Im2vid: Future Video Prediction for Static Image Action Recognition

Badour Ahmad AlBahar

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Engineering

Jia-Bin Huang, Chair
A. Lynn Abbott
Pratap Tokekar

May 9, 2018
Blacksburg, Virginia

Keywords: Human Action Recognition, Static Image Action Recognition, Video Action Recognition, Future Video Prediction.

Im2vid: Future Video Prediction for Static Image Action Recognition

Badour Ahmad AlBahar

(ABSTRACT)

Static image action recognition aims at identifying the action performed in a given image. Most existing static image action recognition approaches use high-level cues present in the image such as objects, object human interaction, or human pose to better capture the action performed. Unlike images, videos have temporal information that greatly improves action recognition by resolving potential ambiguity. We propose to leverage a large amount of readily available unlabeled videos to transfer the temporal information from video domain to static image domain and hence improve static image action recognition. Specifically, We propose a video prediction model to predict the future video of a static image and use the future predicted video to improve static image action recognition. Our experimental results on four datasets validate that the idea of transferring the temporal information from videos to static images is promising, and can enhance static image action recognition performance.

Im2vid: Future Video Prediction for Static Image Action Recognition

Badour Ahmad AlBahar

(GENERAL AUDIENCE ABSTRACT)

Static image action recognition is the problem of identifying the action performed in a given image. Most existing approaches use the high-level cues present in the image like objects, object human interaction, or human pose to better capture the action performed. Unlike images, videos have temporal information that greatly improves action recognition. Looking at a static image of a man who is about to sit on a chair might be misunderstood as an image of a man who is standing from the chair. Because of the temporal information in videos, such ambiguity is not present. To transfer the temporal information and action features from video domain to static image domain and hence improve static image action recognition, we propose a model that learns a mapping from a static image to its future video by looking at a large number of existing images and their future videos. We then use this model to predict the future video of a static image to improve its action recognition. Our experimental results on four datasets show that the idea of transferring the temporal information from videos to static images is promising, and can enhance static image action recognition performance.

Dedication

To my mother, who has always been a continuous source of love and support. Without you, I would have never become the woman, mother, and researcher I am proud to be today.

Acknowledgments

I would like to thank my advisor, Jia-Bin Huang, for his guidance and support. I appreciate all the time and effort he contributed to this journey. Thank you for introducing me to this fascinating realm of research. I would also like to thank Kuwait University for choosing me among many highly qualified individuals for their generous scholarship. Finally, I would like to thank my family whom I am truly grateful to have in my life, for their love and endless support and encouragement. To my parents, thank you for being continuously supportive of me and my goals in life. To my brothers and sisters, thank you for believing in me and for being there when I need you most. To my husband and daughter Al-Zain, who has been my little bundle of joy and happiness throughout this wonderful journey, thank you for your love and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Problem Definition Human action recognition is the problem of identifying the action performed by a human in a given data sample. It has been an active area of research in Computer Vision for a long time. In which, many studies have tackled video action recognition. However, recently, many studies in static image action recognition have emerged. In contrast to video action recognition, static image action recognition lacks the temporal information of videos, which makes the problem even more challenging. Looking at a static image of a human performing an action, it might be difficult to identify what action is being performed. For example, figure 1.1 shows a static image of a human performing an action. The action is ambiguous as it could be either sitting down or standing up. However, videos do not have such ambiguity and one can easily eliminate this confusion. Therefore, we aim to improve static image action recognition by leveraging the future video of the static image and make use of the temporal information static images lack.

Given a static image, we would like to identify the action performed by the human by leveraging the temporal information of the future video of this static image. Recognizing actions in static images can help in image annotation, image organization by human action categories, and image search and retrieval.

Previous Approaches Various high-level cues like human body, body parts, action-related objects, human object interaction, and the scene context have been used to aid

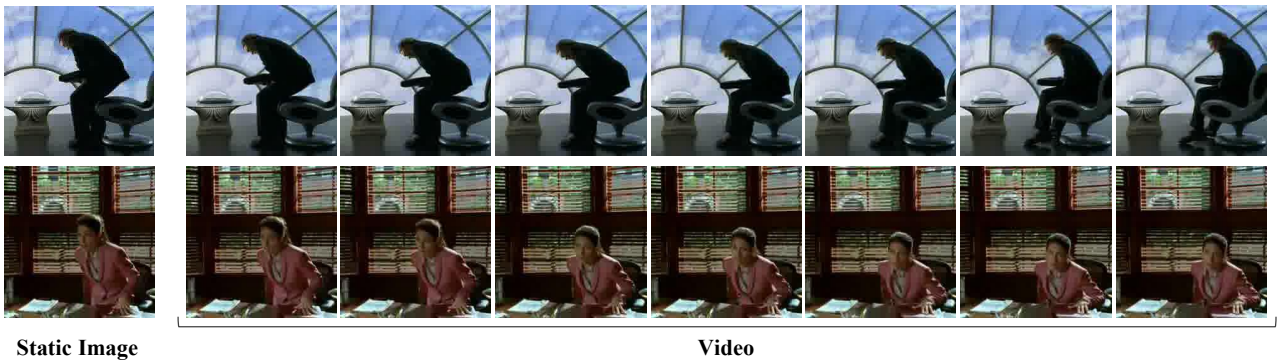

Figure 1.1: Looking at the static image alone, it is difficult to identify the action performed by the human. Is the human sitting down or standing up? However, looking at the future video on the static image eliminates this ambiguity.

in action recognition [15]. Moreover, to enhance static image action recognition results, videos have also been used to augment the training data [5], and to predict the motion of the static image to be used along with the input image in a two stream action recognition model [11].

Our Approach Despite the recent progress in static image action recognition, it is still a recent research area where there is room for improvement. In this paper, we propose to leverage future videos of static images and make use of the missing temporal information to aid in static image action recognition. We propose im2vid, a future video prediction model that predicts the future frames, given a static image. We then use the predicted future video to improve static image action recognition.

Novelty of Our Approach Because static images can be ambiguous, predicting the future video of a given static image can eliminate the ambiguity and the temporal information of the predicted video can aid in static image action recognition.

Results We show that our proposed future video prediction model, im2vid, when trained on UCF101 [34] and tested on Thumos15 [13], generates videos that are significantly of higher visual quality than the current state of the art iVGAN [19]. Moreover, we show that our proposed model, im2vid, can be transfered to other datasets and used to predict future videos of static images to aid in static image action recognition.

Contributions We leverage the large amount of available videos to predict the future video of a given static image and use the predicted video and its temporal information to improve the performance of static image action recognition.

Chapter 2

Literature Review

2.1 Future Video Prediction

There has been various work on future video prediction. Some do pixel level predictions, while others do pixel motion prediction to reduce the dimensionality of the prediction space.

Pixel Level Prediction Due to the high dimensionality of the pixel space, directly predicting the pixels of the RGB future video is very difficult and often results in blurry predictions. Given a number of frames, [23, 24, 28, 47] predict a pixel-level next future frame. In order to predict more future frames, the prediction can be done recursively such that the predicted frames are used as input, which leads to error accumulation. Other works leverage a high-level structure that can enhance the future prediction. [39] decomposes the input into motion and content features to be decoded to generate the future prediction. On the other hand, [8, 40, 45] leverage pose to aid in a better future prediction.

