LARNet: Latent Action Representation for Human Action Synthesis

Naman Biyani[1]
namanb@iitk.ac.in
Aayush J Rana[2]
aayushjr@knights.ucf.edu
Shruti Vyas[2]
shruti@crcv.ucf.edu
Yogesh S Rawat[2]
yogesh@crcv.ucf.edu

[1] IIT Kanpur
Kanpur, India
[2] Center for Research in Computer Vision
University of Central Florida,
Florida, USA

Abstract

We present LARNet, a novel end-to-end approach for generating human action videos. A joint generative modeling of appearance and dynamics to synthesize a video is very challenging and therefore recent works in video synthesis have proposed to decompose these two factors. However, these methods require a driving video to model the video dynamics. In this work, we propose a generative approach instead, which explicitly learns action dynamics in latent space avoiding the need of a driving video during inference. The generated action dynamics is integrated with the appearance using a recurrent hierarchical structure which induces motion at different scales to focus on both coarse as well as fine level action details. In addition, we propose a novel mix-adversarial loss function which aims at improving the temporal coherency of synthesized videos. We evaluate the proposed approach on four real-world human action datasets demonstrating the effectiveness of the proposed approach in generating human actions. Code available at https://github.com/aayushjr/larnet.

1 Introduction

Video generation is a challenging problem with a lot of applications in robotics [18, 33], augmented reality [64, 65], data augmentation [4, 22, 50, 74], and action imitation [1, 13, 37, 39, 55, 60, 62]. It has different variations, such as video prediction [28, 31, 71], video synthesis [64, 65], video interpolation [42, 54], and super-resolution [14, 27, 38]. In this work, we focus on generating human actions conditioned on image of an actor and a target action. A generative approach can be used to predict the future frames in a video, but the problem of predicting future frames from a single input image is ill-posed. Given an input frame, a generated video can have a different sequence of future frames depending on the variation in the performed action. A use of stochastic noise can address this limitation [30, 44, 45, 60, 66], but a joint generative modeling of appearance and motion for video synthesis is very challenging (Figure 1(a)).

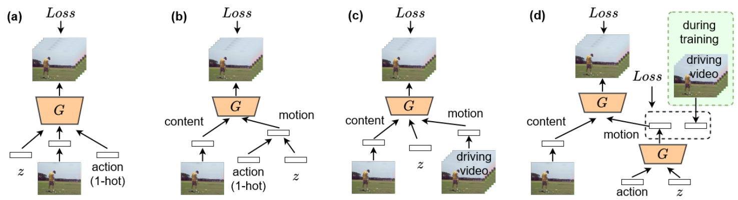

Figure 1: Overview of different approaches for video synthesis. (a) joint generative modeling [30, 44, 45, 60, 66], (b) implicit disentanglement of content and motion [41, 55, 58, 67], (c) motion transfer from a driving video [7, 25, 41], and (d) proposed approach, explicit disentanglement of content and motion without requiring any driving video during inference.

To address this, the existing approaches utilize a decomposition of appearance and motion [34, 41, 55, 58, 67] (Figure 1(b)). This enables the model to independently learn the variations in the performed action and helps in video synthesis. With a similar motivation, there are approaches explicitly using a prior motion which avoids the need of motion modelling and simplifying the complexity of video synthesis (Figure 1(c)). This approach has been found very effective for video synthesis by motion transfer [7, 25, 41].

We want to benefit from both these approaches (Figure 1 (b) and (c)) for conditional human action synthesis. However, each of these come with their own limitations. In the first approach, the decomposition of appearance and motion is not explicit as the only supervision comes from the generated video which limits the potential of this disentanglement. And in the second approach, the explicit use of motion information requires a synchronized driving video during inference which also restricts the motion generating capability of the model.

We propose LARNet, a generative framework which attempts to benefit from both these approaches and simultaneously overcome the above two limitations. LARNet explicitly models the action dynamics in latent space by approximating it to motion from real action videos. This enables effective decomposition of appearance and motion while avoiding the need of any driving video during inference (Figure 1(d)). The disentangled appearance and motion features needs to be integrated effectively for video synthesis. LARNet utilizes a recurrent hierarchical structure for this integration focusing at different scales for capturing both coarse as well fine-level action details. The proposed method is trained end-to-end in an adversarial framework, optimizing multiple objectives. We make the following novel contributions in this work,

1. We propose a generative approach for human action synthesis that leverages the decomposition of content and motion by explicit modeling of action dynamics.

2. We propose a hierarchical recurrent motion integration approach which operates at multiple scales focusing on both coarse level and fine level details.

3. We propose mix-adversarial loss, a novel objective function for video synthesis which aims at improving the temporal coherency in the synthesized videos.

