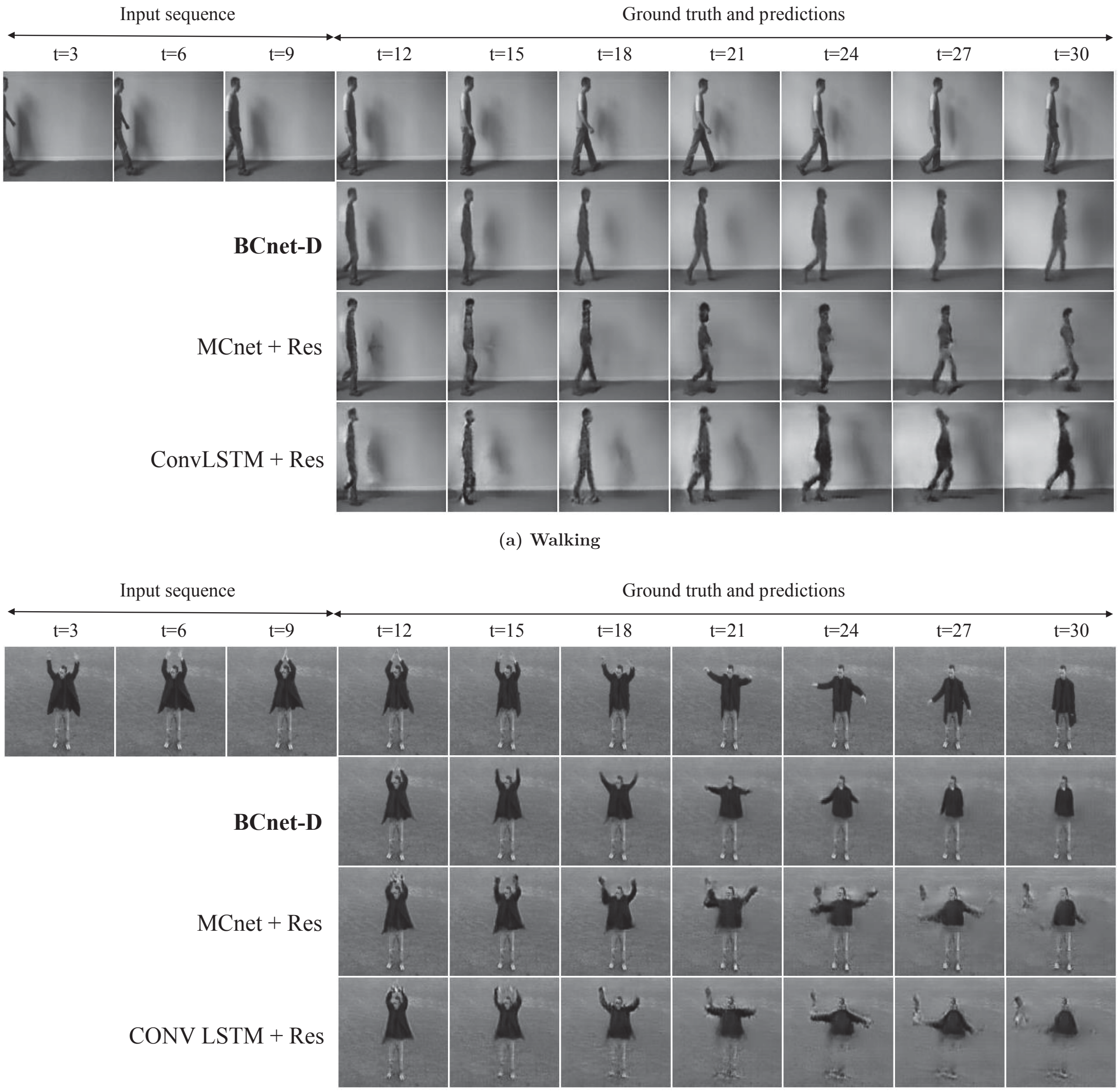
Fig. 7. Frame-wise PSNR and SSIM of different models on Weizmann.

In order to validate the intuitive interpretation of BCnet, we visualize the predicted sequences of forward prediction module F and backward prediction module B. As shown in Figure 5, when F predicts a non-interpretable sequence, B will generate a sequence that deviates even further from the target. The possible reason is that B cannot recognize the motion of input non-interpretable sequence. On the contrary, F can generate more reasonable sequences with the proposed bidirectional constraint. We argue that the additional constraints, which make the sequence generated by B to be close to the target, provide informative signals to F.