Vondrick et al. [42] predicts 32 future frames from a single input image using a generative adversarial network that separates the generation of the foreground form the background, which is assumed to be constant, however, Kratzwald et al. [19], predicts 32 future frames from a single input image using a one-stream video generation model and does not impose such static background assumption.

Pixel Motion Prediction A better approach to future prediction instead of directly predicting the pixel values of the RGB future video, is to predict the motion of those pixels [4, 10, 37, 41, 44] and generate videos that are less blurry and are of better quality.

Almost all of the aforementioned work do future prediction given a clip of video, however, to the best of our knowledge [4, 19, 42, 44, 47] are the only work that do future prediction from a single image. [4, 19, 42] are category specific such that they train their proposed model on a single specific scene and test on that scene. Using multiple variant scenes usually results in poorer performance. In this work, we generalize to multi-scene learning and focus on human action videos. Like [19] we don't impose the static background assumption, however, instead of directly predicting the pixels of the video we predict the motion of those pixels.

2.2 Human Action Recognition

Human action recognition is the problem of identifying the human action performed in a given input.

Video Action Recognition Video action recognition has been extensively studied. Early approaches independently processed each frame and combined the per frame extracted features to classify the actions paying no attention to the temporal information [27, 46]. A better approach to exploit the long term temporal information, is to employ recurrent neural networks, like long short term memory (LSTM) [9, 26]. An even better approach is to use 3D convolutions instead of performing 2D convolutions, which were proposed by Ji et al. [16] to extract both spatial and temporal features. Several approaches using 3D ConvNets for video action recognition were later proposed [35, 36, 38]. A disadvantage of 3D ConvNets is their inability to leverage ImageNet pretraining, to combat this issue, inflated 3D ConvNets

were proposed by Carreira and Zisserman [2], which add a temporal dimension to the 2D architecture by inflating its filters and pooling kernels.

Static Image Action Recognition Recently, static image action recognition has been an active area of research. Unlike video action recognition, there are no temporal information to be used to aid in action recognition. Therefore, various high-level cues like human body [33, 49], body parts [48, 50], objects [32, 52], human-object interaction [7, 21], and context or scene [22, 51] are used. Such high-level cues are characterized by various low-level features and used in a variety of action learning methods to aid in action recognition. For more information, see the survey by Guo and Lai [15].

Moreover, there are approaches that leverage videos to improve static image action recognition. Chen and Grauman [5] uses video to learn human motion in order to augment the training data. By learning the motion from video, this work uses the motion with the input static image to generate more images corresponding to the same action label. Gao et al. [11] however, proposes im2flow, an encoder-decoder network to predict the future flow implied by the static image. They then make use of both streams, the static image and the predicted flow, to enhance static image action recognition. Similarly, we leverage unlabeled videos for static image action recognition. However, instead of augmenting the training data or predicting the implied flow of the given image, we predict the future video to make use of the temporal information videos have for static image action recognition.

2.3 Generative Models

Generative Adversarial Networks (GANs) GANs [12] consist of two different networks that compete against each other. The generator, which takes an input and transforms

it into a randomly selected sample from the data distribution it is learning, and the discriminator, which looks at both the training samples and the generated samples and tries to accurately distinguish between the two. As the generator learns to synthesize more realistic samples, the discriminator learns to accurately classify the real and fake samples. The goal is for the generator to be able to fool the discriminator and produce data that is very realistic and is indistinguishable from real data.

The discriminator is trained to maximize the probability of assigning the correct label to both the training samples and the data synthesized by the generator while the generator is simultaneously trained to minimize this probability. In other words, the discriminator and the generator are playing a two-player minimax game with the value function:

$$\min_G \max_D V(D,G) = \mathbf{E}_{x\sim P_x}[logD(x)] + \mathbf{E}_{\tilde{x}\sim P_g}[log(1-D(\tilde{x}))] \tag{2.1}$$

where G is the generator, D is the discriminator, P_x is the data distribution, and P_g is the generator distribution. In practice, Goodfellow et al. [12] alternates between k steps of optimizing D and one step of optimizing G to avoid optimizing D to completion in the inner loop of training, as it would result in over fitting on finite datasets. Moreover, equation 2.1 may not provide sufficient gradient for G to learn well. Because when G is in early learning, D can reject samples with high confidence. In this case, $log(1-D(\tilde{x}))$ saturates. Therefore, instead of minimizing this log, G is trained to maximize $logD(\tilde{x})$ which provides stronger gradients.

Despite GANs' great success in realistic data generation, they suffer from instability, which makes them hard to train as the discriminator must be synchronized well with the generator during training. Many approaches have been proposed to address the limitations of GANs. Radford et al. [29] uses several layers of deconvolution and also makes use of batch

normalization to learn faster and become more stable. Moreover, since max pooling is not invertible, this work suggests not using any inverse of the pooling operation.

Another work by Salimans et al. [31] sheds light on mode collapse, one of the biggest problems in training GANs, where the generator starts to produce several copies of exactly the same output. This happens when the generator is fully optimized while the discriminator is held constant. The discriminator will describe a single region in space as being the most likely point to be real rather than fake, and all generated results will then be mapped to this point. As a solution, minibatch GAN is proposed, where the discriminator looks at an entire minibatch of samples at a time rather than looking at a single sample and the minibatch as a whole has to look realistic and has to have the correct amount of space in between different samples.

Another work that addresses GANs limitations is by Arjovsky et al. [1], which proposes Wasserstein GAN (WGAN), which significantly improves mode collapse. In addition, the training of the discriminator and generator no longer needs to be carefully balanced. Unlike GANs, which minimize the Jensen-Shannon (JS) divergence 2.2 between the data distribution P_x and the generator distribution P_g, WGANs minimize the Earth Mover's (EM) distance or Wasserstein-1 2.3.

Jensen-Shannon (JS) Divergence

$$JS(P_x, P_g) = KL(P_x||P_m) + KL(P_g||P_m) \tag{2.2}$$

where P_m is $\frac{P_x+P_g}{2}$, and KL is the Kullback-Leibler divergence given by:

$$KL(P_x||P_g) = \int_X log(\frac{P_x}{P_g})P_x d\mu$$

where μ is any measure on X and both distributions are assumed to be continuous with respect to μ.

Earth-Mover (EM) distance or Wasserstein-1

$$W(P_x, P_g) = \inf_{\gamma \in \Pi(P_x, P_g)} E_{(x,y)\sim\gamma}[||x - y||] \tag{2.3}$$

where $\Pi(P_x, P_g)$ is the set of all joint distributions $\gamma(x, y)$, which denotes how much to transport from x to y to make P_x closer to P_g. The final computed distance reflects the the minimum cost of the transform of P_x into P_g.

