

LEARNED FRAME PREDICTION FOR VIDEO COMPRESSION

by

Serkan Sülün

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of
Master of Science

in

Electrical and Electronics Engineering



**KOÇ
UNIVERSITY**

November 22, 2018

**LEARNED FRAME PREDICTION
FOR VIDEO COMPRESSION**

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Serkan Sülün

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

Prof. A. Murat Tekalp

Assoc. Prof. Engin Erzin

Prof. Bilge Günsel

Date: _____



To my family and friends

ABSTRACT

Motion compensation is one of the most essential methods for any video compression algorithm. Video frame prediction is a task analogous to motion compensation. In recent years, the task of frame prediction is undertaken by deep neural networks (DNNs). In this thesis we create a DNN to perform learned frame prediction and additionally implement a codec that contains our DNN. We train our network using two methods for two different goals. Firstly we train our network based on mean square error (MSE) only, aiming to obtain highest PSNR values at frame prediction and video compression. Secondly we use adversarial training to produce visually more realistic frame predictions. For frame prediction, we compare our method with the baseline methods of frame difference and 16x16 block motion compensation. For video compression we further include x264 video codec in the comparison. We show that in frame prediction, adversarial training produces frames that look sharper and more realistic, compared MSE based training, but in video compression it consistently performs worse. This proves that even though adversarial training is useful for generating video frames that are more pleasing to the human eye, they should not be employed for video compression. Moreover, our network trained with MSE produces accurate frame predictions, and in quantitative results, for both tasks, it produces comparable results in all videos and outperforms other methods on average. More specifically, learned frame prediction outperforms other methods in terms of rate-distortion performance in case of high motion video, while the rate-distortion performance of our method is competitive with x264 in low motion video.

ÖZETÇE

Tüm video sıkıştırma algoritmaları içinde en önemli metotlardan biri hareket dengelemedir. Video çerçevesi tahmini, hareket dengeleme ile benzer bir problemdir. Son yıllarda çerçeve tahmini, derin sinir ağları (DSA) tarafından yapılmaktadır. Bu tezde öğrenilmiş çerçeve tahmini yapmak için bir DSA yaratıyoruz ve ek olarak bu DSA'yı kapsayan bir video kodlayıcı üretiyoruz. DSA'mızı iki farklı amaç doğrultusunda iki farklı yöntem ile eğitiyoruz. İlk olarak çerçeve tahmini ve video sıkıştırmada en yüksek doruk sinyal gürültü oranı (PSNR) değerlerini elde etmek amacıyla DSA'mızı ortalama karesel hatayı (OKH) baz alarak eğitiyoruz. Ardından görsel olarak daha gerçekçi çerçeve tahminleri yapmak için çekişmeli eğitim yöntemini kullanıyoruz. Çerçeve tahmininde, yöntemimizi, referans yöntemler olan çerçeve farkı ve 16x16 blok hareket dengeleme ile karşılaştırıyoruz. Video sıkıştırmada, karşılaştırmaya x264 video kodlayıcıyı da dahil ediyoruz. Çerçeve tahmininde, çekişmeli eğitimin, OKH ile eğilmiş DSA'ya göre daha keskin ve gerçekçi çerçeveler ürettiğini; ancak video sıkıştırmada sürekli olarak daha başarısız olduğunu gösteriyoruz. Bu sonuç, çekişmeli eğitimin insan gözüne daha hoş görünen video çerçeveleri üretmesine rağmen video sıkıştırmada kullanılmaması gerektiğini kanıtıyor. Ek olarak, OKH ile eğitilen DSA, çerçeve tahmininde yüksek doğruluklu sonuçlar üretiyor; nicel sonuçlarda, iki problem için de, tüm videolarda diğer yöntemlerle kıyaslanabilir sonuçlar veriyor ve ortalama başarıda diğer yöntemleri geçiyor. Daha detaylı olarak, yüksek hareketli videolarda, öğrenilmiş çerçeve tahmini, bithızı-bozulma performansında diğer yöntemleri geçiyor; ve düşük hareketli videolarda x264 ile yarışabilir bir sonuç üretiyor.

ACKNOWLEDGMENTS

Firstly I would like to thank my supervisor Prof. A. Murat Tekalp for his academic and technical guidance, positivity, patience and unceasing laughter. I also thank my office mates Selin, Tuğtekin, Tolga, Akın and Berker for their intellectual contributions and especially Arda, Ogün and Gonca for their deep friendship and drinking ability. In addition I want to thank my friends far away, Doğan, Felister and Melek for their never ending support. To my family Lale, Ali and Zeynep, I can not thank enough, for their unconditional love and compassion, for granting me freedom to follow any path I want, for spending almost all of their time with me as a child, and for giving me a happy and peaceful home, amongst plants and animals, in a town by the mountains and the sea. Finally I thank my dearest Ceren for her unbounded love.

TABLE OF CONTENTS

List of Tables	ix
List of Figures	x
Nomenclature	xiii
Chapter 1: Introduction	1
Chapter 2: Background and Related Work	4
2.1 Background on Neural Networks and Deep Learning	4
2.1.1 Convolutional Neural Networks	7
2.1.2 Generative Networks	7
2.2 Related Work on Video Frame Prediction	10
2.2.1 Categorization	20
Chapter 3: Methodology	21
3.1 Learned Video Frame Prediction	21
3.1.1 Choice of Architecture and Training Method	21
3.1.2 Generator Network	23
3.1.3 Discriminator Network	25
3.1.4 Training Details	26
3.2 Compression of Predictive Frame Differences	29
3.2.1 Encoder	30
3.2.2 Decoder	30

Chapter 4: Experimental Results	32
4.1 Experimental Setting and Evaluation Methodology	32
4.2 Frame Prediction Results	35
4.3 Baseline Video Codec Configurations	46
4.3.1 x264 Hierarchical Coding	46
4.3.2 x264 Sequential Coding	47
4.4 Predictive Video Compression Results	48
Chapter 5: Conclusion	53
Bibliography	55

LIST OF TABLES

4.1	MPEG test sequences.	32
4.2	Bjontegaard delta PSNR metrics with respect to x264 SEQ-VBR (anchor). Positive values indicate higher performance compared to the anchor.	52



LIST OF FIGURES

2.1	Perceptron	4
2.2	Non-linear functions	5
2.3	A layer with 3 inputs and 3 outputs	6
2.4	A two layer network with 3 inputs 2 outputs	6
2.5	A deeper network with 8 inputs, 4 outputs and 4 layers.	6
2.6	An autoencoder	8
2.7	A variational autoencoder	9
2.8	Generative Adversarial Network’s block diagram	9
2.9	EDSR model	10
2.10	Srivastava’s model	11
2.11	Mathieu’s multiscale architecture	12
2.12	Video Pixel Networks	13
2.13	Video Ladder Networks	14
2.14	Amersfoort’s model	15
2.15	Vondrick’s model	15
2.16	Villegas’ model. (a) Pose predicting LSTM (b) Frame predicting CNN	16
2.17	Wichers’ end-to-end model	16
2.18	Finn’s model	17
2.19	Qualitative results using a single frame as input to Finn’s model . . .	17
2.20	Babaeizadeh’s model	18
2.21	Lee’s model	19
2.22	Denton’s model. (a) Training with vanilla VAE (b) Training with con- ditional VAE (c) Testing with conditional VAE	19

3.1	Block diagram of the generator network: (a) the EDSR network (b) each residual block of the EDSR network shown in purple in (a). . . .	24
3.2	Block diagram of the discriminator network.	26
3.3	The block diagram of the proposed encoder.	31
3.4	Block diagram of the decoder.	31
4.1	Computation of the Bjontegaard delta PSNR (BD-PSNR) metric [1].	35
4.2	PSNR of predicted frames vs. frame number for all test videos.	37
4.3	Coastguard. (a) Predicted frame using LFP-MSE method, (b) Predicted frame using LFP-GAN method, (c) Frame difference of LFP-MSE, (d) Frame difference of LFP-GAN. Difference values are scaled by 5 and level shifted by 128.	38
4.4	Container. (a) Predicted frame using LFP-MSE method, (b) Predicted frame using LFP-GAN method, (c) Frame difference of LFP-MSE, (d) Frame difference of LFP-GAN. Difference values are scaled by 5 and level shifted by 128.	39
4.5	Football. (a) Predicted frame using LFP-MSE method, (b) Predicted frame using LFP-GAN method, (c) Frame difference of LFP-MSE, (d) Frame difference of LFP-GAN. Difference values are not scaled. They are level shifted by 128.	40
4.6	Foreman. (a) Predicted frame using LFP-MSE method, (b) Predicted frame using LFP-GAN method, (c) Frame difference of LFP-MSE, (d) Frame difference of LFP-GAN. Difference values are scaled by 2 and level shifted by 128.	41
4.7	Garden. (a) Predicted frame using LFP-MSE method, (b) Predicted frame using LFP-GAN method, (c) Frame difference of LFP-MSE, (d) Frame difference of LFP-GAN. Difference values are scaled by 3 and level shifted by 128.	42

4.8	Hall Monitor. (a) Predicted frame using LFP-MSE method, (b) Predicted frame using LFP-GAN method, (c) Frame difference of LFP-MSE, (d) Frame difference of LFP-GAN. Difference values are scaled by 5 and level shifted by 128.	43
4.9	Mobile. (a) Predicted frame using LFP-MSE method, (b) Predicted frame using LFP-GAN method, (c) Frame difference of LFP-MSE, (d) Frame difference of LFP-GAN. Difference values are scaled by 2 and level shifted by 128.	44
4.10	Tennis. (a) Predicted frame using LFP-MSE method, (b) Predicted frame using LFP-GAN method, (c) Frame difference of LFP-MSE, (d) Frame difference of LFP-GAN. Difference values are not scaled. They are level shifted by 128.	45
4.11	Compression performance on test videos. Left: PSNR vs. bitrate. Right: Motion vector bitrate vs. total bitrate	51

NOMENCLATURE

DNN: Deep neural network
ANN: Artificial neural network
CNN: Convolutional neural network
GAN: Generative adversarial network
MSE: Mean square error
PSNR: Peak signal to noise ratio
GPU: Graphical processing unit
CPU: Central processing unit
SR: Super-resolution
SISR: Single-image super-resolution
EDSR: Enhanced deep super-resolution network
LSTM: Long short-term memory
SSIM: Structural similarity index measure
VAN: Visual analogy network
FP: Frame prediction
LFP: Learned frame prediction
ReLU: Rectified linear units
Tanh: Hyperbolic tangent
BPG: Better portable graphics
MPEG: Moving picture experts group
VBR: Variable bitrate
CBR: Constant bitrate
QP: Quantization parameter
BD: Bjontegaard delta

FD: Frame difference

MC: Motion compensation

NAL: Network abstraction layer

HRD: Hypothetical reference decoder



Chapter 1

INTRODUCTION

The transition from analog to digital media and the widespread adoption of the Internet as the premiere channel for media communication was made possible by advances in digital video processing. Sending digital video through a channel with limited bandwidth is a fundamental problem faced in video streaming over the Internet. A raw, high definition (1080p) video of 1 second occupies around 370 megabytes, while the highest national average of internet bandwidth is less than 8 megabytes per second. As a solution to this problem, *video compression* deals with reducing the file size while maintaining visual integrity.