B. Ablation Studies

Experimental setting. To assess and appreciate the effect of different loss functions and the contribution of bidirectional constraint, we provide a quantitative verification on the variants of our approach on the KTH human action dataset. The KTH dataset contains six categories of periodic motions which are performed by 25 subjects on simple backgrounds with a static camera. For comparison purposes, we adopt the experiment setup in [2]. All video frames are resized to 128×128 pixels. The subjects 1 to 16 are used for training and 17 to 25 for testing. In the end, we have 9,950 training sequences and 3,361 test sequences. In the training phase, all of the models observe 10 frames and predict the following 10 frames. In the inference, all of the models use 10 frames to generate the following 20 frames. Following [2], we set $\alpha = 0.02$ for training.

Results. Table I shows the comparisons of the average performance of variants of our approach. As Table I shows, the image pixel loss and video adversarial loss (L_I, L_A) facilitate the forward prediction process F to obtain a better performance. BCnet, with various loss functions except L_A, significantly outperform the best performance of F. This result indicates that the bidirectional constraint is conducive to obtaining a better performance and has a good adaptive capacity. In addition, BCnet-D that optimizes the modules F and B alternately acquires the best performance, which further indicates the effectiveness of our approach.



(a) Walking

(b) Hand-Waving

Fig. 8. Prediction samples from KTH dataset. We display predictions starting from the 12^{th} frame, in every 3 time steps.

C. Experiments on KTH and WEIZMAN Action Datasets

Experimental setting. We use the same experimental setting for KTH dataset as Section VI-B. Following [2], we select the walking, running, one-hand waving and two-hands waving sequences from Weizmann dataset to evaluate the generalization ability of our model.

Results. Figure 6 shows the quantitative comparisons of ConvLSTM [7], ConvLSTM+Res, Predictive Coding Network [9], MCnet+Res [2] and our approach on KTH. [1] Besides the state-of-the-art approaches, a baseline method, CopyLast, is also involved in the comparative study. This method simply copies the last observed frame through time as the predicted frame. Our approach outperforms CopyLast significantly, which suggests that our approach can predict the motion of objects rather than just copying the last observed frame.

As shown in Figure 6, our approach consistently outperforms the state-of-the-art approaches. The performance of other approaches deteriorate quickly for long-term predictions. By contrast, the performance of our approach remain stable over time, only with slow and reasonable decline. The reason for this result is that the existing approaches only have forward constraint, which puts a weak penalty on the last a few frames of the predictions. However, the bidirectional constraint in BCnet can increase the penalty on these frames through backward constraint. Therefore, our approach can achieve obviously better result in long-term prediction.

For evaluating the generalization ability of our approach, we directly use the model trained on the KTH dataset to predict on the Weizmann dataset. Figure 7 shows the quantitative comparisons of ConvLSTM+Res, MCnet+Res [2] and our approach on Weizmann. As shown in Figure 7, our approach outperforms ConvLSTM+Res and MCnet+Res. The reason for this result is that the backward constraint used by our approach can regularize the network to improve the

[1] We use the results reported in [2] for [7], [2]. We also reproduce [9] since they did not test the model on KTH.

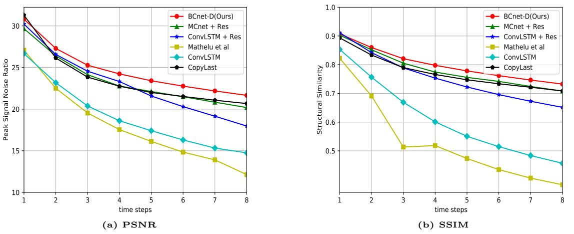


Fig. 9. Frame-wise PSNR and SSIM of different models on UCF-101.

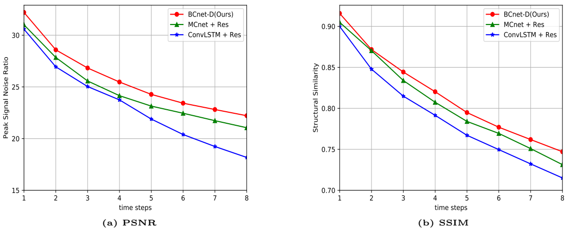


Fig. 10. Frame-wise PSNR and SSIM of different models on THUMOS-15.