We validate our approach on several real-world human action datasets, showing its effectiveness in generating human action videos.

2 Related Work

Video Prediction Video prediction task predicts future frames by conditioning on the input frame(s)[20, 28, 30, 69, 71]. Using future frames as ground-truth leads to conditioned supervised learning approach which gives better results in contrast to unconditional video generation [8, 18, 28, 39]. GAN based approaches often relies on a sequence of input frames as priors to reduce ambiguity [15, 19, 62, 71]. Our approach uses only the first input frame and action class name as prior for the prediction task similar to [28, 60].

Video Synthesis Although GANs have been successful in image synthesis task [5, 11, 32, 63, 72], synthesizing a high resolution realistic video is still challenging due to the temporal complexity and resource requirements [13, 44, 55, 60, 61]. GANs use RNN architectures [36, 55], progressive generative models [14, 17, 35] or decoupled two-stream approach [53, 60] to address this. Unconditional video GANs rely on various forms to improve on spatio-temporal consistency such as random noise input [13, 55], two-stream learning [55, 60], multi-scale approach [44] and increasing computing power [13]. In contrast, conditional video synthesis task is able to generate higher quality videos and easily learn the data distribution. Conditional GANs have many variants that use text [2, 36], speech [10, 40, 75], class label [13, 55, 67], pose information [34, 62, 70], semantics [64, 65], the entire video [4, 46, 49, 61] or only first frame [29, 60]. Our approach uses first frame and class label embedding for conditioning and is evaluated against prior approaches [51, 55, 60, 66, 67].

Motion Transfer Video generation using conditional GANs is also done using additional motion or pose information from image sequences [7, 25, 41]. [41] uses optical flow and synthesizes realistic images. [7, 70] use pose information for motion transfer between videos. Extracting this motion can be a limitation for these approaches. Our method instead uses a generative approach where the motion can be synthesized instead of using a driving video.

3 Approach

Given an input image x^0 of an actor and an action class y_a, our goal is to predict a video v with T frames $x^1, x^2, \ldots, x^T$ depicting how the action will be performed. To solve this, we propose LARNet consisting of two main parts; 1) *action dynamics module*, and 2) *video synthesis module* (Figure 2). The *action dynamics module* generates latent action representation e_m and the *video synthesis module* synthesizes the action video v by integrating the appearance and generated motion features.

3.1 Action Dynamics Module

It has been shown that decoupling appearance and motion component of a video provides more flexibility and improves overall quality for video synthesis [26, 55, 67]. Motivated by this, we propose to model the action dynamics as a latent representation conditioned on the action, independent of the appearance, by using a *motion generator* G_m, which utilizes the action class label y_a to estimate the latent action representation e_m.

Generating action dynamics merely based on a class label can be challenging. Therefore we develop a generative approach where we propose to approximate the generated action representation e_m to motion features $\hat{e}_m$ extracted from a real action video $\hat{v}$. We use a 3D

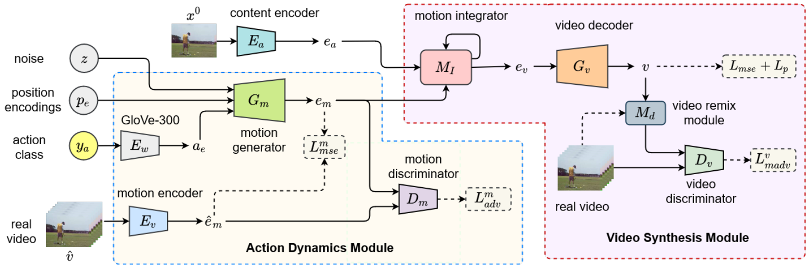

Figure 2: Overview of the proposed framework. Given an actor image x^0, action class y_a, position encoding p_e, and noise z, the network generates corresponding action video v. The motion generator G_m generates action representation e_m in latent space in the action dynamics module. Next, the motion integrator M_I recurrently integrates e_m with the appearance e_a in a latent space to produce video features e_v which is used to synthesize the action video v. The complete network is trained end-to-end with the help of multiple objective functions.

convolution based network E_v to extract motion features from a video [6]. To account for the temporal as well as action dynamics variations, we provide a position encoding p_e and stochastic noise $z \sim \mathcal{N}(0,1)$ to G_m. For position encoding p_e, we use the relative position of the starting frame of the action video $\hat{v}$ which is computed as a ratio of the frame position to the total number of frames in the video. The position encoding makes the learned action representation aware of the temporal variation in the action.