The most essential technique used in video compression is *motion compensation*, which is done by dividing each video frame in blocks, estimating how these blocks move between frames and finally using this information to reconstruct each frame. Naturally this reconstruction is imperfect, so the *residual error* needs to be transmitted alongside the *motion vectors*. These two components constitute a large portion of the compressed video file. The sizes of these components can be reduced at the cost of less accurate reconstruction, creating videos with worse visual quality.

Until recently, the task of motion compensation, like all other signal processing problems, has been tackled using handcrafted methods. But in the last 8 years the human-engineered signal processing methods are being outperformed. With the advent of the modern graphical processing units (GPUs), artificial intelligence is more popular than ever. Using the computational power of these new GPUs, it only takes hours to create models that would require months ten years ago. These models of artificial neural networks (ANNs) can have hundreds of times more parameters than

their older versions, including many layers of neurons and so we call them deep neural networks (DNNs). The usage of DNNs has prompted a subfield of machine learning, named *deep learning*.

Having a vast number of neurons has greatly increased the computational and functional capacities of the ANNs. Comparing to human-engineered methods, DNNs are much more effective in modeling the real world phenomena. Almost every day, we use a DNN as a solution of a new engineering problem. Furthermore, techniques using DNNs are becoming the state-of-the-art in many fields of engineering.

A special type of ANN called *convolutional neural network* (CNN) which models the human visual system, became a standard tool in computer vision, image and video processing tasks. Using these networks to create novel samples of data such as sounds and images, gave rise to another category of ANNs, named *generative networks*. These networks can be used in a common task in video processing and computer vision, named *video frame prediction*, namely generating future frame(s) of a video sequence, given its past frame(s).

This problem is never deterministic. Consider a video of a person waving her hand. It is not possible to know if she will raise her hand by 30, 40 or 50 centimeters. In fact, any movement that she will make can be represented as a probability distribution. If our goal is to maximize our prediction's accuracy, the best prediction would be the expected value (average) of this distribution. In the context of image generation, producing an averaged image yields a blurry output.

In order to create sharp images that look pleasing to the human eye, the goal must be different than maximizing accuracy measured by the mean square error (MSE). Moreover there is no handcrafted metric that perfectly reflects human evaluation. *Generative adversarial networks* (GANs) tackle this issue by using a second ANN, named *discriminator*, as a substitute for the human critic. The first ANN, named *generator* is trained to maximize scores given by the discriminator. Many of the works on video frame prediction use stochastic methods such as GANs, to create realistic frames, at the cost of reduced prediction accuracy.

Motion is almost always present in video sequences, hence this problem of video frame prediction is analogous to motion compensation. However, in frame prediction we do not employ the original (ground-truth) frame to be compressed even at the encoder. Learned frame prediction computes a predicted frame by only using the previous frames. Our goal is to perform predictive video compression by replacing the traditional motion compensation algorithm with a generative DNN for learned frame prediction. This method makes motion vectors redundant since a single DNN can perform frame prediction for any video. Consequently, compressed video occupies less space, or this extra space can be used to increase video quality.

We create and train DNNs to perform frame prediction on natural video sequences, with the highest possible accuracy. We also train a GAN to obtain sharper and more realistic images. Moreover, we build a codec encapsulating these DNNs, for the task of video compression and present a comparative analysis including the state-of-the-art video compression methods.

Chapter 2

BACKGROUND AND RELATED WORK

2.1 Background on Neural Networks and Deep Learning

Artificial neural networks (ANNs) are self-learning computational systems inspired by the animal brain. We use these networks to build approximations of unknown functions that take *inputs* and create *outputs*.

In the simplest case of a single neuron, *perceptron*, multiplies all inputs $\{x_1, x_2, \dots, x_n\}$ with its *weights* $\{w_1, w_2, \dots, w_n\}$, and then sums them up. Next, a *bias* term b is added and the result is then put through a non-linear function $f(\cdot)$ to obtain the final output y , as displayed in Figure 2.1 and formulized in Equation 2.1.

$$y = f\left(\sum_{i=1}^n w_i x_i + b\right) \quad (2.1)$$

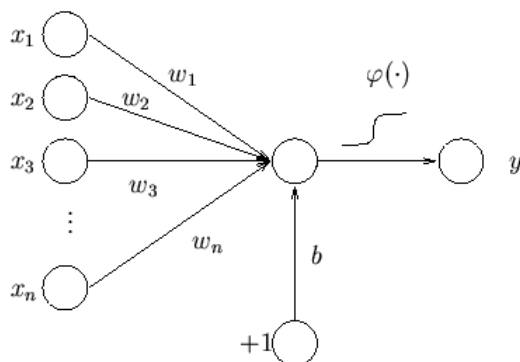


Figure 2.1: Perceptron

Since multiplying by a weight and adding a bias are linear operations, including a non-linear function is essential for better approximations. Figure 2.2 shows some popular examples of non-linear functions.

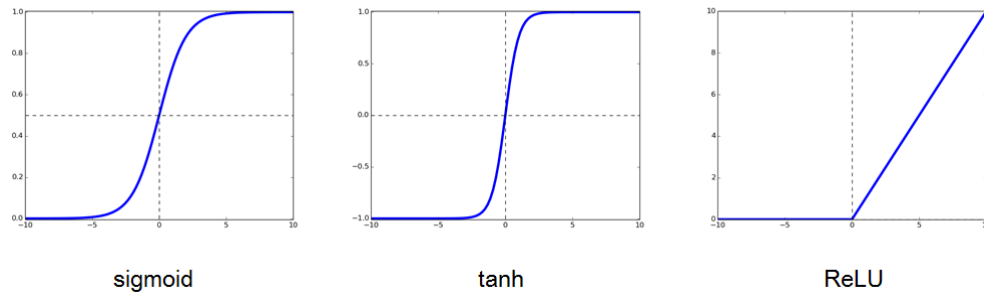


Figure 2.2: Non-linear functions

In the case of multiple outputs, neurons can be stacked together, to create a *layer*, as shown in Figure 2.3 and formulized in Equation 2.2 and 2.3.

$$y_j = f\left(\sum_{i=1}^n w_{ij}x_i + b_j\right) \quad j = \{0, 1, \dots, n\} \quad (2.2)$$

Or in matrix notation:

$$\begin{aligned}
 y &= f(Wx + b) \\
 y &= \begin{bmatrix} y_1 & y_2 & \dots & y_m \end{bmatrix}^T \\
 b &= \begin{bmatrix} b_1 & b_2 & \dots & b_m \end{bmatrix}^T \\
 x &= \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}^T \\
 W &= \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \dots & w_{mn} \end{bmatrix}
 \end{aligned} \quad (2.3)$$

We call these layers of neurons a *fully connected layer* since every input affects every output. In a single layer with n inputs and m outputs, there are $m \times n$ weights and m biases.

To approximate higher order functions, we stack these layers together and create a *neural network*. Figure 2.4 shows an example.

This operation can be written as:

$$y = g(W_2(f(W_1 + b_1)) + b_2) \quad (2.4)$$

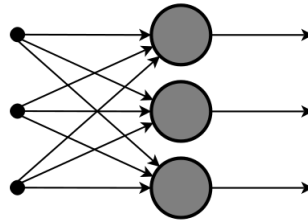


Figure 2.3: A layer with 3 inputs and 3 outputs

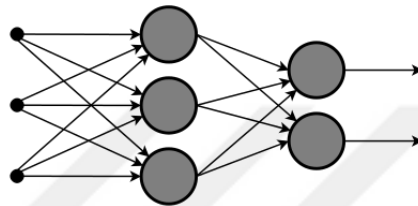


Figure 2.4: A two layer network with 3 inputs 2 outputs

where g is another non-linear function.

To create more complex models, we need to stack more layers, as in Figure 2.5. The number of layers indicate how *deep* a neural network is.

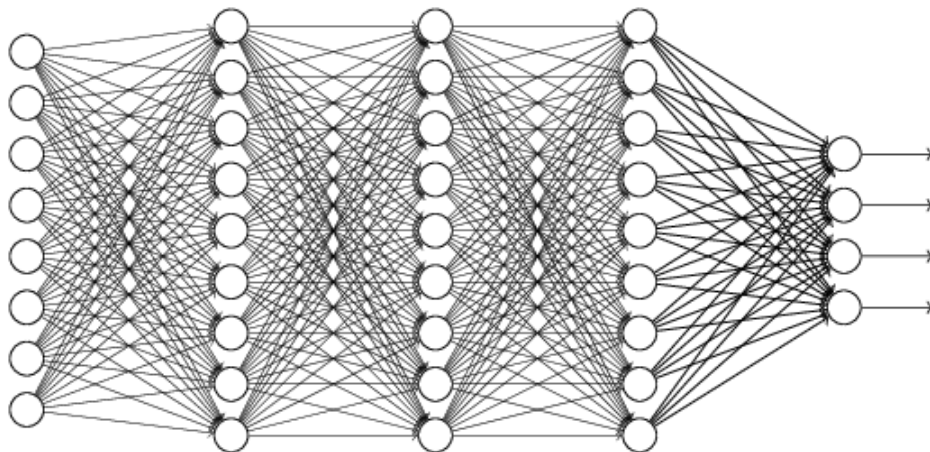


Figure 2.5: A deeper network with 8 inputs, 4 outputs and 4 layers.