WGAN Because EM's distance has better properties than JS divergence when optimized, and because it is intractable to compute the infimum over $\Pi(P_x, P_g)$, Arjovsky et al. [1] proposes to use Kantorovich-Rubinstein duality:

$$W(P_x, P_g) = \sup_{||f||_L \leq 1} E_{x\sim P_x}[f(x)] - E_{x\sim P_g}[f(x)]$$

where the supremum is over all the 1-Lipschitz functions f. Therefore, the WGAN minimizes the EM's distance using the value function:

$$\min_G \max_C V(C, G) = \mathbf{E}_{x\sim P_x}[C(x)] + \mathbf{E}_{\tilde{x}\sim P_g}[C(\tilde{x})] \tag{2.4}$$

instead of a discriminator (D), a critic (C) is used because no classification is being performed. C once again denotes the set of 1-Lipschitz functions. This value function of WGAN 2.4 compared to the value function of GAN 2.1, improves the stability of learning and correlates well with the generators convergence and sample quality.

To enforce this Lipschitz constraint, weight clipping is performed. Such that the weights of the critic are forced to lie within a range $[-w, w]$. This enforces a k-Lipschitz function, were k depends on the clipping limit w as well as the critic's architecture. The authors note that the weight clipping is far from ideal. When the clipping parameter is large, we suffer from slow convergence. On the other hand, when the clipping parameter is small, we suffer from vanishing gradients.

Gradient Penalty To better enforce the Lipschitz constraint, Gulrajani et al. [14] uses a gradient penalty. Knowing the data distribution P_x and the generator distribution P_g, $P_{\hat{x}}$ is defined implicitly as uniformly sampling pairs from these distributions along a straight line. The gradient penalty enforces the norm of the critic's gradient $\nabla_{\hat{x}} C(\hat{x})$ over $\hat{x} \sim P_{\hat{x}}$ to be a unit gradient norm with a penalty coefficient λ:

$$\min_G \max_C V(C, G) = E_{\tilde{x} \sim P_g}[C(\tilde{x})] - E_{x \sim P_x}[C(x)] + \lambda E_{\hat{x} \sim P_{\hat{x}}}[(||\nabla_{\hat{x}} C(\hat{x})||_2 - 1)^2] \tag{2.5}$$

Spectral Normalization A recent approach to further improve the stability of GANs is spectral normalization, which is a weight normalization method proposed by Miyato et al. [25]. This approach is more stable with lower computational cost than WGAN with gradient penalty. This work forces the Lipschitz constraint by normalizing the spectral norm of the discriminator's weight matrix (W):

$$\hat{W} = \bar{W}_{SN}(W) = \frac{W}{\sigma(W)} \tag{2.6}$$

where $\sigma(W)$ is the spectral norm of the W, which is equal to the largest singular value of W. Therefore, equation 2.6, forces $\sigma(\hat{W})$ to be equal to 1. Having, $\sigma(\hat{W}) = 1$ sets an upper bound on the discriminator such that $||f||_{lip} \leq \prod_{l=1}^{L+1} \sigma(\hat{W}^l) \leq 1$. Hence, the Lipschitz

constraint is satisfied.

In this work, we use WGAN with gradient penalty and note that spectral normalization can be used as future work to improve the the stability of the generative model and the sample quality.

Chapter 3

Method

In this work, we aim to improve static image action recognition by leveraging the large amount of available unlabeled videos. In contrast to videos, the lack of temporal information in static images makes the problem of static image action recognition challenging. We propose im2vid, a future prediction model that predicts the future video of a static image to leverage the predicted temporal information to improve static image action recognition. See figure 3.1. Our model consists of two main parts. First, the future prediction model takes a static image as input and predicts a future video. Second, a video action recognition model takes the predicted video and outputs an action class. Below we discuss the two models in detail.

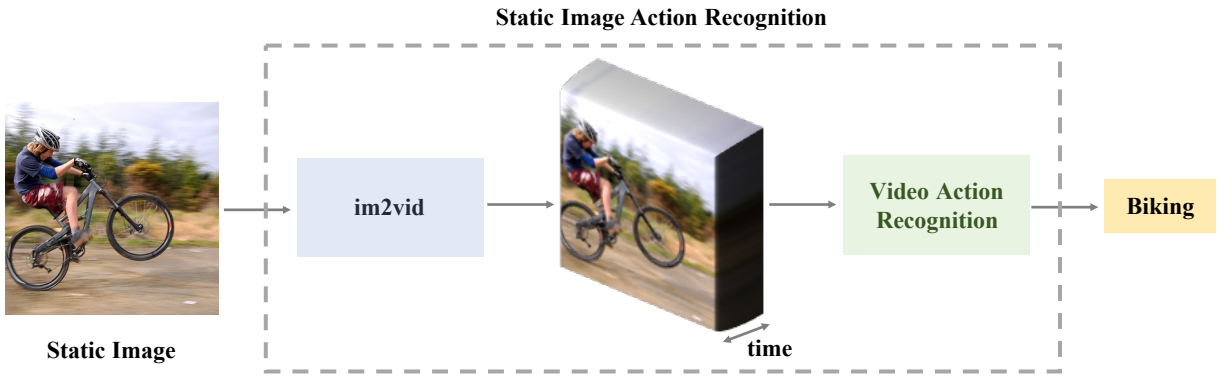

Figure 3.1: To improve static image action recognition, we propose im2vid to predict the future video of a given static image to be used to perform video action recognition and predict the action class of the given static image.

3.1 Static Image Action Recognition

Intuition Static image action recognition is the problem of identifying the action performed in a given static image. Due to the lack of temporal information, identifying the action of a static image is quite challenging. Having a large amount of unlabeled videos, we can leverage their temporal information to aid in static image action recognition. Like Chen and Grauman [5] which uses unlabeled videos to augment the static image training data, and Gao et al. [11] which uses unlabeled videos to hallucinate the motion of a static image to be used in a two-stream action recognition model, we propose to leverage unlabeled videos to predict the future video of a given static image to enhance static image action recognition.

Network Architecture To identify the action of a given static image, we use im2vid to generate the future video of a static image, and then use the I3D video action recognition model [3] to get the action class of the generated video as shown in figure 3.2.

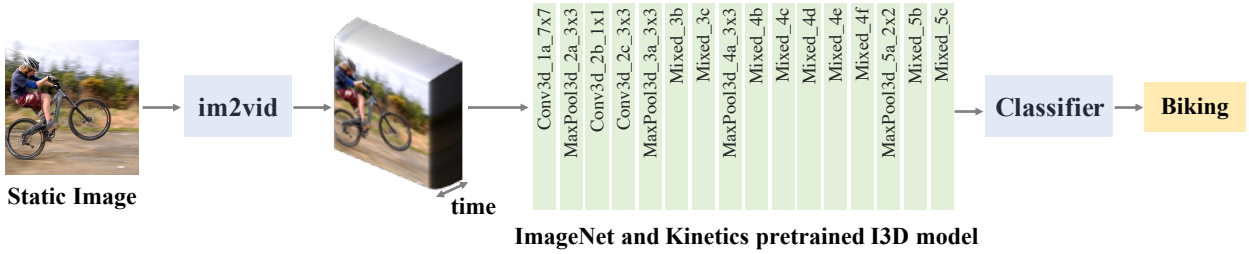

Figure 3.2: Static image action recognition network architecture.

Training We freeze the weights of the I3D model shown in figure 3.2 to extract the video features and train a classifier for the action classes of the target dataset.