We use action semantics instead of a 1-hot encoding where the action name is converted to word embeddings $a_e = E_w$ with the help of GloVe-300 representation [43]. The semantic encodings perform slightly better than 1-hot and enable the model to also synthesize novel actions. The motion generator G_m is a 3D convolution based network which takes the semantic embeddings a_e, position encoding p_e and stochastic noise z as input and generates latent action representation $e_m = G_m(a_e, p_e, z)$.

3.2 Video Synthesis Module

The *video synthesis module* consists of two components; 1) *motion integrator*, and 2) *video decoder*. The generated action representation e_m is integrated with the appearance prior e_a in a latent space to produce video features e_v using a motion integrator M_I. We propose a recurrent approach which utilizes the generated action representation e_m and transforms the appearance e_a one step at a time according to the learned action. The motion integration module M_I has a recurrent structure based on convolutional Gated Recurrent Unit (Conv-GRU) [3], which takes the encoded prior e_a as input along with the generated action representation e_m and predicts integrated video features e_v. Formally, $e_v = M_I(E_a(x^0), G_m(a_e, p_e, z))$ where, x^0 is the actor image and E_a is the image encoder where we use a 2D conv network [52].

The motion integrator takes the appearance latent representation e_a^{t-1} at each time step and transforms it to e_a^t using the latent action representation. First the foreground f^t and background b^t is separated based on the appearance e_a^t and action features e_m^t using learnable 2D kernels W_f and W_b respectively, $b^t = \sigma(W_b * < e_a^{t-1}, e_m^t >$. Then the background features b_f^t are extracted based on prior latent appearance e_a^{t-1}. The foreground features f_f^t are transformed using action kernels W_a and action features e_m^t. Both foreground and background features are combined to get the generated video features e_v^t for time step t, given as

$$e_v^t = (b^t \odot e_a^{t-1}) + (1 - b^t) \odot [tanh(W_a * < e_m^t, (\sigma(W_f * < e_a^{t-1}, e_m^t >)) \odot e_a^{t-1} >)]. \quad (1)$$

Lastly, the generated video features at each time step are combined together to form integrated video features e_v and video is generated via a video decoder G_v. The integrated latent video features e_v are used to generate a video v where the actor present in the image prior x^0 performs the target action y_a. Formally this can be described as,

$$v = G_v(M_I(E_a(x^0), G_m(a_e, p_e, z))). \quad (2)$$

where the generated latent action representation $e_m = G_m(a_e, p_e, z)$ is integrated with the latent appearance $e_a = E_a(x^0)$ to generate the required video v with the help of a video decoder G_v which is a 3D convolution based network.

Hierarchical Motion Integration To improve on fine action details lost during action encoding, we propose to integrate the motion with appearance at multiple scales using a hierarchical motion integrator, generating coarse to fine features accordingly. In each level, the motion integrator M_I takes the latent appearance features e_a and action representation e_m along with generated video features from previous level and generates video features e_v which are then passed to a video decoder to generate higher resolution features. Similarly, the action generator G_m is trained to generate action features at multiple resolutions.

3.3 Training Objective

We use mean squared error loss L_{mse}^m and adversarial loss L_{adv}^m to learn a latent action representation. We use a 3D convolution based discriminator D_m to differentiate between the generated representation $e_m = G_m(a_e, p_e, z)$ and the motion representation $\hat{e}_m = E_v(\hat{v})$ extracted from a real video. The adversarial objective L_{adv}^m is determined using a Wasserstein loss formulation with a gradient norm penalty [21] for a stable network training.

$$L_{adv}^m = \mathop{\mathbb{E}}_{x \sim \mathbb{P}_g} [D_m(x)] - \mathop{\mathbb{E}}_{\widetilde{x} \sim \mathbb{P}_r} [D_m(\widetilde{x})] + \lambda \mathop{\mathbb{E}}_{\hat{x} \sim \mathbb{P}_{\hat{x}}} [(||\nabla_{\hat{x}} D_m(\hat{x})||_2 - 1)^2]. \quad (3)$$

Here $\mathbb{P}_g$ represents generated representation, $\mathbb{P}_r$ represents extracted representations, and $\mathbb{P}_{\hat{x}}$ represents sampling along straight lines between pairs of points sampled from $\mathbb{P}_g$ and $\mathbb{P}_r$. λ is the penalty coefficient which we set to 10 according to [21].