2.1.1 Convolutional Neural Networks

Convolutional neural networks (CNN) [2] are a class of ANNs inspired by the animal visual cortex and mostly used in visual data analysis. In CNNs, weights are grouped into *kernels* (or *filters*). These kernels are *convolved* with the input data. Video and colored image data are 3 dimensional, having *width*, *height* and *channels*. A single convolutional layer which takes a 3 dimensional input has 4 dimensional weights, with *input channels*, *output channels*, *kernel width* and *kernel height*. We define convolution as

$$y(x, y, n) = \sum_{c=0}^C \sum_{p=0}^k \sum_{q=0}^k I(x+p, y+q, c) \cdot w(c, n, p, q) \quad (2.5)$$

Here $y(x, y, n)$ denotes the pixel value of the output image y 's n th channel, at coordinates (x, y) , and k denotes the side length of a square kernel. Just as fully connected layers, we stack convolutional layers to create deeper networks.

2.1.2 Generative Networks

Networks with a fully connected layer at its output layer produce a vector. In classification of a colored image, the output is a one dimensional vector of probabilities for each class, while the input is three dimensional. For each input sample, the goal is to find a suitable class *label*. We train the network based on labeled data, which is called *supervised learning*.

In contrast, generative networks specialize in creating outputs similar to its inputs. Since the goal is to produce data rather than labels, we train them based on unlabeled data, which is called *unsupervised learning*.

Autoencoders

Autoencoders are generative networks aiming to create an output that is the same as the input. This is usually done to create a compact representation of the data. Autoencoders consist of two modules named *encoder* and *decoder*. The encoder takes the input and creates an output which is smaller in size. The output of the encoder

is called *code*. The decoder takes the code and creates an output with the same size as the input. Figure 2.6 shows an example.

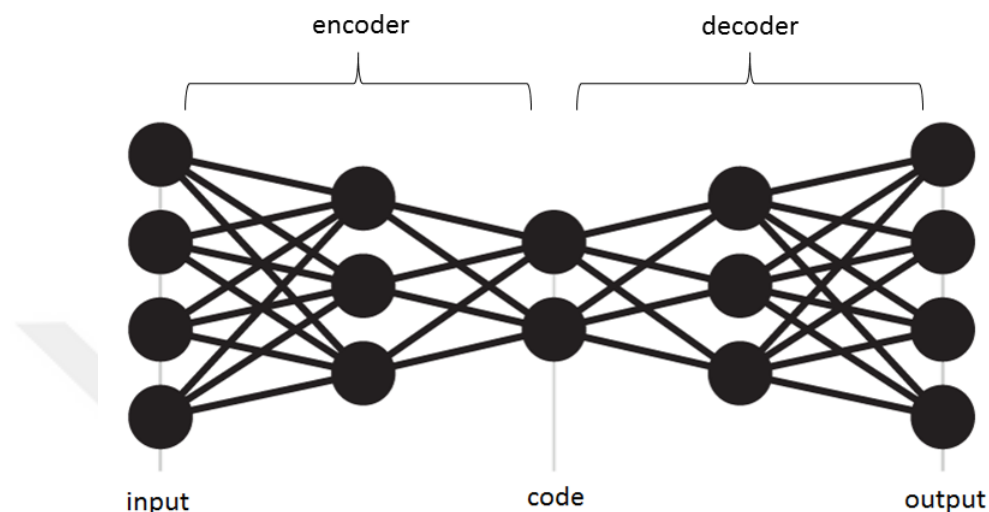


Figure 2.6: An autoencoder

Variational Autoencoders

Variational Autoencoders [3] are capable of creating novel outputs that are similar to the input data. They achieve this by modeling the code as a Gaussian random variable. Figure 2.7 shows an exemplary model. Half of the encoder's output constructs the mean and the other half constructs the standard variation of the code. Using these two parameters, we sample a random variable and feed it to the decoder. During deployment, we don't use the encoder, instead we simply obtain the sample from the Gaussian distribution and feed it into the decoder.

Generative Adversarial Networks

Generative adversarial networks (GANs) [4] are able to create authentic samples of data. This is achieved by using two competing networks, the *generator* and the *discriminator*.

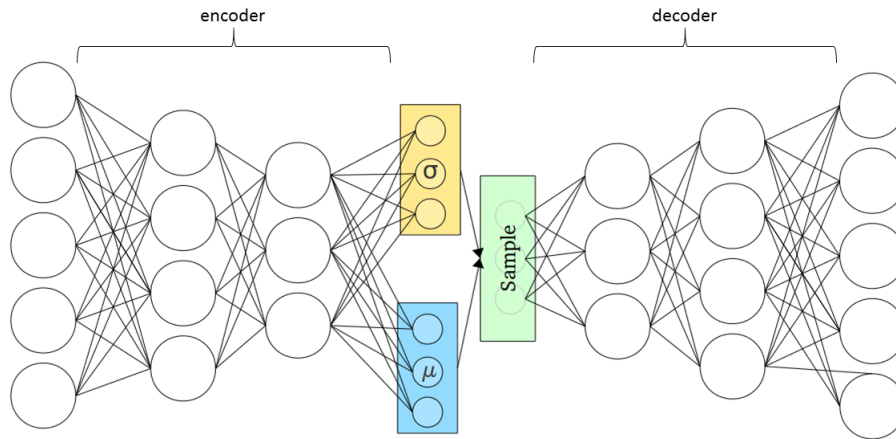


Figure 2.7: A variational autoencoder

The discriminator's task is to differentiate between the original data (real) and data generated by the generator (fake). Hence its output is a single number between 0 and 1, indicating the probability that its input is real. During training, we use both real and fake samples as inputs, with their target labels being 1 and 0 respectively.

The generator's task is to create outputs that will trick the discriminator. It is trained based on the discriminator's feedback, where the target value is 1 (real). To make sure that it produces different outputs, we use random noise as its input.

A block diagram can be seen in Figure 2.8.

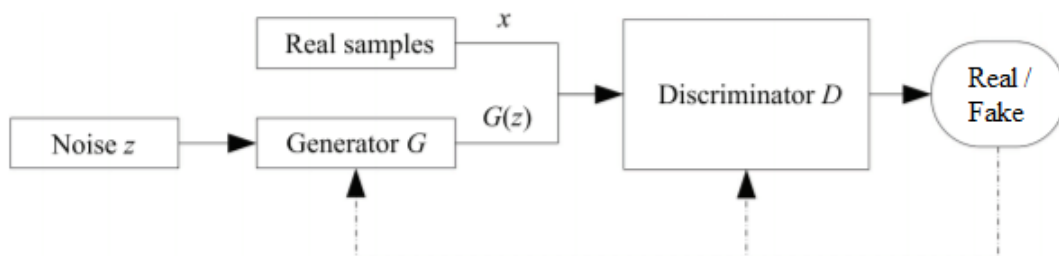


Figure 2.8: Generative Adversarial Network's block diagram

2.2 Related Work on Video Frame Prediction

We first present a paper written for the task of image super-resolution and not for video frame prediction, since our model has a very similar structure as explained in Section 3.1.2. Lim et al. have created a deep convolutional residual network named Enhanced Deep Super-Resolution Network (EDSR) [5], winning the NTIRE 2017 Challenge on Single Image Super-Resolution [6]. Their network takes a single colored low resolution image as input and reproduces this image at a higher resolution. To minimize error, they use a deep residual convolutional network. All intermediate layer's activation sizes (height and width) are the same as the input's, to prevent a loss of spatial information. In order to create the output image at a higher resolution they use an upsampling block with a pixel shuffling layer [7]. Their model can be seen in Figure 2.9.

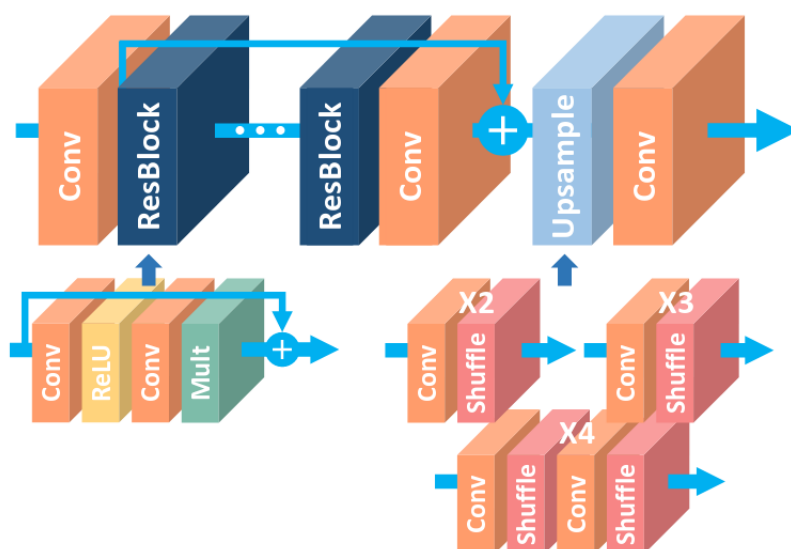


Figure 2.9: EDSR model

Next, works on video frame prediction are presented chronologically and finally they are classified under various categories. We can classify these models based on their training losses and methods. The most commonly used losses are mean square

error, ℓ^1 distance and cross-entropy loss, but some works define a novel loss function e.g by taking a difference of image gradients. Stochastic methods use variational encoders or GANs, employing KL-divergence or adversarial loss, respectively. Conceptual loss is also used, by comparing frames at the feature space, using a pretrained network. Another way to categorize these networks is by their output domain, namely pixel level, patch level and frame level generation. Finally, we can categorize them by their method of generation, frame reconstruction and frame transformation. We note that many of these works fit into multiple categories. We also encourage interested readers to examine two surveys [8, 9] written on video frame prediction using DNNs.

Srivastava et al. [10] work on two tasks, input frame reconstruction and future frame generation using long short-term memory (LSTM) autoencoders. They also come up with a composite model, performing reconstruction and prediction at the same time, outperforming the model that only predicts the future frames. They further improve their results by conditioning the decoder using the output from previous time-step. Their best performing model can be seen in Figure 2.10.