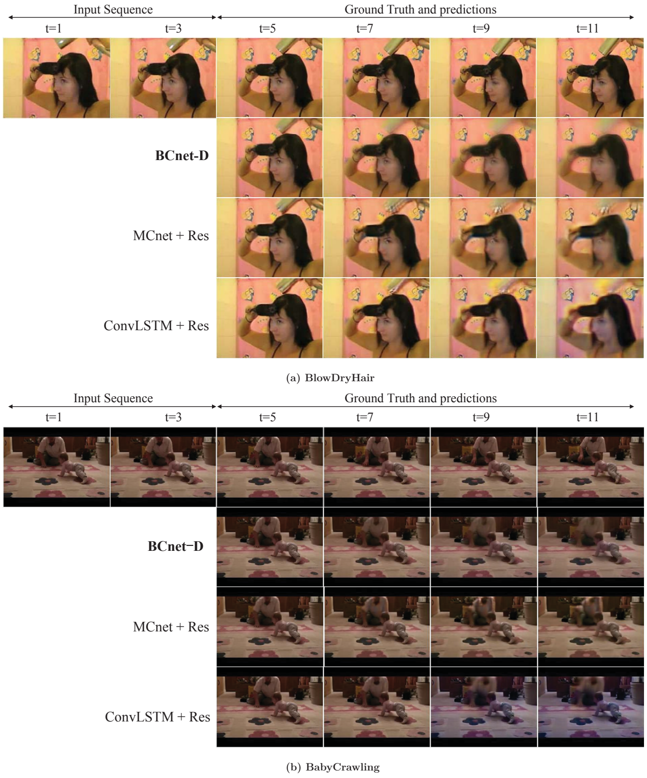
generalization ability.

Figure 8 presents qualitative results of multi-step predictions. From the figure, we can see that ConvLSTM+Res fails to preserve the shape of human body. In addition, there are severe motion confusion and human body deformation in the generated frames of ConvLSTM+Res and MCnet+Res. Our approach predicts human motion and preserves human shape more accurately. Specially, in Figure 8 (a), due to the similarity between the motion of walking and running, MCnet+Res generates the walking sequence in the earlier stage but the running sequence in the later stage. On the contrary, our method can consistently generate the walking sequence, which further suggests that our approach can alleviate the motion confusion.

D. Experiments on UCF-101 and THUMOS-15 Datasets

Experimental setting. UCF-101 is a challenging dataset collected from YouTube. Compared with the previous datasets, this dataset contain much more diversity in objects, backgrounds and camera motions. Specially, it contains 13,320 videos belonging to 101 classes. Following [2], we split 9,537 videos for training and 3,783 videos for testing. We employ the same network architecture as that on the KTH dataset, but resize frames to 240×320 pixels. We train the network to observe 4 frames and predict 4 frames. In the inference, we predict 8 frames. Following [2], α is set to 0.001 for training. In addition, we evaluate the models on the THUMOS-15 dataset for verifying the generalization ability of the different models.

Results. Figure 9 shows the comparisons of CopyLast, ConvLSTM [7], ConvLSTM+Res, BeyondMSE [8], MCnet+Res [2] and our approach on UCF-101. [2] Notice that CopyLast has a higher quantitative score, because of the small motion of the videos in UCF-101. And it outperforms most other approaches, which can not precisely locate the motion. Our approach still outperforms the other approaches by a large margin, because the bidirectional constraint can restrict the motion in a reasonable range.



(a) BlowDryHair

(b) BabyCrawling

Fig. 11. Prediction samples from UCF-101 dataset. We display predictions starting from the 5^{th} frame, in every 2 time steps.

Figure 10 shows the comparisons against ConvLSTM+Res and MCnet+Res on THUMOS-15. Our approach still achieves the best performance, which implies that our model can generalize well to new datasets.

Figure 11 visualizes sample video sequences from the UCF-101 test set. Our approach can generate video frames of impressive quality. However, motion confusion and object deformation trouble the images predicted by other approaches. For example, in Figure 11 (a), the face in the sequence predicted by MCnet+Res tends to move to right following the movement of the hair, resulting in the deformation of the face. Our method can capture the motion of the hair accurately and preserve the shape of the face.

E. Unsupervised Representation Learning

To show the effectiveness of our model on unsupervised video representation learning, we replace the decoder of the forward prediction module F with a classification layer (i.e., fully-connected layer + softmax loss). Our model is then fine-tined and tested with an action recognition loss on the UCF-101 dataset (split-1). This is equivalent to using frame prediction as a pre-training task. As demonstrated in Table II, our approach outperforms random initialization by a large margin and also shows superior performance to other representation learning approaches. To predict the future

[2] We use the results reported in [2] for [7], [2], [8].

TABLE II

CLASSIFICATION ACCURACY OF ACTION RECOGNITION ON UCF-101.

Method	Accuracy
Random	39.1
Unsupervised Video [1]	43.8
Shuffle and Learn [33]	50.2
DVF [17]	52.4
BCnet-D	53.0

frames, BCnet has to encoder both appearance and motion information, which implicitly mimics a two-stream CNN.

VII. CONCLUSION AND FUTURE WORK

In this paper, we propose bidirectional constraint network for pixel-level prediction of future frames in natural video sequences. The bidirectional constraint network couples a forward spatial-temporal prediction module with a backward spatial-temporal prediction module. Experiment results demonstrate that our method can relieve the motion confusion and outperform state-of-the-art approaches on multiple datasets.