Mix-adversarial Loss Differentiating between real and generated videos for a discriminator is often easier during initial stages and gradually gets harder as training progresses, which also causes generator saturation. We propose a remix strategy where the generated and real videos are fused together by a video remix module M_d which stochastically remixes their frames. This mix-video v_m is used as fake instead of the generated video for adversarial learning. The key idea is to introduce temporal inconsistency in the generated video, which improves discriminator performance and also forces the generator to synthesize a temporally coherent video. The loss objective L_{madv}^v for mix-adversarial learning is,

$$L_{madv}^v = \mathop{\mathbb{E}}_{x \sim \mathbb{P}_g, \widetilde{x} \sim \mathbb{P}_r} [D_v(Mix(x, \widetilde{x}))] - \mathop{\mathbb{E}}_{\widetilde{x} \sim \mathbb{P}_r} [D_v(\widetilde{x})] + \lambda \mathop{\mathbb{E}}_{\hat{x} \sim \mathbb{P}_{\hat{x}}} [(||\nabla_{\hat{x}} D_v(\hat{x})||_2 - 1)^2] \quad (4)$$

Here $Mix()$ represents frame mixing of generated and real videos, $\mathbb{P}_g$ and $\mathbb{P}_r$ represents distribution of generated and real videos respectively, $\mathbb{P}_{\hat{x}}$ represents sampling along straight lines between pairs of points sampled from $\mathbb{P}_g$ and $\mathbb{P}_r$, and λ is the penalty coefficient.

Method	Driving Video	PSNR ↑	SSIM ↑	FID ↓	FVD ↓
VGAN [60]		15.8	0.74	181.29	15.36
MoCoGAN [55]		-	-	229.26	16.37
G3AN [67]		-	-	183.08	17.13
Monkey-Net [51]	✓	-	-	215.23	22.61
Imaginator [66]		26.1	0.93	157.31	15.90
LARNet† (Ours)	✓	28.1	0.94	166.34	13.45
LARNet†† (Ours)	✓	28.4	0.94	171.34	13.60
LARNet (Ours)		**28.4**	**0.94**	**164.53**	**12.91**

Table 1: Comparison with existing conditional video synthesis methods on the NTU-RGB+D dataset. † and †† uses motion from a driving video where †† uses a driving video instead of generated action during inference while † is trained using a driving video without action dynamics module.

We use *MSE* loss L_{mse} between generated and real videos to push generator to create realistic videos and a perceptual loss L_p to improve the its perceptual quality [27]. The proposed framework is trained end-to-end and the overall training objective is,

$$L = \lambda_1 L^m_{mse} + \lambda_2 L^m_{adv} + \lambda_3 L_{mse} + \lambda_4 L^v_{madv} + \lambda_5 L_p, \tag{5}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, and λ_5 are weights which are determined experimentally.

4 Experiments

We demonstrate the effectiveness of the proposed approach and highlight the benefits of its main components (action representation learning, hierarchical motion integrator, and mix-adversarial loss) via quantitative and qualitative evaluation.

We experimented with four real-world human action datasets including NTU-RGB+D [48], Penn Action [73] , KTH [47] and UTD-MHAD [9] with a resolution of 112x112.

Evaluation Metrics We evaluate the quality of the generated videos using frame level Structural Similarity Index Measure (SSIM) [68] and Peak Signal to Noise Ratio (PSNR) [24] against the ground-truth video. Apart from these, we also evaluate the realism of the generated videos using video level FVD [56, 57], frame level FID [23] scores.

Baselines Our first baseline, BaseNet-1, does not use action dynamics module and the proposed motion integrator. It directly uses the action class instead and performs a joint content and motion learning (Figure 1 (a)). A second baseline, BaseNet-2, utilizes the action dynamics module without any explicit supervision (Figure 1 (b)). Our third baseline, LARNet-Base, uses a supervision on generated action representation (Figure 1 (d)).

4.1 Evaluation on Human Actions

We further evaluate LARNet on four different real-world human action datasets. The computed PSNR, SSIM, FID, and FVD scores are shown in Table 1 and 2. The generated videos on four different action datasets using LARNet are shown in Figure 3. We observe that the generated videos capture the action dynamics for a wide range of human actions. This is true even for those actions where only a slight movement of arms is involved, such as 'hand waving' and 'eating'. We also observe that the quality of the generated action videos is

Method	Dataset	PSNR ↑	SSIM ↑	FID ↓	FVD ↓
G3AN [67]	Penn	-	-	63.1	24.24
Imaginator [66]	Penn	19.3	0.69	64.8	13.88
LARNet (Ours)	Penn	**23.6**	**0.80**	**52.1**	**8.45**
G3AN [67]	UTD	-	-	87.6	17.8
Imaginator [66]	UTD	28.3	0.93	92.3	19.3
LARNet (Ours)	UTD	**29.7**	**0.94**	**77.1**	**16.2**
G3AN [67]	KTH	-	-	173.7	24.1
Imaginator [66]	KTH	25.1	0.82	127.6	15.5
LARNet (Ours)	KTH	**26.6**	**0.87**	**104.3**	**15.3**

Table 2: Comparison with existing conditional video synthesis methods.