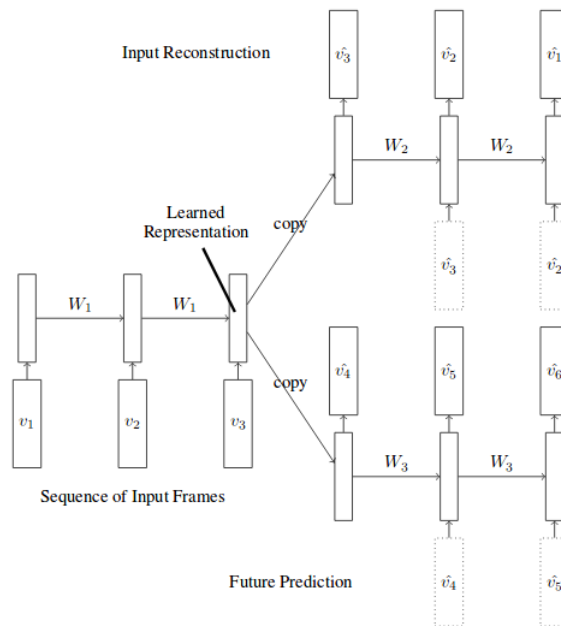


Figure 2.10: Srivastava's model

Mathieu et al. [11] aim at generating sharper images. Their first contribution is to use a multi-scale architecture, concatenating the output of the smaller scale's network with the input of the larger scale's network, as seen in Figure 2.11. Their second contribution is to use different loss functions, complementary to the mean square loss. They introduce gradient difference loss by calculating the difference between the gradients of predicted and ground-truth images. They also use a multi-scale discriminator network to incorporate adversarial loss. The training loss is a weighted sum of mean square, gradient difference and adversarial losses.

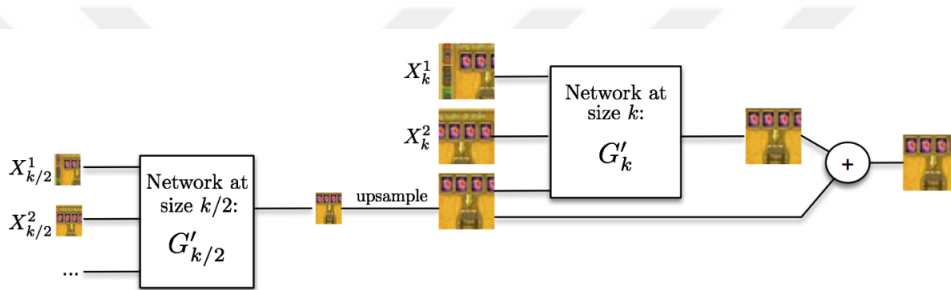


Figure 2.11: Mathieu's multiscale architecture

Kalchbrenner et al. [12] model the pixel values in four dimensions (time, height, width, color) directly and causally. Their model, named Video Pixel Networks, has an encoder-decoder structure. The encoder part consists of CNNs with dilated convolutions, followed by a convolutional LSTM. This part of the network models temporal dependencies, and its output conditions the decoder. The decoder consists of a PixelCNN [13] which models individual pixels sequentially. Using masked convolutions, PixelCNN can use previously generated pixels within the same image and even the pixel at the same location from the previous color channel, as inputs. Implicitly, the decoder can access the previous frames since the output of the encoder is passed to the decoder. Their model can be seen in Figure 2.12.

Cricri et al., in their work named Video Ladder Networks [14] use a multi-scale encoder-decoder architecture. Encoder and decoder consist of cascaded upscaling and downscaling convolutional layers respectively. Encoder's activations at each scale are

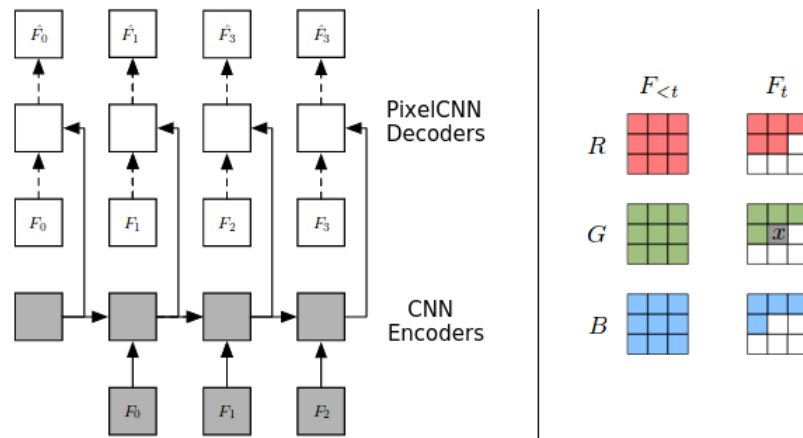


Figure 2.12: Video Pixel Networks

transmitted to the decoder’s convolutional layer at the corresponding scale, through two lateral connections, a recurrent, and a feedforward (skip) connection. The recurrent connections employ convolutional LSTMs, capable of using information from the past frames. The feedforward connection only sends features from the current frame, hence it is useful for the modeling of the static image parts. Their baseline architecture can be seen in Figure 2.13. They extend this model by replacing the single convolutional layers with residual convolutional blocks.

Instead of modeling pixels, Amersfoort et al. [15] build a CNN that only predicts the affine transformations between the patches from consecutive frames. The network takes the affine transformation matrices from the previous frames as input and outputs a predicted affine transformation matrix for the next frame. This network is trained using mean square loss between the predicted and ground-truth affine transformation matrices. The predicted affine transformation is later applied on the current frame to generate the next one. Their model can be seen in Figure 2.14.

Vondrick et al. [16] also aim at predicting frame transformations but train the entire model end-to-end. A CNN takes the input frames and outputs a transformation matrix. The transformation matrix is used to warp the last frame. Using this matrix, for each pixel, a weighted average of its neighboring pixels is computed. This is

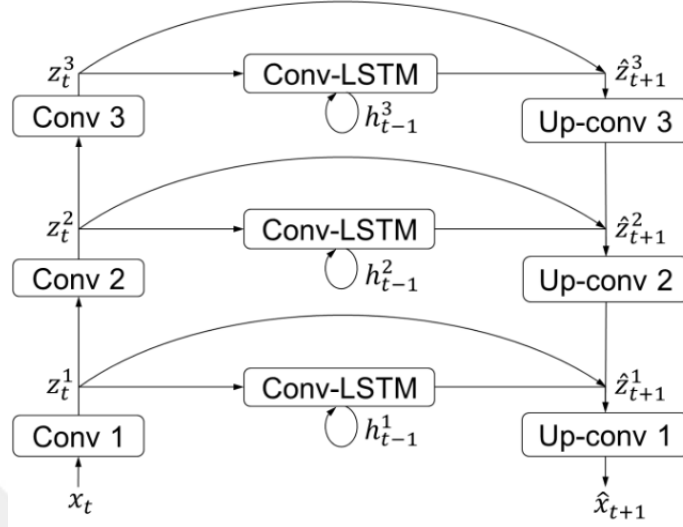


Figure 2.13: Video Ladder Networks

implemented as the standard convolution operation. The model is seen in Figure 2.15. To obtain sharp predictions, the network is trained using adversarial loss only. They have measured the model’s performance by using it as a pretraining method for object classification tasks, and with subjective evaluation. No peak signal-to-noise ratio (PSNR) or structural similarity index measure (SSIM) results are reported.

Villegas et al. [17] make long term predictions on human videos, while avoiding the frame by frame accumulation of error, namely drift. With the help of a pretrained network [18], they extract human pose information from each frame. Using the pose information as input, an LSTM autoencoder is trained, specifically on future pose prediction. The last frame, its corresponding pose and the future pose are fed into a visual analogy network (VAN) [19], that is a CNN with arithmetic operations, to generate the future frame. The loss function for this network consists of the sum of ℓ^2 , perceptual and adversarial losses. Perceptual loss is obtained by measuring the distance between the generated and ground-truth images, in a feature space created by a pretrained classification network [20]. The only recurrent predictions are done by the pose predicting network, hence a predicted pose at an arbitrary future time-step can be used to generate the corresponding frame, without generating its earlier

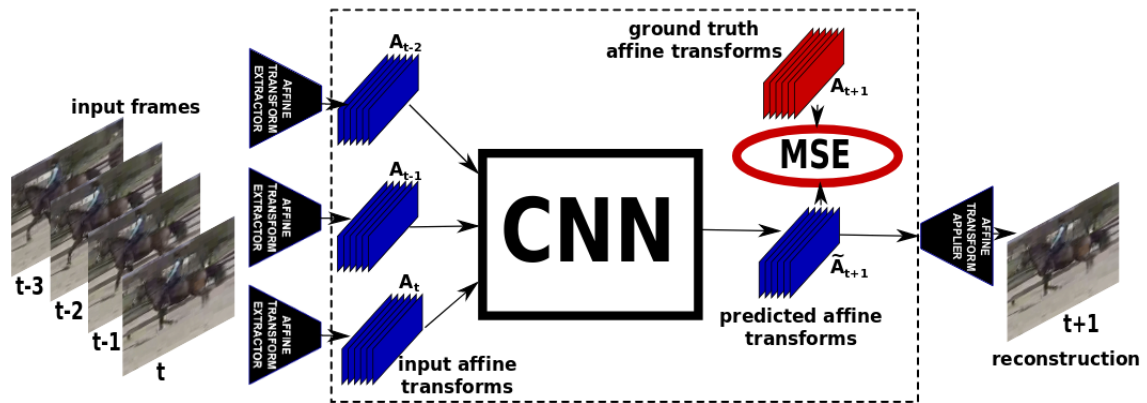


Figure 2.14: Amersfoort's model

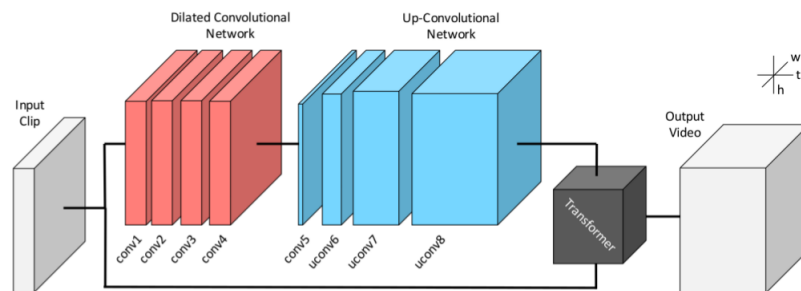


Figure 2.15: Vondrick's model

frames. The model can be seen in Figure 2.16.