In the future, we would like to extend the bidirectional constraint on other video prediction models. Even a native application of bidirectional constraint based on ConvLSTM module already shows encouraging results. We believe that a much better performance can be achieved with more advanced prediction models. Another possible future direction is to apply the similar idea to solve other video representation learning problems.

Acknowledgement This work is supported in part by National Basic Research Program of China (973 Program): 2015CB351802, and Natural Science Foundation of China (NSFC): 61876171, 61702486 and 6180618.

REFERENCES

[1] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *International Conference on Machine Learning (ICML)*, 2015, pp. 843–852.

[2] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," in *International Conference on Learning Representations (ICLR)*, 2017.

[3] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra, "Video (language) modeling: a baseline for generative models of natural videos," *arXiv preprint arXiv:1412.6604*, 2014.

[4] S. H. J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] T. Xue, J. Wu, K. L. Bouman, and W. T. Freeman, "Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks," in *Neural Information Processing Systems (NIPS)*, 2016, pp. 91–99.

[6] N. Kalchbrenner, A. v. d. Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu, "Video pixel networks," in *International Conference on Machine Learning (ICML)*, 2017.

[7] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. K. Wong, and W. c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Neural Information Processing Systems (NIPS)*, 2015, pp. 802–810.

[8] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," in *International Conference on Learning Representations (ICLR)*, 2016.

[9] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," in *International Conference on Learning Representations (ICLR)*, 2017.

[10] X. Jingwei, N. Bingbing, L. Zefan, C. Shuo, and Y. Xiaokang, "Structure preserving video prediction," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[11] W. Byeon and Q. Wang, "Contextvp: Fully context-aware video prediction," in *European Conference on Computer Vision (ECCV)*, 2018.

[12] M. Oliu, J. Selva, and S. Escalera, "Folded recurrent neural networks for future video prediction," in *European Conference on Computer Vision (ECCV)*, 2018.

[13] V. Patraucean, A. Handa, and R. Cipolla, "Spatio-temporal video autoencoder with differentiable memory," *arXiv preprint arXiv:1511.06309*, 2015.

[14] N. Sedaghat, "Next-flow: Hybrid multi-tasking with nextframe prediction to boost optical-flow estimation in the wild," *arXiv preprint arXiv:1612.03777*, 2016.

[15] B. D. Brabandere, X. Jia, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in *Neural Information Processing Systems (NIPS)*, 2016, pp. 667–675.

[16] X. Liang, L. Lee, W. Dai, and E. P. Xing, "Dual motion gan for future-flow embedded video prediction," in *International Conference on Computer Vision (ICCV)*, 2017.

[17] Z. Liu, R. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *International Conference on Computer Vision (ICCV)*, 2017.

[18] C. Vondrick and A. Torralba, "Generating the future with adversarial transformers," in *Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2992–3000.

[19] L. Yijun, F. Chen, Y. jimei, W. zhaowen, L. Xin, and Y. Ming-Hsuan, "Flow-grounded spatial-temporal video prediction from still images," in *European Conference on Computer Vision (ECCV)*, 2018.

[20] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Neural Information Processing Systems (NIPS)*, 2015, pp. 2017–2025.

[21] N. Sundaram, T. Brox, and K. Keutzer, "Dense point trajectories by gpu-accelerated large displacement optical flow," in *European Conference on Computer Vision (ECCV)*, 2010, pp. 438–451.

[22] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T. Liu, and W.-Y. Ma, "Dual learning for machine translation," in *Neural Information Processing Systems (NIPS)*, 2016, pp. 820–828.

[23] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *International Conference on Computer Vision (ICCV)*, 2017.

[24] Z. Yi, H. Zhang, T. Gong, Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *International Conference on Computer Vision (ICCV)*, 2017.

[25] Y. Huang, W. Wang, and L. Wang, "Video super-resolution via bidirectional recurrent convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[26] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 4489–4497.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Neural Information Processing Systems (NIPS)*, 2014, pp. 2672–2680.

[28] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *International Conference on Pattern Recognition (ICPR)*, 2004, pp. 32–36.

[30] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[31] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *International Conference on Computer Vision (ICCV)*, 2005, pp. 1395–1402.

[32] A. Gorban, H. Idrees, Y. Jiang, A. R. Zamir, I. Laptev, M. Shah, and R. Sukthankar, "Thumos challenge: Action recognition with a large number of classes," in *Computer Vision and Pattern Recognition WorkShop (CVPRW)*, 2015.

[33] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: unsupervised learning using temporal order verification," in *ECCV*, 2016, pp. 527–544.