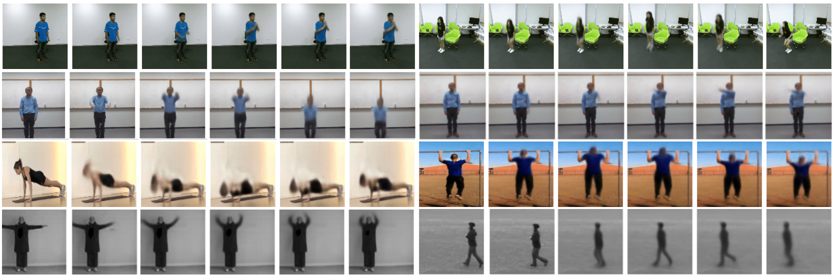

Figure 3: Generated action videos on four different datasets using LARNet including NTU-RGB+D (row 1), UTD (row 2), Penn Action (row 3), and KTH (row 4).

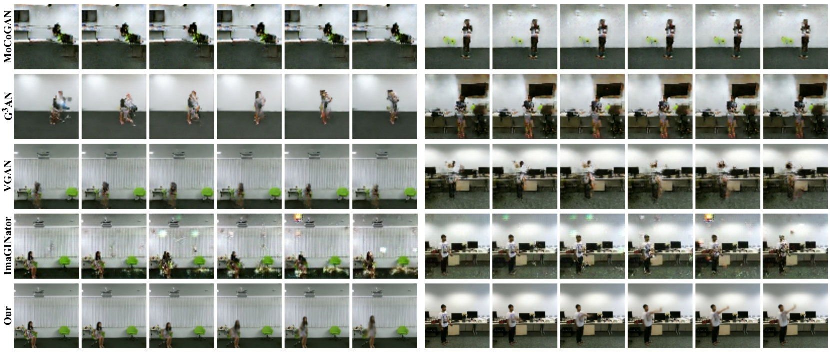

Figure 4: Generated results compared with VGAN [60], MoCoGAN [55] , G3AN [67], and Imaginator [66] on NTU-RGB+D dataset.

much better for UTD when compared with other datasets such as NTU-RGB+D. This can be explained by the complex scene structure and lot of action variations in NTU-RGB+D dataset. Next, we compare the quantitative and qualitative performance of LARNet with existing conditional video synthesis methods, including VGAN [60], MoCoGAN [55], G3AN [67], and Imaginator [66].

Quantitative Comparison We first compare the performance on NTU-RGB+D dataset, which is one of the largest human action dataset, in Table 1. Our method outperforms all

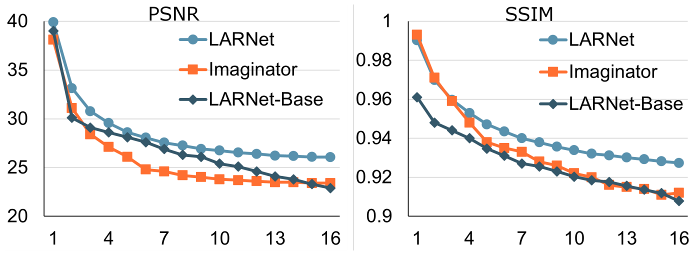

Figure 5: Variation of frame quality in the generated videos with time.

the other approaches in terms of PSNR, SSIM, and FVD scores. We observe that Imaginator [66] has a slightly better performance in terms of FID score which could be due to the use of frame level adversarial loss. However, it is important to note that FID only measures frame level quality whereas FVD is more focused on video dynamics. LARNet outperforms other methods in terms of FVD score.

Next, we compare the performance on small scale datasets including Penn Action, UTD-MHAD and KTH to evaluate the generalization capability of LARNet. We compare with G3AN [67] and Imaginator [66] in Table 2. Even on small sized datasets LARNet consistently outperforms these two methods on all four metrics.

The proposed method generates 16 consecutive frames at a time. In the qualitative results, we observe that the visual quality of frames degrade over time as the action is being generated. To analyze this further, we compare the quality of generated frames independently at each time-step with Imaginator and our baseline model. We utilize PSNR and SSIM scores for this comparison and it is shown in Figure 5. We observe that as we move temporally, the quality degrades for all models but with LARNet the quality is preserved much better than Imaginator which is mainly accredited to the hierarchical motion integrator which helps in preserving the fine level details.

Qualitative Comparison In Figure 4, we show some generated videos for comparison with the existing methods. We observe that LARNet not only can keep better content information than other methods, but also captures the video dynamics for a wide range of actions. Although the other methods are able to generate a good background, they are not able to capture the fine level action details (such as motion of hands). These results show that LARNet can consistently generate the background content of still objects while synthesizing reasonable action dynamics, which clearly outperforms other methods.