Inference To validate our approach, we set an upper bound by using the ground truth video for action recognition for video datasets, while for static image datasets, we duplicate

the static image to generate a mock video.

3.2 Im2vid: Future Video Prediction

Intuition We propose im2vid, a video prediction model that predicts a future video given a static image. Unlike [4, 19, 42] which are scene specific, we generalize to multiple scenes and train on different action classes. Moreover, like [19], we do not impose a static background assumption and hence do not require stabilized videos for training.

RGB from Flow Because we know that future frames will have similar pixels to the given static image but at different position, instead of re-predicting those pixels we reduce the dimensionality of the output space, and predict the motion of the pixels instead of directly predicting their values. This will reduce the blurriness and improve the quality of the final prediction. We also note that to avoid error accumulation over time, we do not do recursive prediction of the motion, such that we can independently predict any frame at time t in the future. That is, for any time t in the future the predicted frame does not depend on any prior predicted frames at time $< t$.

Warping from Flow Using the motion of the pixels, we can warp the pixels of the given image to get the future frames, however, because some pixels move in and out of the frames, those future frames will have missing regions which we call holes. To reflect those holes, we use the predicted motion to introduce a mask to know which pixels have moved within the frame and which pixels have moved out or into the frame.

Completion Network To fill up the holes, we introduce a completion network that directly generates the pixels of the missing regions that haven't been warped by the motion.

Network Architecture Our future video prediction model takes a human action static image as input and maps it to a future video as shown in figure 3.3. To obtain the future

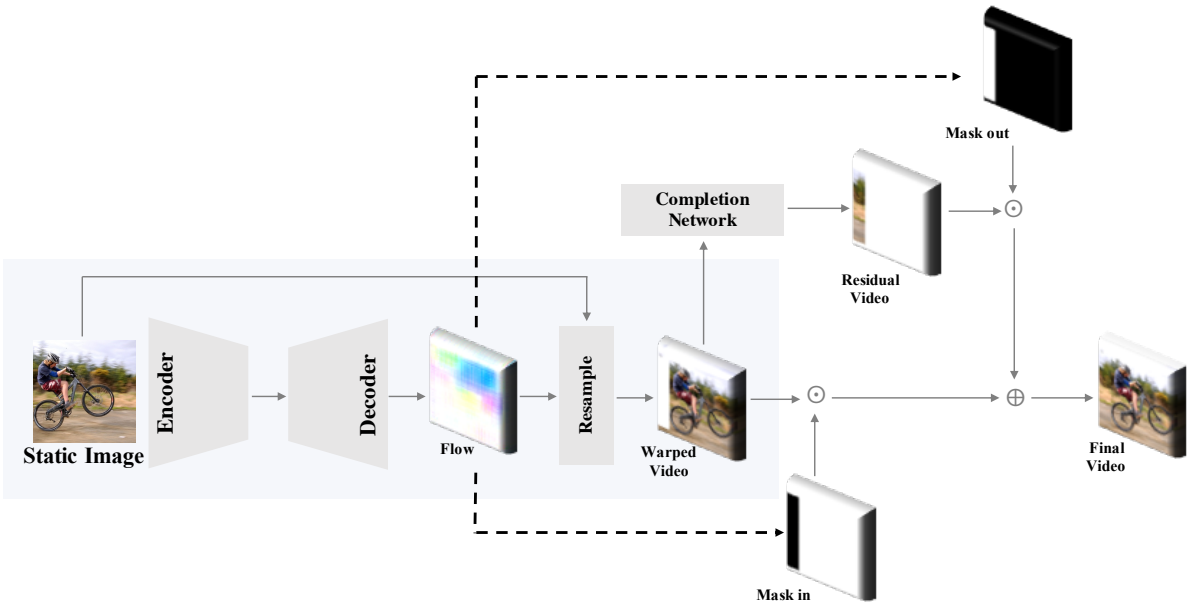

Figure 3.3: Im2vid network architecture.

video from the static image, instead of predicting the future in pixel space directly we predict the motion of the pixels to get a less blurry output. To do so, given the input image we use an encoder-decoder network to generate the future flow and use this flow to warp the input image into a warped future video. Because pixels move in and out of the frames, the warped video will have missing regions which we call holes. We propose a binary mask to reflect such holes. To fill up those holes, we propose CompNet, an encoder-decoder completion network with skip connections. We also propose CompNet + input a variation of CompNet that takes the static image as an additional input. Both models are shown in figure 3.4. Both completion networks take the warped video as input and generate a residual video containing

patches to fill up the holes. Such that the final video $\hat{V}$:

$$\hat{V} = WV \cdot M_{in} + RV \cdot M_{out}$$

for the warped video WV, residual video RV, mask reflecting the warped regions M_{in}, and mask reflecting the holes $M_{out} = 1 - M_{in}$.

Training During training, the model is given a static image and its corresponding ground truth future with the first frame matching the given image. Given the static image, our model generates a warped video minimizing the $L2$ and the adversarial loss between the warped video and the ground truth video. For the adversarial loss, we follow the approach of [19], which uses the WGAN framework with a gradient penalty originally proposed by [14] with a penalty coefficient λ as:

$$L_{adv} = E_{\tilde{x}\sim P_g}[C(\tilde{x})] - E_{x\sim P_x}[C(x)] + \lambda E_{\hat{x}\sim P_{\hat{x}}}[(||\nabla_{\hat{x}} C(\hat{x})||_2 - 1)^2] \tag{3.1}$$

where P_x is the data distribution, P_g is the generator distribution, and $P_{\hat{x}}$ is implicitly defined as uniformly sampling pairs from these distributions along a straight line. The L2 loss enforces the warped video to look like the ground truth video, while the adversarial loss enforces the warped video to look as realistic as the ground truth samples. We note that the absence of this deep supervision on the warped video produced non-satisfactory results. To generate the final video, we minimize the $L2$ and the adversarial loss as shown in equation 3.1 between the final video and the ground truth video, forcing the model to learn a final video that looks realistic and looks like the ground truth video.

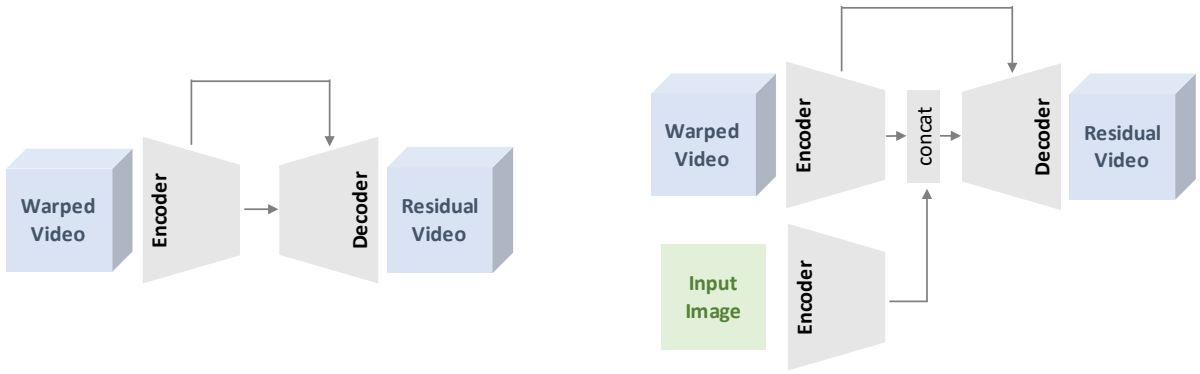

Figure 3.4: The model on the left is CompNet, which is an encoder-decoder network with skip connection. A variation of CompNet, is compNet + input shown on the right, which takes the static image as input.