Following up the work of Villegas, Wichers et al. [21] replace the pre-trained pose extracting network with a trainable encoder, enabling end-to-end training. An LSTM is used to predict the encoding of the next frame, given the encoding of the current frame. The model can be seen in Figure 2.17. They have performed three experiments. Firstly, they trained the entire network end-to-end. With this method, they have observed that VAN produces blurry images when the predictor LSTM is not confident. To address this issue they have trained the predictor and VAN separately, obtaining sharper outputs. To achieve even sharper results, they have included adversarial loss during the training of the predictor. It should be noted that a direct quantitative evaluation of generated images using standard metrics such as

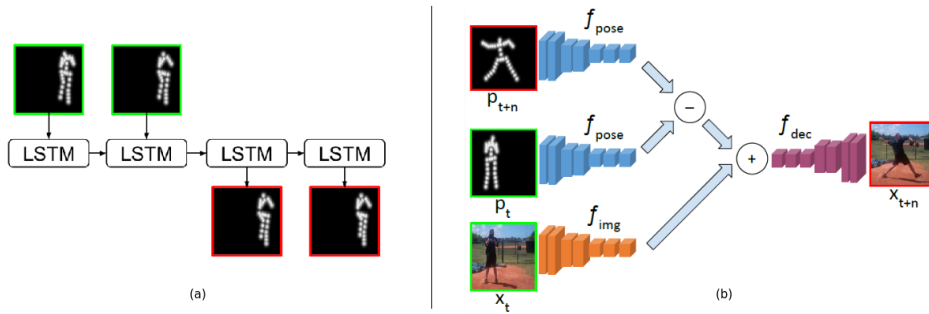


Figure 2.16: Villegas' model. (a) Pose predicting LSTM (b) Frame predicting CNN

PSNR or SSIM is not presented.

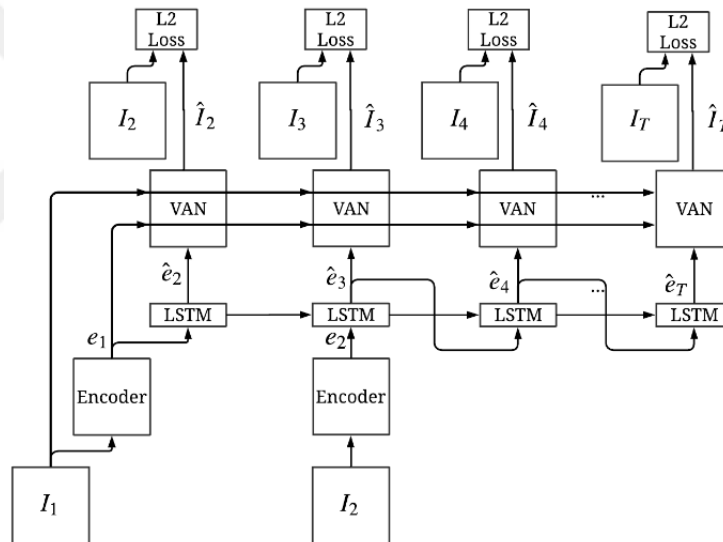


Figure 2.17: Wichers' end-to-end model

Finn et al. [22] focus on predicting object motions and then transforming the video frames rather than reconstructing them. In order to model motion explicitly, they use convolutional LSTMs. Their models are used on Human3.6M and robotic pushing datasets. For robotic pushing dataset, robot states and actions are also used as inputs by concatenating with the activation of the middle layer. The size of the activation is compatible with the state-action input, which is smaller than the rest of the activations. This is achieved by using an encoder-decoder structure, with

decreasing activation sizes until the middle of the network and increasing activation sizes afterwards. Their models create motion transformation kernels to apply motion transformation to the input image, and they create compositing masks to combine multiple predictions into a single image. Their model, can be seen in Figure 2.18.

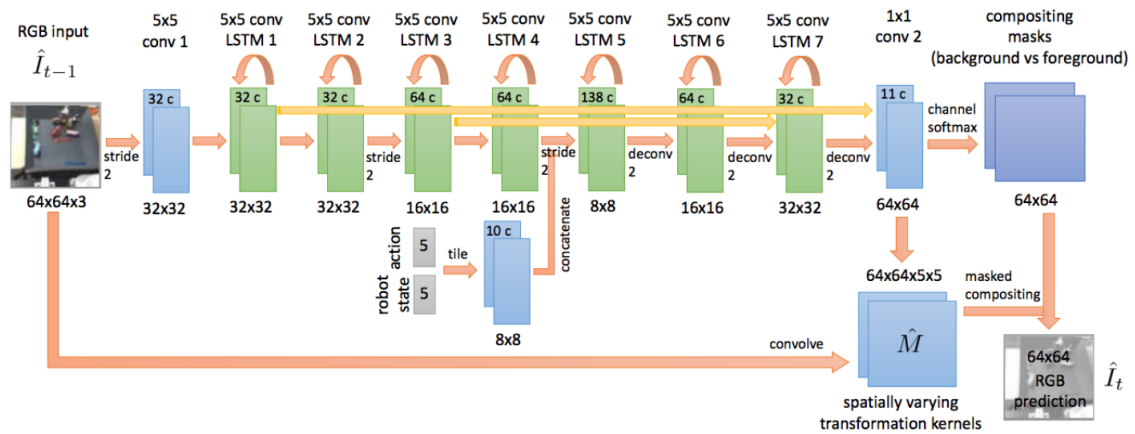


Figure 2.18: Finn's model

In a follow-up work, Babaeizadeh et al. [23] show that, given a single input frame, Finn's model produces a blurry output, including artifacts from all possible future outcomes, as show in Figure 2.19. To avoid this, they supplement Finn's model with a variational encoder. The output of the encoder is used to sample a latent variable which is fed into the generator. This sampling procedure is equivalent to choosing a single mode from the multimodal distribution representing possible future outcomes. The model is shown in Figure 2.20.

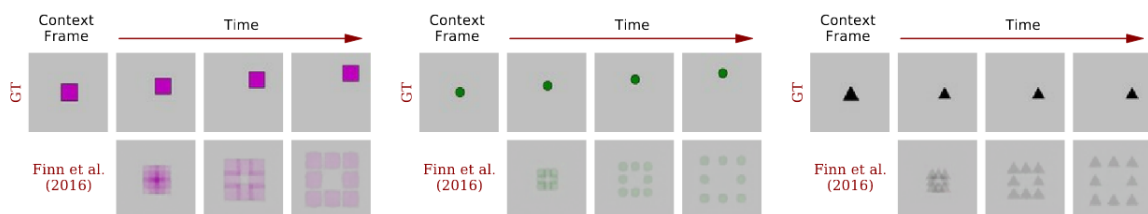


Figure 2.19: Qualitative results using a single frame as input to Finn's model

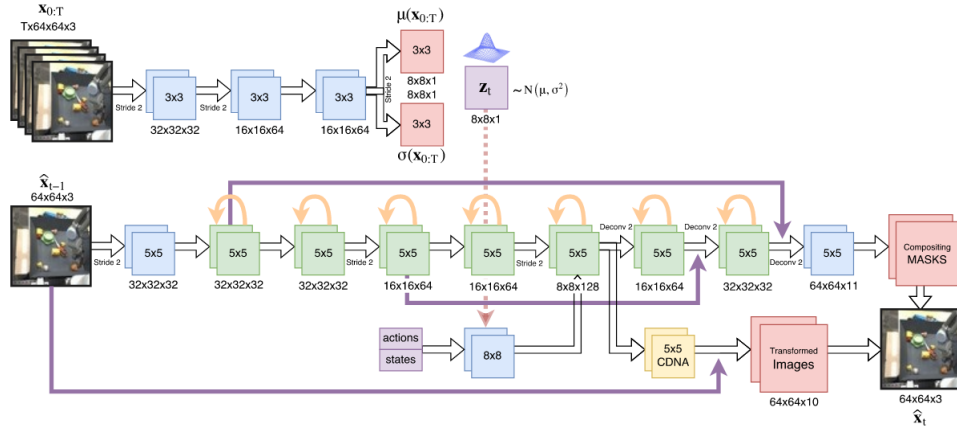


Figure 2.20: Babaeizadeh's model

In the next follow-up work, Lee et al. [24] supplement Babaeizadeh's model with an adversarial loss, to obtain more realistic results. The model can be seen in Figure 2.21.

Denton et al. [25] use a variational encoder to deal with uncertainty. The output of the encoder is a probability distribution of the potential outcomes for the predicted frame. A latent variable is sampled from this distribution, and together with the input frame, they are fed into the decoder. They use convolutional LSTMs to enable training at each time-step for a variable number of frames.

They present two models, using a vanilla variational encoder and a conditional variational encoder. The former models the uncertainty as a zero mean unit variance Gaussian distribution while the latter creates a Gaussian distribution conditioned on the previous frames. Their models at training and test times can be seen in Figure 2.22.