4.2 Ablation Study

We perform several ablation experiments to analyze the effectiveness of various components and loss functions in our approach. While the main experiments are done at 112x112 resolution, all the ablations are performed on NTU-RGB+D dataset at a resolution of 56x56.

Effectiveness of Explicit Action Representation To evaluate the effectiveness of explicit action representation, we first train the proposed method without any motion generator (BaseNet-1). Next, we add the motion generator, but without any explicit supervision

Approach	PSNR ↑	SSIM ↑	FID ↓	FVD ↓
BaseNet-1	26.1	0.919	67.1	14.18
BaseNet-2	25.2	0.912	65.3	14.21
LARNet-Base	26.41	0.921	66.5	14.15
LARNet-MI-1	26.83	0.927	64.9	14.11
LARNet-MI-3	**27.25**	**0.931**	**63.5**	**14.02**
LARNet-MI-3 + [L^v_{adv}]	27.23	0.933	63.1	13.89
LARNet-MI-3 + [L^v_{adv}, L^m_{adv}]	27.32	0.937	62.8	13.86
LARNet-MI-3 + [L^v_{madv}, L^m_{adv}]	**27.39**	**0.939**	**62.2**	**13.71**

Table 3: Quantitative comparisons to study the effect of various components of LARNet and the effects of different loss terms on NTU-RGB+D dataset.

Approach	PSNR ↑	SSIM ↑	FID ↓	FVD ↓
$LARNet_{1-hot}$	27.11	0.93	168.34	14.09
$LARNet_{GloVe}$ [43]	**27.40**	**0.94**	165.53	**13.34**
$LARNet_{BERT}$ [16]	27.36	**0.94**	**161.33**	13.67

Table 4: Comparison of using different encodings for the action labels from NTU-RGB+D dataset in our LARNet model. $LARNet_{1-hot}$ uses only one-hot encoding for the labels. $LARNet_{GloVe}$ uses the GloVe-300 text encoding for labels. $LARNet_{BERT}$ uses the BERT [16] text encoding for labels.

(BaseNet-2). Finally, we add a loss on the generated action representation for explicit modeling (LARNet-Base). The comparison is shown in Table 3 and we can observe that adding explicit supervision outperforms both the variants in all metrics.

Effectiveness of Hierarchical Motion Integration To study the effect of hierarchical recurrent motion integrator M_I on LARNet, we experimented with two different hierarchies on top of LARNet-base model. LARNet-MI-1 refers to recurrent motion integrator with single hierarchy and LARNet-MI-3 refers to a recurrent hierarchical motion integrator with three levels. A comparison of these two models is shown in Table 3. We observe that adding a three level motion integrator improves the PSNR and SSIM values as it focuses on both coarse level as well as fine level action dynamics.

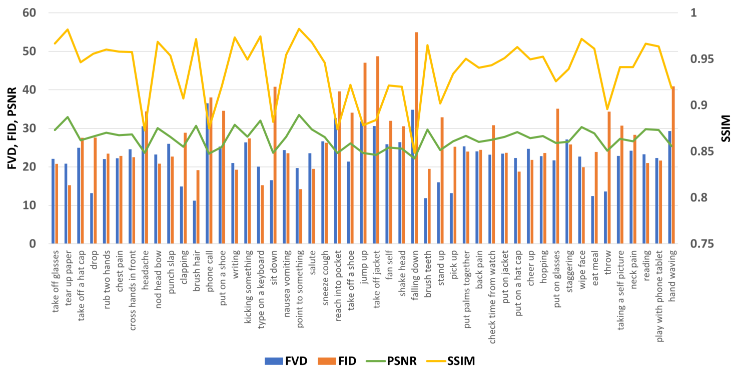

Figure 6: Per class FVD, FID, PSNR and SSIM analysis on the NTU-RGB+D dataset. The left axis represents FVD, FID and PSNR scores while the right axis represents the SSIM scores per class.

Influence of Loss Functions To further demonstrate the effect of different loss functions we add a normal adversarial loss on LARNet-MI-3 model (+L^{v}_{adv}) for synthesized video. Next, we add an adversarial loss (+[L^{v}_{adv}, L^{m}_{adv}]) on generated motion dynamics. And finally, we use a mix-adversarial loss (+[$L^{v}_{madv}, L^{m}_{adv}$]) instead of normal adversarial loss on the generated video. A comparison of these loss functions is shown in Table 3. We observe that the adversarial loss terms improves FID and FVD scores, which are indicators of realism in the videos. Further, we also observe that the proposed mix-adversarial loss outperforms the classical adversarial loss in all four evaluation metrics.