Inference To validate our approach, we compare the visual quality of the generated video with iVGAN [19] and we test the effect of the generated video on static image action recognition performance.

Chapter 4

Experimental Results

We evaluate our work for both future video prediction and static image action recognition.

4.1 Im2vid: Future Video Prediction

Datasets We train our future prediction model on UCF101 [34] shown in figure 4.1. We use all $13,320$ videos of 101 different actions by dividing each video into 32 frames per clip. If the video has a number of frames that are not a multiple of 32, such that the last clip would have less than 32 frames, we complete this final clip by using frames from its preceding clip. We finally end up with $81,624$ total number of clips. To evaluate our future prediction model, we use the test data of Thumos15 dataset [13]. We extract 10 clips of 32 frames from each of the $5,613$ videos, if the video is long enough, and hence end up with a total of $55,035$ clips of 32 frames.

Evaluation metrics We evaluate the visual quality of our predicted future video and compare the results with iVGAN's prediction [19]. Similar to the work proposed in this area, we use two different metrics to evaluate the visual quality, PSNR and inception score. Inception score was found to have high correlation with human annotators when evaluated on a large enough dataset [30]. We compute both metrics by treating the predicted frames as independent images.

Figure 4.1: The 101 human action classes of UCF101 video dataset.

Implementation details We implement our work in TensorFlow. Like iVGAN [19], we use a video of 32 frames with shape $32 \times 64 \times 64 \times 3$ such that the first frame is the input image and the whole 32 frames are the final predicted video. To get a 64×64 input image, we crop the center square of the original input and then resize to 64×64. We use Adam optimizer with $beta1$ as 0.5 and $beta2$ as 0.999. Moreover, we use a batch size of 32 with learning rate 0.0001 and train for $150k$ steps with 4 iteration for each discriminator. We also use default value of the penalty coefficient $\lambda = 10$. We trained both our work and iVGAN

[19] on UCF101 [34] and we use the same implementation parameters for both.

Quantitative evaluation Table 4.1 shows the visual quality results of our work when using the completion networks CompNet and CompNet + input compared to iVGAN [19] using the inception score and PSNR metrics. Results show that our method significantly outperforms iVGAN.

Table 4.1: Visual quality results of our work and iVGAN [19] on Thumos15 dataset [13].

Method	Inception Score Ground Truth $= 9.629 \pm 0.305$ higher is better	PSNR higher is better
iVGAN [19]	2.188 ± 0.015	12.02
Our Work		
CompNet	8.445 ± 0.193	19.08
CompNet + input	7.789 ± 0.149	18.56

Qualitative evaluation We show the qualitative evaluation of our work when using the completion networks CompNet and CompNet + input and iVGAN on Thumos15 dataset in figure 4.2. Results show that our work greatly outperforms iVGAN and produces sharper non-blurry results.

4.2 Static Image Action Recognition

Datasets To perform static image action recognition, we use our future prediction model, im2vid, to predict the future video of a given static image and use the predicted future to perform video action recognition using [3]. We evaluate our approach on two static image datasets, Willow action [6] shown in figure 4.3 and stanford10 [5] a subset of stanford40 [52].

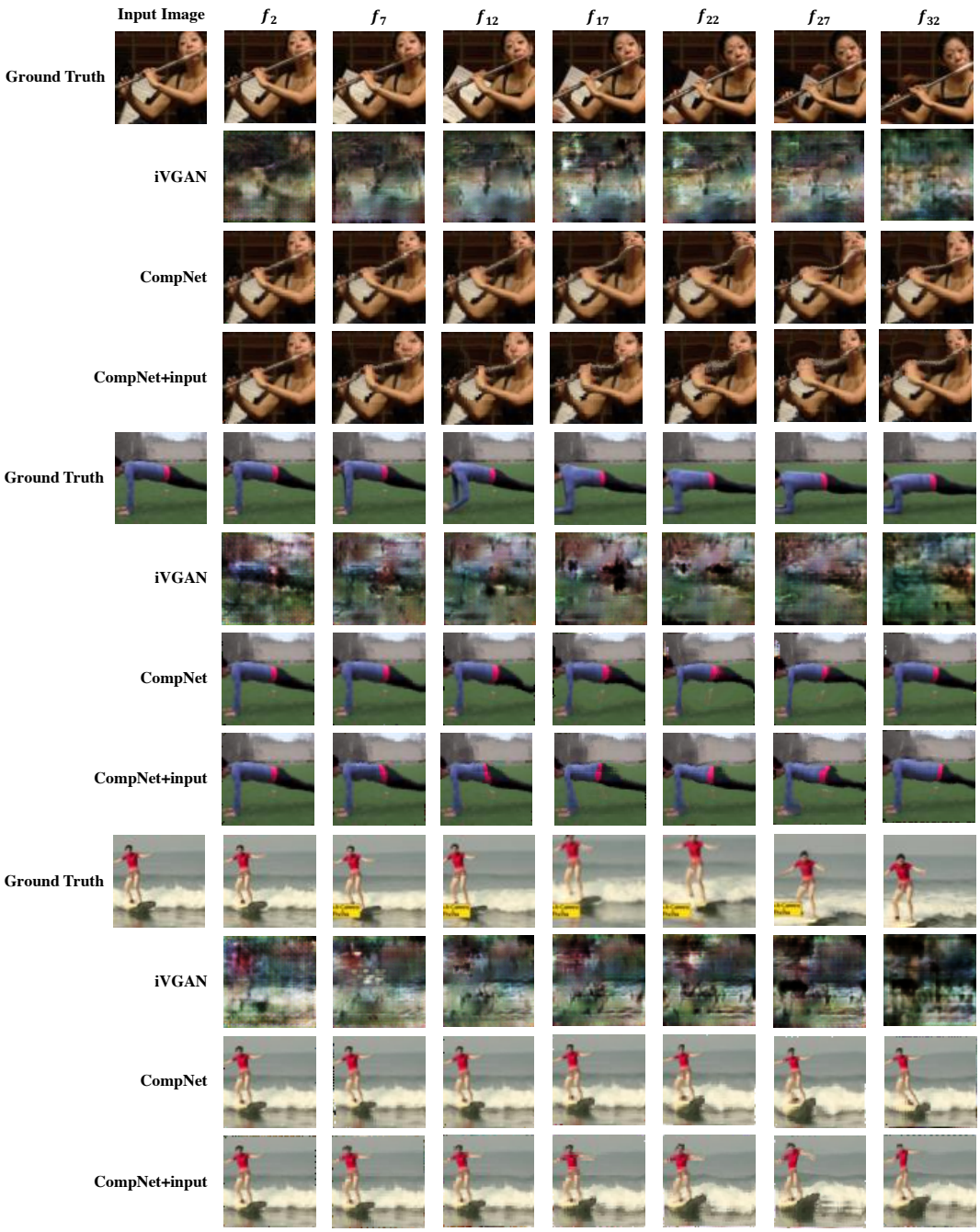

Figure 4.2: Qualitative evaluation on Thumos15 [13]. The first row shows the ground truth data, second row shows iVGAN's [19] prediction, third row shows our work with CompNet, and the final fourth row shows our work with CompNet + input. On the other hand, the first column is the input image, and the rest of the columns represent different time step future frames of the input image.