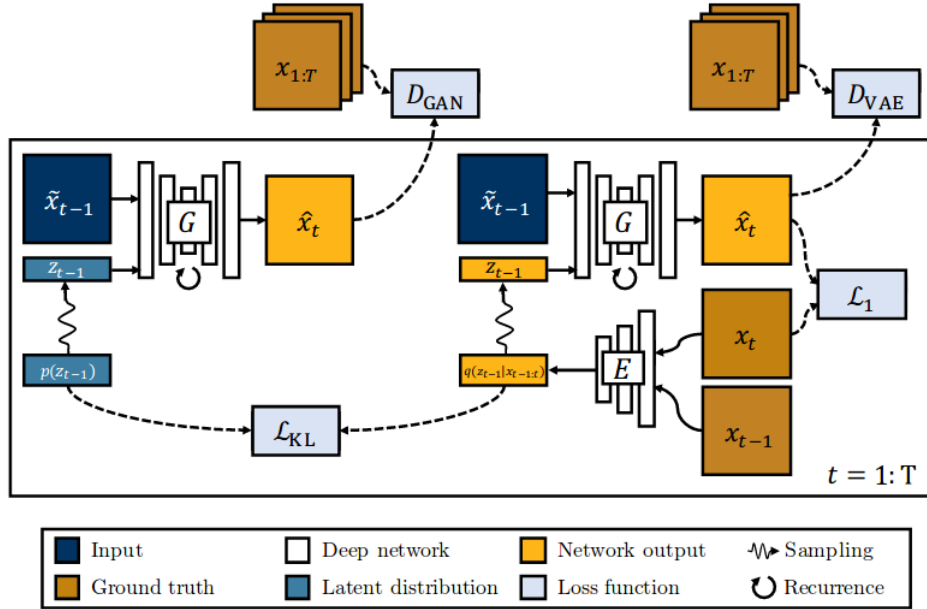


Figure 2.21: Lee's model

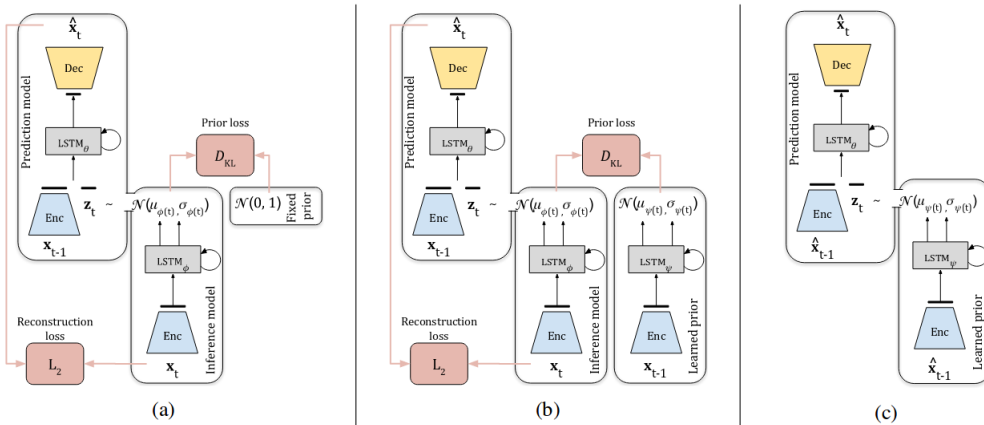


Figure 2.22: Denton's model. (a) Training with vanilla VAE (b) Training with conditional VAE (c) Testing with conditional VAE

2.2.1 Categorization

In this section, related works are classified under categories defined in the beginning of Section 2.2.

Loss/training method

Mean square error: Srivastava [10], Mathieu [11], Amersfoort [15], Villegas [17], Wichers [21], Finn [22], Babaeizadeh [23], Lee [24], Denton [25]

ℓ^1 distance: Mathieu [11]

Cross entropy: Srivastava [10], Kalchbrenner [12], Cricri [14]

Gradient difference: Mathieu [11]

Perceptual loss: Villegas [17]

Adversarial loss/GAN: Mathieu [11], Vondrick [16], Villegas [17], Lee [24]

KL divergence/Variational: Babaeizadeh [23], Lee [24], Denton [25]

Network output domain

Frame level: Mathieu [11], Vondrick [16], Villegas [17], Wichers [21], Finn [22], Babaeizadeh [23], Lee [24], Denton [25]

Patch level: Srivastava [10], Amersfoort [15]

Pixel level: Kalchbrenner [12]

Method

Frame reconstruction: Srivastava [10], Mathieu [11], Kalchbrenner [12], Cricri [14], Wichers [21], Denton [25]

Frame transformation: Amersfoort [15], Vondrick [16], Villegas [17], Finn [22], Babaeizadeh [23], Lee [24]

Chapter 3

METHODOLOGY

The methodology proposed in this thesis for learned predictive video compression consists of two steps: learned video frame prediction, which is discussed in Section 3.1, and compression of the predictive frame differences, which is described in Section 3.2.

3.1 *Learned Video Frame Prediction*

We first discuss the rationale for our choice of network architecture and training method for the purpose of frame prediction in predictive video compression in Section 3.1.1. Next, we present the details of our generator and discriminator network architectures in Sections 3.1.2 and 3.1.3, respectively. We end this section by explaining the details of training procedures using standard mean square loss only and combined mean square and adversarial loss in Section 3.1.4.

3.1.1 Choice of Architecture and Training Method

Video frame prediction is a stochastic problem, where the target frame can be modeled by a multimodal probability distribution, since there are more than one possible outcomes for the predicted frame. Standard methods aiming to minimize the mean square error converge to the expected value of this multimodal distribution, which is an average of all possible outcomes; hence, minimization of mean square error generally yields a blurry predicted image.

Many approaches, which are reviewed in Section 2.2, have been proposed in recent years to overcome this problem and generate sharper and more realistic looking predicted images. A majority of these works use recurrent models to process an arbitrary number of input and output frames. Until recently, LSTMs were the top choice

of architecture to solve sequence learning problems, especially in speech and natural language processing, where recurrent models yield amazing results in speech to text conversion and machine translation. However, training LSTMs for video processing can still be problematic.

With the introduction of the ResNet architecture, which overcomes the vanishing gradients problem by clever bypass connections, it is easier to train deeper CNNs to learn temporal dynamics from a fixed number of past video frames. Feeding a fixed number of past frames to a CNN requires only modification of the input layer. Hence, the additional computational cost of processing a fixed-size block of input frames (instead of a single input frame) is insignificant compared to the computational cost of the rest of the network. In theory, LSTMs can remember entire history of a video from the beginning. However, in practice, we obtain as good if not better performance (due to training issues) by batch processing of a fixed size block of past frames using a CNN rather than processing all frames sequentially using a recurrent architecture.

Another important factor that affects video compression performance is the choice of training methodology and loss function. It has been observed in the literature that stochastic sampling methods such as variational autoencoders, and adversarial training procedures similar to those used in training GANs do produce sharper predicted images with more spatial detail in frame prediction problems. However, they do so at the expense of higher mean square error when compared to the actual next frame. This raises the question: How do we evaluate the prediction performance of a neural network? In computer vision applications such as surveillance or robotics planning, where the predicted images are to be viewed or used as final product and/or there is no ground-truth available, then sharper and more natural looking images are preferred. In contrast, in video compression the predicted image is only an intermediate result (not to be viewed), and the goal is to minimize the bitrate required to send the difference between the predicted and actual frames (ground-truth). In this case, it turns out that the entropy of the predictive frame difference is lower when we train the network using the mean square loss even if the predicted images look blurry.

In conclusion, we chose to employ a deep convolutional network, where the activation sizes stay constant from the input to the output layer; hence, avoiding autoencoder (encoder-decoder) architectures, recurrent architectures, and stochastic sampling approaches. We train the CNN using the mean square loss function in order to obtain the minimum mean square prediction error for video compression task. In the following, we propose to employ a modified version of the Enhanced Deep Super-Resolution (EDSR) network [5] as a generator for the task of video frame prediction. We only modified the input and output layers of the EDSR network as described below. Next, we propose a discriminator network for the purpose of adversarial training of the generator network. Finally, we discuss the details of training procedures used.

3.1.2 Generator Network

Learned frame prediction (FP) problem is similar to the learned single-image super-resolution (SISR) problem in that they both generate new images using generative networks. In the SISR problem, the generative network predicts a high resolution image based on a learned model of high frequency content of images given a low resolution image, while in the FP problem, the generative network predicts the next frame based on a learned model of temporal dynamics given N previous frames.

The Enhanced Deep Super-Resolution (EDSR) network [5] is the winner of the NTIRE 2017 Single Image Super-Resolution Challenge [6]. Unlike many other works, which employ an autoencoder architecture for image generation, in the EDSR network the size of all intermediate layer activations is the same as that of the input layer. That is, they avert loss of spatial information by avoiding an encoder-decoder architecture. This property of the EDSR network architecture and its success in the SISR problem form the basis of our choice for generator architecture.

The EDSR network takes a single color image and outputs a single color image with double the height and width. In contrast, our network's input layer takes N past grayscale (Y-channel) frames and our output layer produces a single grayscale frame, whose size (height and width) is the same as the input frames. Therefore, we

modified both the input and output layers of the EDSR network. Since the input and output images are all the same size, we don't need an upscaling layer that was used by the EDSR network.

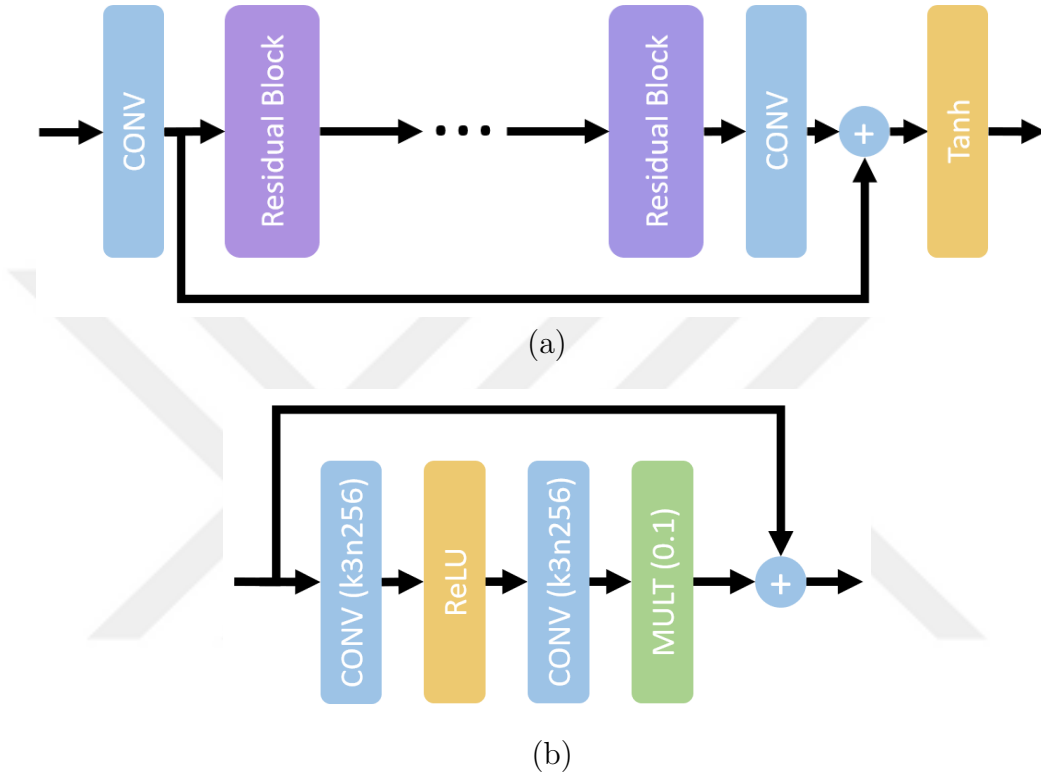


Figure 3.1: Block diagram of the generator network: (a) the EDSR network (b) each residual block of the EDSR network shown in purple in (a).