Effectiveness of different label encoding The effect of changing the label encoding on LARNet is shown in Table 4, where we compare among using one-hot encoding, word encoding from GloVe-300 [43] and word encoding from BERT [16] models. We observe that the word encoding performs better than simple one-hot encoding in all metrics. While BERT encoding has better FID score, it performs slightly worse in FVD and PSNR metrics.

4.3 Analysis and Discussion

The proposed model generates video conditioned on the action type and to illustrate its effectiveness we use the same actor to synthesize different actions. In Figure 3 (row 2), we have shown two different generated videos using the same actor. We can observe that the action is distinctly visible in all the generated videos which demonstrates the capability of LARNet to synthesize diverse action videos.

Per Class Analysis We observe the per class quantitative performance on NTU-RGB+D dataset in Figure 6. It is observed that classes with lower FVD and FID scores (lower is better) also have higher PSNR and SSIM scores (higher is better), showing a correlation of improved performance across the metrics for those classes.

Limitations and Challenges Video synthesis has been challenging as it needs an understanding of the action dynamics. Despite the recent efforts, the problem of video synthesis is still far from being solved. The proposed approach successfully generates videos with visible actions, however, modelling complex actions remains a challenge. Using significantly high computational resources have shown great improvement in this task by using large scale TPUs [12, 59]. Our training was limited to a single 24Gb GPU which we believe will scale well with the availability of higher computational resources.

5 Conclusion

In this work, we present a novel approach for generating human actions from an input image. The proposed framework predicts human actions conditioned on action semantics and utilizes a generative mechanism which estimates latent action representation. The latent action representation is explicitly learned with the help of a similarity and adversarial loss formulation. This learned latent representation is then used to generate an action video which is optimized using multiple objectives, including a novel mix-adversarial loss. We perform extensive experiments on multiple human action datasets demonstrating the effectiveness of various components of the proposed approach.

References

[1] Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Towards high resolution video generation with progressive growing of sliced wasserstein gans. *arXiv preprint arXiv:1810.02419*, 2018.

[2] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, pages 1995–2001, 2019.

[3] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. *International Conference on Learning Representations, ICLR*, 2016.

[4] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *Proceedings of the European conference on computer vision (ECCV)*, pages 119–135, 2018.

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[7] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5933–5942, 2019.

[8] Yu-Wei Chao, Jimei Yang, Brian Price, Scott Cohen, and Jia Deng. Forecasting human dynamics from static images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 548–556, 2017.

[9] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)*, pages 168–172. IEEE, 2015.

[10] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[11] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[12] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv*, pages arXiv–1907, 2019.

[13] Aidan Clark, Jeff Donahue, and Karen Simonyan. Efficient video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019.

[14] Ugur Demir, Yogesh S Rawat, and Mubarak Shah. Tinyvirat: low-resolution video action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7387–7394. IEEE, 2021.

[15] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. *arXiv preprint arXiv:1802.07687*, 2018.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[17] Bin Duan, Wei Wang, Hao Tang, Hugo Latapie, and Yan Yan. Cascade attention guided residue learning gan for cross-modal translation. *arXiv preprint arXiv:1907.01826*, 2019.

[18] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *arXiv preprint arXiv:1605.07157*, 2016.

[19] Katerina Fragkiadaki, Jonathan Huang, Alex Alemi, Sudheendra Vijayanarasimhan, Susanna Ricco, and Rahul Sukthankar. Motion prediction under multimodality with conditional stochastic networks. *arXiv preprint arXiv:1705.02082*, 2017.

[20] Jean-Yves Franceschi, Edouard Delasalles, Mickaël Chen, Sylvain Lamprier, and Patrick Gallinari. Stochastic latent residual video prediction. In *International Conference on Machine Learning*, pages 3233–3246. PMLR, 2020.

[21] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[22] Shir Gur, Sagie Benaim, and Lior Wolf. Hierarchical patch vae-gan: Generating diverse videos from a single sample. *Advances in Neural Information Processing Systems*, 33, 2020.

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017.

[24] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369. IEEE, 2010.

[25] Qiyang Hu, Adrian Waelchli, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. Video synthesis from a single image and motion stroke. *arXiv preprint arXiv:1812.01874*, 2018.

[26] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[27] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[28] Bernhard Kratzwald, Zhiwu Huang, Danda Pani Paudel, Acharya Dinesh, and Luc Van Gool. Improving video generation for multi-functional applications. *arXiv preprint arXiv:1711.11453*, 2017.

[29] Bernhard Kratzwald, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Towards an understanding of our world by ganing videos in the wild. *arXiv preprint arXiv:1711.11453*, 2017.