Willow has 7 action classes with 911 images, 427 are used for training, while the remainder 484 are used for testing. On the other hand, stanford10 has 10 classes which are a subset of stanford40 classes with 1,000 train images and 1,672 test images.

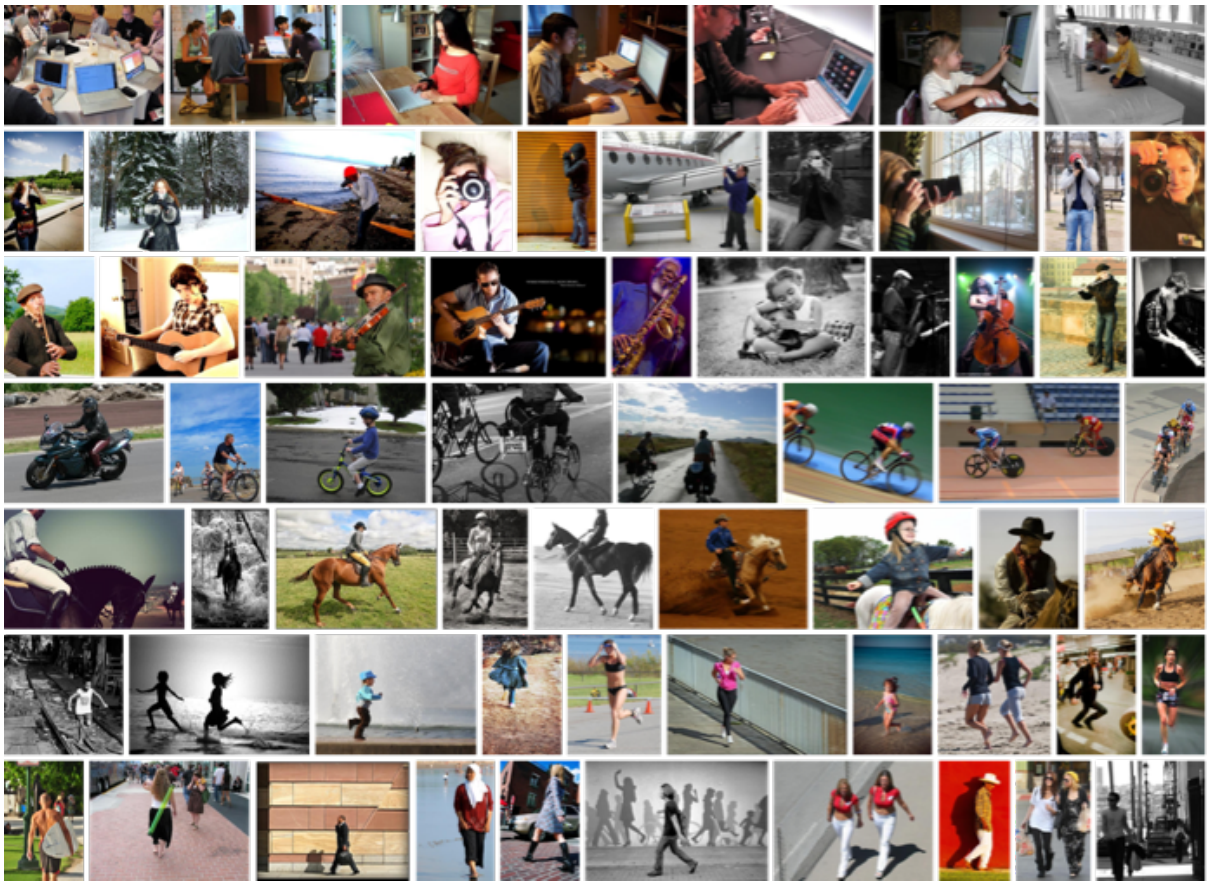

Figure 4.3: This is Willow dataset. A dataset for human action recognition in static images. Each row represents one of the 7 action classes which are Interacting with computer, Photographing, Playing Instrument, Riding Bike, Riding Horse, Running, and Walking.

Moreover, we evaluate our approach on both HMDB51 [20] shown in figure 4.4 and Penn-Action [53]. However, to convert those video datasets to static images, we take the center frame of each video following the approach of [11]. We also use the reported testing and training splits; therefore, we end up with 3,570 training images and 1,530 testing images for HMDB-static and 1,258 training images and 1068 testing images for PennAction-static.

Figure 4.4: The 51 human action classes of HMDB51 video dataset.

Evaluation metrics We evaluate the static image action recognition accuracy of our proposed approach and compare the results with im2flow [11]. As an upper bound, we show the recognition accuracy of the ground truth 32 frame video of HMDB-static and PennAction-static starting at the center frame. In addition, we duplicate the static image of Willow and stanford10 32 times to mock a still video of 32 frames and report the action recognition of this mock video.

Implementation details We implement our work in TensorFlow. To perform static image action recognition, we input the $64 \times 64 \times 3$ image into our future prediction model to generate a future video of shape $32 \times 64 \times 64 \times 3$. We then upsample to $32 \times 224 \times 224 \times 3$ to use the video action recognition model proposed by [3] to classify the action of the generated video. To use the I3D model, we freeze the ImageNet and Kinetics pre-trained weights and train a classifier for the different classes of each dataset to perform action recognition. We train the classifier using 224×224 frames for 25 epochs with a batch size of 8. We use a learning rate of 0.001 for all datasets except for HMDB51, for which we use a learning rate of 0.0001. For fair comparison of the video datasets upper bound and the static dataset mock video with our predicted video, we crop the center of each frame and downsample to 64×64 then we

upsample again to 224×224.

Quantitative evaluation Due to the randomness of the initialized classifier variables, we compute the average of 5 accuracies and report their mean and variance. Table 4.3 shows the action recognition accuracy of our work when using CompNet and CompNet + input compared to recognition results of the models shown in table 4.2 which are reported by Gao et al. [11].

Table 4.2: The methods whose action recognition results are reported by Gao et al. [11].

Method	Action Recognition Model Trained On
Single Stream	
Appearance	Static images only.
Motion (Walker et al. [43])	Optical flow computed by Walker et al. [43].
Motion (Gao et al. [11] on UCF)	Flow computed by im2flow [11] trained on UCF.
Motion (Gao et al. [11] on HMDB)	Flow computed by im2flow [11] trained on HMDB.
Motion (Gao et al. [11] on UCF+HMDB)	Flow computed by im2flow [11] trained on both UCF and HMDB.
Ground-truth Motion	Ground truth optical flow computed from ground truth videos.
Two-Stream	
Appearance + Appearance	Two separate appearance streams
Appearance + Motion (Walker et al. [43])	Appearance stream and optical flow stream computed by Walker et al. [43].
Appearance + Motion (Gao et al. [11] on UCF)	Appearance stream and flow stream computed by im2flow [11] when trained on UCF.
Appearance + Motion (Gao et al. [11] on UCF on HMDB)	Appearance stream and flow stream computed by im2flow [11] when trained on HMDB.
Appearance + Motion (Gao et al. [11] on UCF on UCF+HMDB)	Appearance stream and flow stream computed by im2flow [11] when trained on both UCF and HMDB.
Appearance + Ground-truth Motion [11]	Appearance stream and ground truth optical flow stream computed from ground truth videos.