The architecture of the generator network is depicted in Figure 3.1. Our network has 32 residual blocks, each having 2 convolutional layers with kernel size 3, padding 1 and channel depth 256 and rectified linear units (ReLU) as the activation functions between hidden layers. The model is fully convolutional so it can process inputs with an arbitrary size (height and width). The height and width of all intermediate layer activations and the output layer are the same as those of the input layer. At the end of each residual block, residual scaling with 0.1 is applied [26]. At network's output hyperbolic tangent (Tanh) is used as the nonlinear function, ensuring output values between -1 and 1. At test time, the outputs are scaled between 0 and 255 and

converted to 8-bit unsigned integers to generate predicted video frames.

We observe that the best results with EDSR in the NTIRE 2017 competition were obtained by using their largest network. Since our ultimate goal is to obtain the best rate-distortion performance in video compression, we aim at obtaining the lowest mean square prediction error. As will be demonstrated in Chapter 4, the performance of the modified EDSR network to predict frames is surprisingly good. We observed in our FP experiments that using larger networks consistently provide lower mean square prediction error similar to that observed in SISR experiments. This is plausible since the task of image/video generation is highly complex so overfitting is rarely observed.

3.1.3 Discriminator Network

A discriminator network is needed for adversarial training of the generator network in order to compare performances of training with MSE only and adversarial training, both in terms of visual quality of predicted frames as well as the resulting rate-distortion performance in predictive video compression.

Even though large and deep residual networks with a fully connected output layer have been used as the discriminator network in some adversarially trained super-resolution models, such as the SRGAN model [27], in our experiments we obtained better results when a smaller discriminator network is used. The discriminator undertakes a simpler task compared to the generator network, where the output and target of the discriminator network are single numbers, while the generator has a much more complicated task of producing full frames. As a result, using similar size networks typically results in the discriminator outperforming the generator, causing an instability in the process of training. To this effect, we used a small discriminator network with no fully connected layers.

The block diagram of our proposed discriminator network architecture is depicted in Figure 3.2. There are three convolutional layers with kernel size of 7 and varying channels depths. Even though the discriminator is fully convolutional, by using pooling layers with kernel size 2 and stride 2, and avoiding padding, we reduce the

activation sizes throughout the network and obtain a single value in the output. We note that this is only possible when an input with size 48x48 is used, but this does not create any problems since the discriminator is only used during training. As recommended in [28], for a stable training, we have used average pooling instead of maximum pooling, to avoid sparse gradients. Between the hidden layers, leaky rectified linear unit (Leaky ReLU) with a slope of 0.2 is used as the nonlinearity, as advised in [29]. The final nonlinear function is a sigmoid, enabling an output value between 0 and 1, regarded as a probability of the input image’s authenticity (higher means more realistic).

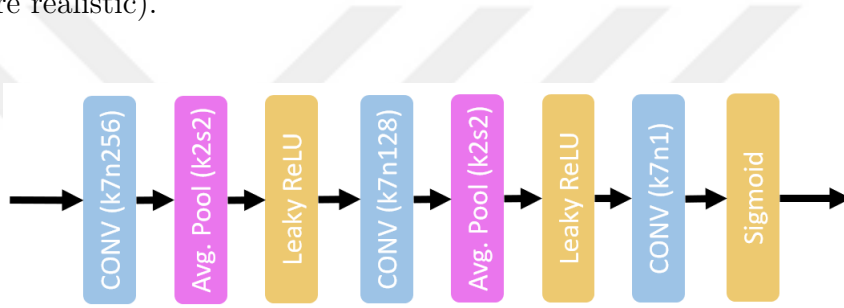


Figure 3.2: Block diagram of the discriminator network.

3.1.4 Training Details

We first trained the generator network using the mean square loss. Then, we trained the discriminator network together with our pretrained generator network in order to perform adversarial training, where the discriminator is trained with adversarial loss only, while the generator is trained with a weighted combination of mean square and adversarial losses.

Training Dataset

We used UCF101 [30] dataset, created by the University of Central Florida, as our training set. This dataset contains 13,320 videos of human actions, collected from Youtube. We converted the videos into grayscale (Y-channel) and extracted sequences

of patches with size 48×48 pixels and duration 9 frames. We predict the 9th frame given the first 8, and use the 9th frame of each patch sequence as the ground-truth.

While extracting patch sequences, we selected the video, the starting frame and patch location randomly, with the condition that there is motion between all consecutive pairs of patches within selected patch sequences. The pixel values range between 0 and 255, and we set the motion detection threshold as mean square difference of 7 between successive pairs of patches. With probability 0.05, we ignored this threshold and extracted the patch sequence anyway completely randomly. Overall, we extracted about 2 million patch sequences as our training dataset. We saved the extracted patch sequences as 8-bit unsigned integer to enable faster data transfer from the CPU to the GPU. At training time, we transfer the patches to the GPU, and then convert them to 32-bit floats by scaling the pixel values between -1 and 1.

With a training set of this size, we don't use the concept of training by *epochs*, meaning that we don't make multiple passes over the dataset. Instead, we pick a set of patch sequences randomly from the entire patch sequence dataset to construct minibatches at training time, leading to the concept of training by *steps* or *iterations*, where one training step corresponds to processing one minibatch.

Training with mean square loss

We first trained our model based only on mean square loss. Since the model takes 8 frames as input and predicts the 9th frame, we compute the mean square error over the 9th patch, which is defined by

$$L_{MS} = \frac{1}{N} \sum_{i=1}^N (y_i - x_i)^2 \quad (3.1)$$

where x and y denote the generated and the ground-truth (original) 9th patch, respectively, and i is the index looping over all N pixels in the ground-truth patch.

We used Adam optimizer [31] with an initial learning rate of $1e-4$ and a batch size of 32. If the training loss did not decrease for 6000 steps, we halved the learning rate. We trained our model for 400,000 steps (iterations), which lasted about 8 days using

a single NVIDIA GeForce GTX 1080Ti GPU on a HP Server with Intel Xeon Gold CPU @ 2.30GHz with 24 cores.

Training with combined mean square and adversarial losses

We next fine tune the parameters of the generator network by the process of adversarial training, which employs the discriminator network. To this effect, we jointly trained the randomly initialized discriminator and the pretrained generator networks. The minibatch sizes are 16 for the generator and 32 for the discriminator. A minibatch for training the discriminator network consists of 16 original (ground-truth) and 16 generated (fake) patch-sequence samples.

Original samples are actual training patch sequences. Generated samples are composed of the first 8 frames of original patch sequences that are input to the generator network concatenated with the predicted 9th patch that is the output of the generator network. Thus, a generated sample includes the first 8 original patches and a generated image as the 9th patch. By feeding sequences instead of single frames into the discriminator, we enforce the discriminator to make its predictions based on sequence continuity. This procedure is similar to that was used by Mathieu et al. [11]. We use target labels of 0 and 1 for the generated and original samples, respectively. Then, the discriminator network is trained using the binary cross entropy loss, given by

$$L_{BCE} = -y \log x - (1 - y) \log(1 - x) \quad (3.2)$$

where x is the discriminator network's output (prediction) and y is the binary ground-truth label.

The generator network is trained using a weighted combination of mean square and adversarial losses. While calculating the adversarial loss for the generator network, we only feed generated sequences to the discriminator and use a target label of 1, since the generator's aim is to create realistic images. Thus, the adversarial loss for

the generator network is defined as

$$L_{ADV}^G = -\log x_{disc} \quad (3.3)$$

Hence generator's combined loss becomes

$$L = \lambda_{MS} \left(\frac{1}{N} \sum_{i=1}^N (y_i - x_i)^2 \right) - \lambda_{ADV} \log x_{disc} \quad (3.4)$$

where x_i is a single pixel in generator's output, x_{disc} is discriminator's output, N is the number of pixels, λ_{MS} and λ_{ADV} are weights for the mean square and adversarial losses, respectively. We have used $\lambda_{MS} = 0.95$ and $\lambda_{ADV} = 0.05$ similar to that was done in [11].

The learning rates are constant at 1e-6 and 1e-5 for the generator and discriminator, respectively. The adversarial training for 300000 steps took about 5 days.

3.2 Compression of Predictive Frame Differences

Predictive coding is a common practice to exploit temporal redundancy in video compression. The standard method of exploiting temporal redundancy in the video compression community is motion compensation, where motion vectors computed at the encoder between variable sized blocks of the current (original) frame and a reconstructed reference frame are sent to the decoder as side information. While this techniques has been very successful for over three decades, it has some drawbacks: 1) Rate-distortion optimization to determine the best block partitioning and estimation of motion vectors requires significant computational resources at the encoder, which makes real-time encoding a challenge, and 2) sending motion vectors as side information incurs a bitrate overhead, which is significant in low bitrate video coding.

We propose to replace the motion compensation module in predictive video compression with learned frame prediction using a neural network. More specifically, we propose to encode the difference image between the ground-truth and predicted frames using a still image encoder. In this thesis, we employed the Better Portable Graphics (BPG) still image codec [32] for this purpose.

3.2.1 Encoder

The block diagram of the proposed video encoder is depicted in Figure 3.3. The first K frames are input to the BPG encoder as I pictures without prediction. Starting with frame $K + 1$, the predicted frame is subtracted from the original frame, and the difference is encoded using the BPG encoder. In order to input exactly the same past frames into the neural networks in the encoder and decoder, the encoder has a BPG decoder in the feedback loop which produces the decoded frame difference. The decoded differences are added to the next frame predictions in order to produce decoded frames that are identical to the ones at the decoder, which become the input to the neural network for frame prediction at the next time step.