[30] Manoj Kumar, Mohammad Babaeizadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, and Durk Kingma. Videoflow: A conditional flow-based model for stochastic video generation. In *International Conference on Learning Representations (ICLR)*, 2020.

[31] Yong-Hoon Kwon and Min-Gyu Park. Predicting future frames using retrospective cycle gan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1811–1820, 2019.

[32] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE conference on CVPR*, 2017.

[33] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018.

[34] Xianhang Li, Junhao Zhang, Kunchang Li, Shruti Vyas, and Yogesh S Rawat. Pose-guided generative adversarial net for novel view action synthesis. *IEEE Winter Conference on Applications of Computer Vision*, 2022.

[35] Yitong Li, Martin Renqiang Min, Dinghan Shen, David E Carlson, and Lawrence Carin. Video generation from text. In *AAAI*, 2018.

[36] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6329–6338, 2019.

[37] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P Xing. Dual motion gan for future-flow embedded video prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1744–1752, 2017.

[38] Alice Lucas, Santiago Lopez-Tapia, Rafael Molina, and Aggelos K Katsaggelos. Generative adversarial networks and perceptual losses for video super-resolution. *IEEE Transactions on Image Processing*, 28(7):3312–3327, 2019.

[39] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[40] Gaurav Mittal and Baoyuan Wang. Animating face using disentangled audio representations. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 3290–3298, 2020.

[41] Katsunori Ohnishi, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Hierarchical video generation from orthogonal information: Optical flow and texture. *arXiv preprint arXiv:1711.09618*, 2017.

[42] Sunghyun Park, Kangyeol Kim, Junsoo Lee, Jaegul Choo, Joonseok Lee, Sookyung Kim, and Edward Choi. Vid-ode: Continuous-time video generation with neural ordinary differential equation. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2412–2422, 2021.

[43] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[44] Masaki Saito and Shunta Saito. Tganv2: Efficient training of large models for video generation with multiple subsampling layers. *arXiv preprint arXiv:1811.09245*, 2018.

[45] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[46] Kara Marie Schatz, Erik Quintanilla, Shruti Vyas, and Yogesh S Rawat. A recurrent transformer network for novel view action synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, 2020, Proceedings*, pages 410–426. Springer, 2020.

[47] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE, 2004.

[48] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on CVPR*, 2016.

[49] Sarah Shiraz, Krishna Regmi, Shruti Vyas, Yogesh S Rawat, and Mubarak Shah. Novel view video prediction using a dual representation. *International Conference on Image Processing*, 2021.

[50] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019.

[51] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019.

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[53] Ximeng Sun, Huijuan Xu, and Kate Saenko. A two-stream variational adversarial network for video generation. *arXiv preprint arXiv:1812.01037*, 2018.

[54] Quang Nhat Tran and Shih-Hsuan Yang. Efficient video frame interpolation using generative adversarial networks. *Applied Sciences*, 10(18):6245, 2020.

[55] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[56] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[57] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

[58] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. In *International Conference on Learning Representations (ICLR)*, 2017.

[59] Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V Le, and Honglak Lee. High fidelity video prediction with large stochastic recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 81–91, 2019.

[60] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *NeurIPS*, 2016.

[61] Shruti Vyas, Yogesh S Rawat, and Mubarak Shah. Multi-view action recognition using cross-view video prediction. In *Computer Vision–ECCV 2020: 16th European Conference, 2020, Proceedings*, pages 427–444. Springer, 2020.

[62] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3332–3341, 2017.

[63] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. *arXiv preprint arXiv:1711.11585*, 2017.

[64] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems*, pages 1144–1156, 2018.

[65] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Bryan Catanzaro, and Jan Kautz. Few-shot video-to-video synthesis. *Advances in Neural Information Processing Systems*, 32:5013–5024, 2019.

[66] Yaohui WANG, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal gan for video generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[67] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3an: Disentangling appearance and motion for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5264–5273, 2020.

[68] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[69] Jingwei Xu, Huazhe Xu, Bingbing Ni, Xiaokang Yang, and Trevor Darrell. Video prediction via example guidance. In *International Conference on Machine Learning*, pages 10628–10637. PMLR, 2020.

[70] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[71] Yufei Ye, Maneesh Singh, Abhinav Gupta, and Shubham Tulsiani. Compositional video prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10353–10362, 2019.

[72] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[73] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2248–2255, 2013.

[74] Yumeng Zhang, Gaoguo Jia, Li Chen, Mingrui Zhang, and Junhai Yong. Self-paced video data augmentation with dynamic images generated by generative adversarial networks. *arXiv preprint arXiv:1909.12929*, 2019.

[75] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9299–9306, 2019.