Table 4.3: Action recognition results of our work and im2flow [11] on HMDB-static [20], PennAction-static [53], Willow [6], and stanford10 [5].

Method	HMDB-static	PennAction-static	Willow	Stanford10
Appearance	35.1	73.1	65.1	81.3
Motion (Walker et al. [43])	4.96	21.2	18.8	19.0
Motion (Gao et al. [11] on UCF)	13.9	51.0	35.7	46.4
Motion (Gao et al. [11] on HMDB)	-	42.4	30.6	42.4
Motion (Gao et al. [11] on UCF+HMDB)	-	51.5	35.9	48.4
Ground-truth Motion [11]	20.0	52.4	-	-
Appearance + Appearance	35.5	73.4	65.8	81.3
Appearance + Motion (Walker et al. [43])	35.9	73.1	65.9	81.5
Appearance + Motion (Gao et al. [11] on UCF)	**37.1**	74.5	67.4	82.1
Appearance + Motion (Gao et al. [11] on UCF on HMDB)	-	74.3	67.1	81.9
Appearance + Motion (Gao et al. [11] on UCF on UCF+HMDB)	-	**74.5**	**67.5**	**82.3**
Appearance + Ground-truth Motion [11]	39.5	77.4	-	-
Our Work				
CompNet	36.458 ± 0.001	77.772 ± 0.009	71.198 ± 0.009	81.435 ± 0.004
CompNet + input	35.621 ± 0.004	76.292 ± 0.007	71.694 ± 0.008	81.711 ± 0.004
Ground-truth Video	61.386 ± 0.004	95.412 ± 0.001	-	-
Mock Video	-	-	70.826 ± 0.004	81.172 ± 0.004

Qualitative evaluation We show an action recognition sample in figure 4.5. The figure shows the static input image, the predicted video, and the final predicted action class.

Figure 4.5: The qualitative results of static image action recognition.

Effect of Resolution We note that using the higher resolution 224×224 instead of the lower resolution, where we crop the center of each frame and downsample to 64×64 then we upsample again to 224×224 gives different accuracies. For the case of the ground truth video dataset, the original resolution's action recognition accuracy results are slightly better than the upsampled. However, for the case of the static image replication mock video, original results are significantly better than the upsampled. To show the effect of the resolution, table 4.4 shows the accuracy of the original $32 \times 224 \times 224 \times 3$ compared to $32 \times 224 \times 224 \times 3$

upsampled from $32 \times 64 \times 64 \times 3$.

Table 4.4: The effect of resolution on static image action recognition performance.

Input	HMDB-static	PennAction-static	Willow	Stanford10
$224 \times 224 \times 3$	63.765 ± 0.005	96.629 ± 0.001	83.843 ± 0.005	90.885 ± 0.002
$64 \times 64 \times 3 \rightarrow 224 \times 224 \times 3$	61.386 ± 0.004	95.412 ± 0.001	70.826 ± 0.004	81.172 ± 0.004

Chapter 5

Discussion

Im2vid: Future Video Prediction Both visual quality quantitative results (see table 4.1) and qualitative results (see figure 4.2) of our model with both variations of the completion network show significant improvement compared to iVGAN by Kratzwald et al. [19] the current state of the art. Both our work and iVGAN do not impose a static background assumption, however, our work does pixel motion prediction instead of direct pixel prediction. Unlike iVGAN, predicting the motion of the pixels enables our model to generalize to multiple scenes and actions. Moreover, unlike what we expected, the visual quality quantitative results in table 4.1 show that the CompNet completion network has better visual quality than CompNet + input.

Static Image Action Recognition Table 4.3 shows that using im2vid output video for static image action recognition improved the recognition performance compared to the state of the art for both PennAction and Willow datasets. However, no such improvement can be seen for HMDB and stanford10. In fact, the recognition results of HMDB and PennAction are very far away from the ground truth results, while the recognition results of Willow and stanford10 barely exceed the mock video baseline. Regardless of the results, it can be clearly seen that the ground truth video action recognition performance significantly exceeds the current state of the art, reinforcing the fact that our idea of leveraging the future video to improve static image action recognition performance is very promising.

Chapter 6

Future Work

Intuition As a future work to further enhance our work, we propose to impose a video action recognition loss on our im2vid prediction model. Having a video action recognition loss enables our model to transfer the action related features from the large amount of existing unlabeled videos to static images and therefore enhance and improve the static image action recognition performance. Moreover, we propose to use spectral normalization proposed by Miyato et al. [25] instead of WGAN with gradient penalty to improve the stability of the generative model and enhance the generation quality.

Training We propose to train our proposed model im2vid like before, but this time we can introduce an additional I3D [3] perceptual loss. Perceptual losses [17] minimize the differences between several high-level feature representations extracted from pretrained networks. In our work, we can minimize the loss between four or more of the 17 different feature representations of the final video and the ground truth video extracted from the ImageNet and kinetics pretrained I3D network as shown in figure 6.1. We do this to force our model to learn action related features to enhance and improve the action recognition results.

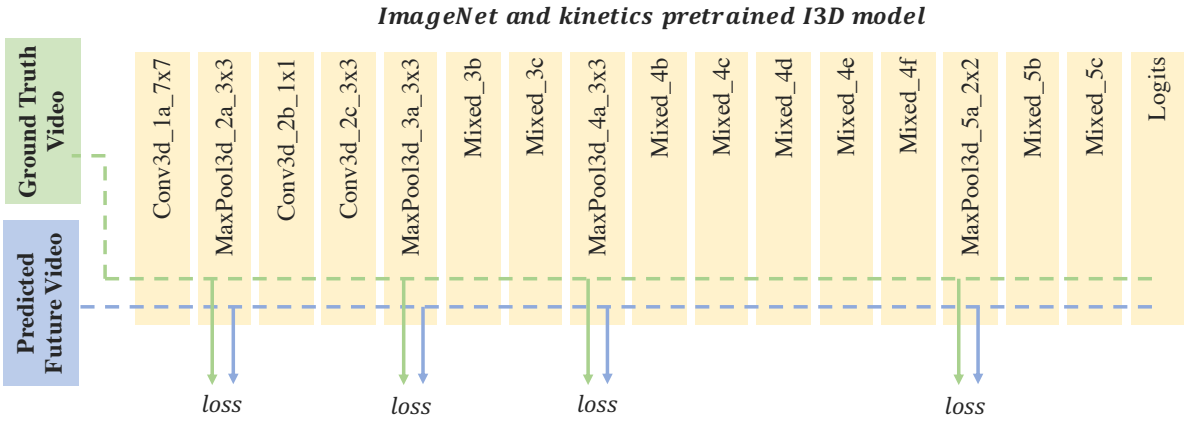

Figure 6.1: I3D perceptual loss.

Chapter 7

Conclusions

Summary of Our Work In this work, we propose to leverage the large amount of available videos to transfer the temporal information from video domain to static image domain to enhance static image action recognition. We propose a multi-scene future video prediction model that predicts the future video of a given static image. Instead of directly predicting the pixels of the future video, we predict the motion of those pixels. To reflect the missing regions that result from pixels moving in and out of the frames we use a mask, and propose a completion network to fill those regions. We then use the predicted video to improve static image action recognition performance.

7.1 Limitations and Future Work

Prediction Quality We note that our future prediction model is often able to learn highly plausible motion, however, it is not fully capturing all the action classes as some of the predicted motion might be slightly random. To combat this issue, we propose to train on a larger dataset like Kinetics [18].

Video Prediction Action Recognition Loss Having the static image action recognition ground truth and mock video results significantly exceeding the state of the art as shown in table 4.3 shows that the idea of leveraging future video to improve static image action

recognition performance is quite promising. Despite the significantly improved quality of our predicted video compared to the state of the art [19], our static image action recognition results are quite far away from the ground truth results, and do not show significant improvement from the mock video results. One way to combat this issue is to better capture and transfer the action features from videos to static images. We are currently working on enforcing an action recognition loss to help transfer the action features from videos to static images and improve the performance of static image action recognition.

Multi-Stream Architecture To further enhance the action recognition, we can experiment with two stream action recognition models. For example, we could use the original static image and the predicted video or the predicted video and the predicted flow. Moreover, we can experiment with having all three streams the static image, the predicted video, and the predicted flow.

Resolution Effect We also note that having a higher resolution might improve the recognition results as shown in table 4.4, which we do not do due to memory limitations.

Multi-modality It is important to note that the future is ambiguous and is multi-modal. A single static image might have multiple futures. We can introduce this multi-modality, and compute the static image action recognition from k possible futures and experiment the effect of multi-modality of the future on static image action recognition.

Flow Supervision Because the warped video supervision introduced significant improvement, we can experiment with adding flow supervision instead of the warped video supervision and measure its effect on both the prediction and action recognition performance.

Bibliography

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017.

[4] Baoyang Chen, Wenmin Wang, Jinzhuo Wang, and Xiongtao Chen. Video imagination from a single image with transformation generation. *CoRR*, abs/1706.04124, 2017. URL http://arxiv.org/abs/1706.04124.

[5] Chao-Yeh Chen and Kristen Grauman. Watching unlabeled video helps learn new human actions from very few labeled snapshots. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 572–579. IEEE, 2013.

[6] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*, 2010. updated version, available at http://www.di.ens.fr/willow/research/stillactions/.

[7] Vincent Delaitre, Josef Sivic, and Ivan Laptev. Learning person-object interactions for action recognition in still images. In *Advances in neural information processing systems*, pages 1503–1511, 2011.

[8] Emily Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. *arXiv preprint arXiv:1705.10915*, 2017.

[9] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[10] 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.

[11] Ruohan Gao, Bo Xiong, and Kristen Grauman. Im2flow: Motion hallucination from static images for action recognition. *arXiv preprint arXiv:1712.04109*, 2017.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[13] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://www.thumos.info/, 2015.

[14] 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 5769–5779, 2017.

[15] Guodong Guo and Alice Lai. A survey on still image based human action recognition. *Pattern Recognition*, 47(10):3343–3361, 2014.

[16] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.

[17] 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.

[18] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[19] Bernhard Kratzwald, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Towards an understanding of our world by ganing videos in the wild. *CoRR*, abs/1711.11453, 2017. URL http://arxiv.org/abs/1711.11453.

[20] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[21] Dieu Thu Le, Raffaella Bernardi, and Jasper Uijlings. Exploiting language models to recognize unseen actions. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 231–238. ACM, 2013.

[22] Li-Jia Li and Li Fei-Fei. What, where and who? classifying events by scene and object recognition. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[23] William Lotter, Gabriel Kreiman, and David D. Cox. Deep predictive coding networks

for video prediction and unsupervised learning. *CoRR*, abs/1605.08104, 2016. URL http://arxiv.org/abs/1605.08104.

[24] Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *CoRR*, abs/1511.05440, 2015. URL http://arxiv.org/abs/1511.05440.

[25] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1QRgziT-.

[26] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 4694–4702. IEEE, 2015.

[27] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International journal of computer vision*, 79(3):299–318, 2008.

[28] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L. Lewis, and Satinder P. Singh. Action-conditional video prediction using deep networks in atari games. *CoRR*, abs/1507.08750, 2015. URL http://arxiv.org/abs/1507.08750.

[29] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[30] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and

Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[31] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *CoRR*, abs/1606.03498, 2016. URL http://arxiv.org/abs/1606.03498.

[32] Fadime Sener, Cagdas Bas, and Nazli Ikizler-Cinbis. On recognizing actions in still images via multiple features. In *European Conference on Computer Vision*, pages 263–272. Springer, 2012.

[33] Gaurav Sharma, Frédéric Jurie, and Cordelia Schmid. Discriminative spatial saliency for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3506–3513. IEEE, 2012.

[34] K. Soomro, A. Roshan Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. In *CRCV-TR-12-01*, 2012.

[35] Graham W Taylor, Rob Fergus, Yann LeCun, and Christoph Bregler. Convolutional learning of spatio-temporal features. In *European conference on computer vision*, pages 140–153. Springer, 2010.

[36] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 4489–4497. IEEE, 2015.

[37] Joost van Amersfoort, Anitha Kannan, Marc'Aurelio Ranzato, Arthur Szlam, Du Tran, and Soumith Chintala. Transformation-based models of video sequences. *arXiv preprint arXiv:1701.08435*, 2017.

[38] Gul Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[39] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. 2017.

[40] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. *arXiv preprint arXiv:1704.05831*, 2017.

[41] Carl Vondrick and Antonio Torralba. Generating the future with adversarial transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[42] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *CoRR*, abs/1609.02612, 2016. URL http://arxiv.org/abs/1609.02612.

[43] Jacob Walker, Abhinav Gupta, and Martial Hebert. Dense optical flow prediction from a static image. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 2443–2451. IEEE, 2015.

[44] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *European Conference on Computer Vision*, 2016.

[45] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. *arXiv preprint arXiv:1705.00053*, 2017.

[46] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In

Computer Vision (ICCV), 2013 IEEE International Conference on, pages 3551–3558. IEEE, 2013.

[47] Tianfan Xue, Jiajun Wu, Katherine L Bouman, and William T Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *NIPS*, 2016.

[48] Weilong Yang, Yang Wang, and Greg Mori. Recognizing human actions from still images with latent poses. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2030–2037. IEEE, 2010.

[49] Bangpeng Yao and Li Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 9–16. IEEE, 2010.

[50] Bangpeng Yao and Li Fei-Fei. Action recognition with exemplar based 2.5 d graph matching. In *European Conference on Computer Vision*, pages 173–186. Springer, 2012.

[51] Bangpeng Yao and Li Fei-Fei. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1691–1703, 2012.

[52] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1331–1338. IEEE, 2011.

[53] 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.