3.2.2 Decoder

The decoder also runs a neural network with the same model parameters to predict the next frame given the previous K decoded frames. The block diagram of the proposed video decoder is presented in Figure 3.4. For the first K frames, the decoder receives compressed full frames and the output of BPG decoder becomes the decoded image. After the first K frames, the decoder receives the encoded frame differences, decodes them using the BPG decoder, and adds them to the predictions of the next frames generated by the neural network. In our experiments we used $K = 8$. Since the neural network model is sent to the decoder only once, only the encoded frame differences are transmitted and the proposed video codec is free from motion vector overhead.

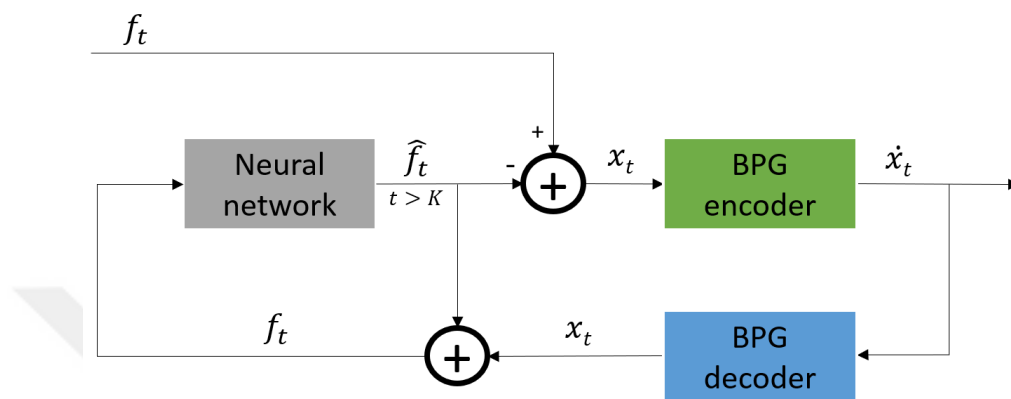


Figure 3.3: The block diagram of the proposed encoder.

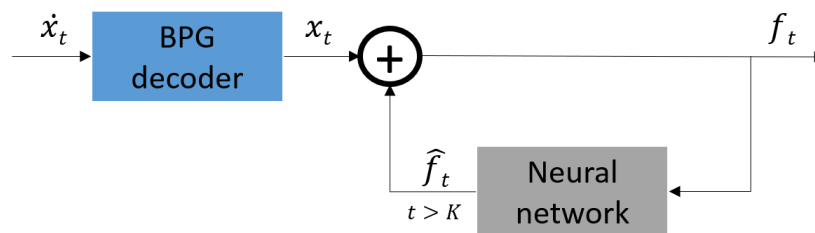


Figure 3.4: Block diagram of the decoder.

Chapter 4

EXPERIMENTAL RESULTS

This chapter presents experimental results to evaluate the performance of the proposed learned frame prediction method and its use for predictive video compression. We discuss the experimental setting and our evaluation methodology in Section 4.1. Learned frame prediction results are presented in Section 4.2. We provide the configuration parameters of the baseline video codecs used for comparison in Section 4.3. Finally, Section 4.4 presents our predictive video coding results and compare them with the results of the baseline video codecs.

4.1 *Experimental Setting and Evaluation Methodology*

Our test dataset consists of 8 MPEG test sequences, shown in Table 4.1. We converted all videos into grayscale (Y-channel), which are stored in lossless png format.

Table 4.1: MPEG test sequences.

Sequence	Pixels x Lines	Number of frames	Frames per sec.
Coastguard	352 x 288	300	25
Container	352 x 288	300	25
Football	352 x 240	125	25
Foreman	352 x 288	300	25
Garden	352 x 240	115	25
Hall Monitor	352 x 288	300	25
Mobile	352 x 240	140	25
Tennis	352 x 240	112	25

We first perform next frame prediction for all frames (except the first 8 frames) of these sequences using our two learned models, one trained with mean square loss only, and the other trained with a combination of mean square and adversarial losses. While the training is patch based, we perform full frame prediction during test time. The prediction results are evaluated in terms of the PSNR, defined by

$$PSNR = 10 \log_{10} \left(\frac{255^2}{L_{MS}} \right) \quad (4.1)$$

where L_{MS} denotes the mean square difference between learned frame prediction results and the original (ground-truth) frames as defined in Equation. 3.1. The PSNR of consecutive frame differences (without any prediction) and PSNR of predicted frames using 16×16 block motion compensation are used as baselines to compare the performance of our learned prediction models.

We then perform lossy predictive video compression on Y-channels of these test sequences using our two learned frame prediction models (instead of the standard motion compensation) in fixed quantization parameter (QP), i.e. variable bitrate (VBR) setting. For comparison, we compress the same videos using x264 codec library [33] as a baseline. Since our networks perform sequential prediction and compression, we configured the x264 codec for both sequential (IPP...) and hierarchical coding as well as both CBR and VBR modes for fair comparison. In order to obtain PSNR vs. bitrate curves for all videos, we used integer QPs ranging from 25 to 35 in the VBR mode. For CBR coding using x264 codec, we employed the rate-control option. Employing BPG to encode the residual images, we also performed compression with our baseline frame prediction methods, 16×16 block motion compensation and frame difference.

In summary, we compare performance of the following predictive coding settings:

- Learned frame prediction trained with mean square loss only (LFP-MSE)
- Learned frame prediction trained with a combination of mean square and adversarial losses (LFP-GAN)
- Sequential (IPP...) x264 with variable bitrate (SEQ-VBR)

- Hierarchical x264 with variable bitrate (HIE-VBR)
- Sequential (IPP...) x264 with constant bitrate (SEQ-CBR)
- Hierarchical x264 with constant bitrate (HIE-CBR)
- 16x16 block motion compensation (MC)
- Frame difference (FD)

To evaluate the performance of these methods, we compute rate-distortion (PSNR vs. bitrate) curves for all methods and videos. We use the *PSNR*, defined by Eqn. 4.1, as the distortion metric, where L_{MS} denotes the mean square error, as defined in Eqn. 3.1, between the decompressed video (after adding the encoded prediction error to the predicted frames) and the original (ground-truth) frames.

Since we employ the BPG codec to encode predictive frame differences in a fixed QP setting in our method, we compute the resulting average *bitrate* per second for each video by multiplying the compressed file size for each difference image with the frame rate and then averaging the resulting bits/second values over the sequence. The output of x264 encoder is a single file. In order to compute its bitrate, we divide the file size by the number of frames and multiply the result with the frame rate.

We compare rate distortion curve of each method with that of the anchor method using the Bjontegaard delta (BD-PSNR) metric [34], which measures the difference of areas between two PSNR vs. bitrate curves, as shown in Figure 4.1. The bitrates are expressed in logarithmic scale and continuous curves are formed by fitting a third order polynomial to available data points. The limits of integration are set as the higher one of the minimum bitrates and the lower one of the maximum bitrates. Area under the curve of the anchor method is subtracted from the area under the curve of the each method; hence, a positive value means the method performs better than the anchor.

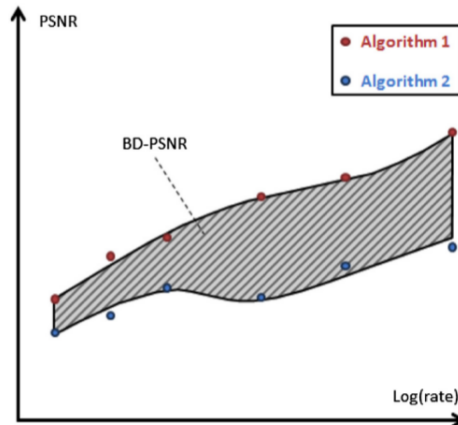


Figure 4.1: Computation of the Bjontegaard delta PSNR (BD-PSNR) metric [1].

4.2 Frame Prediction Results

This section presents experimental results for both quantitative and qualitative evaluation of our two learned video frame prediction models: the LFP-MSE and LFP-GAN models without considering the effect of compression. That is, in this section all prediction results are obtained by using original (uncompressed) previous frames.

We first perform quantitative evaluation in terms of prediction PSNR results by comparing each output frame obtained by our models (generator network) to the corresponding frame of the original video sequence (ground-truth). In Figure 4.2, these values are plotted against the frame number for each video. We compare our methods to two baselines. The first baseline method is motion compensation (MC) with 16x16 blocks (patches), half pixel precision and exhaustive search with a motion vector limit of 31. The second baseline method is frame difference (FD), that is using the current frame as a prediction for the next frame.

As seen in Figure 4.2, in each plot, both curves follow the same trends, with LFP-GAN producing inferior results. Since PSNR is inversely proportional to mean square error it is expected that training to minimize mean square error *only* would produce the highest PSNR values. To train the LFP-GAN, we include the additional term of adversarial loss, hence we sacrifice mean square error for adversarial loss. As

a result of this, the residual images for LFP-GAN have more information, namely higher entropy, so the compressed residual images have larger file sizes.

It can also be seen that, in container, a video with smooth surfaces and very slow motion, the gap between LFP-MSE and LFP-GAN is higher. This can be due to the fact that LFP-GAN aims at creating sharp patterns even when the original frame has smooth surfaces. Moreover, since the video has slow motion, LFP-MSE does not yield a high mean square error, hence LFP-GAN's attempt at reducing adversarial loss at the cost of increasing mean square loss becomes a dominating factor.

We further observe that FD performs the worst for all videos. MC performs the best only at the presence of fast motion, such as the entire football video and parts of coastguard, foreman and tennis videos. Especially in football, a video with fast and complex motion, MC is capable of dealing with this complexity by relocating blocks of pixels independently. It is noted that MC algorithm has access to the target frame and it always relocates the block which yields the highest PSNR value.

Next, we present our qualitative results for the prediction of the 9th frames for each video. Through Figures 4.3 to 4.10, LFP-MSE (left) and LFP-GAN's (right) outputs can be seen on the upper row. Their difference with respect to the original frames are presented in the lower row. While the gray levels mean low difference, white and black levels denote high positive and negative difference. Despite its lower PSNR values, LFP-GAN produces sharper and more realistic looking images. In a detailed observation, it is seen that LFP-GAN produces higher error especially at the regions where LFP-MSE generates blurriness while LFP-GAN produces sharper surfaces. We can say that, in these regions, LFP-GAN takes a more stochastic approach.

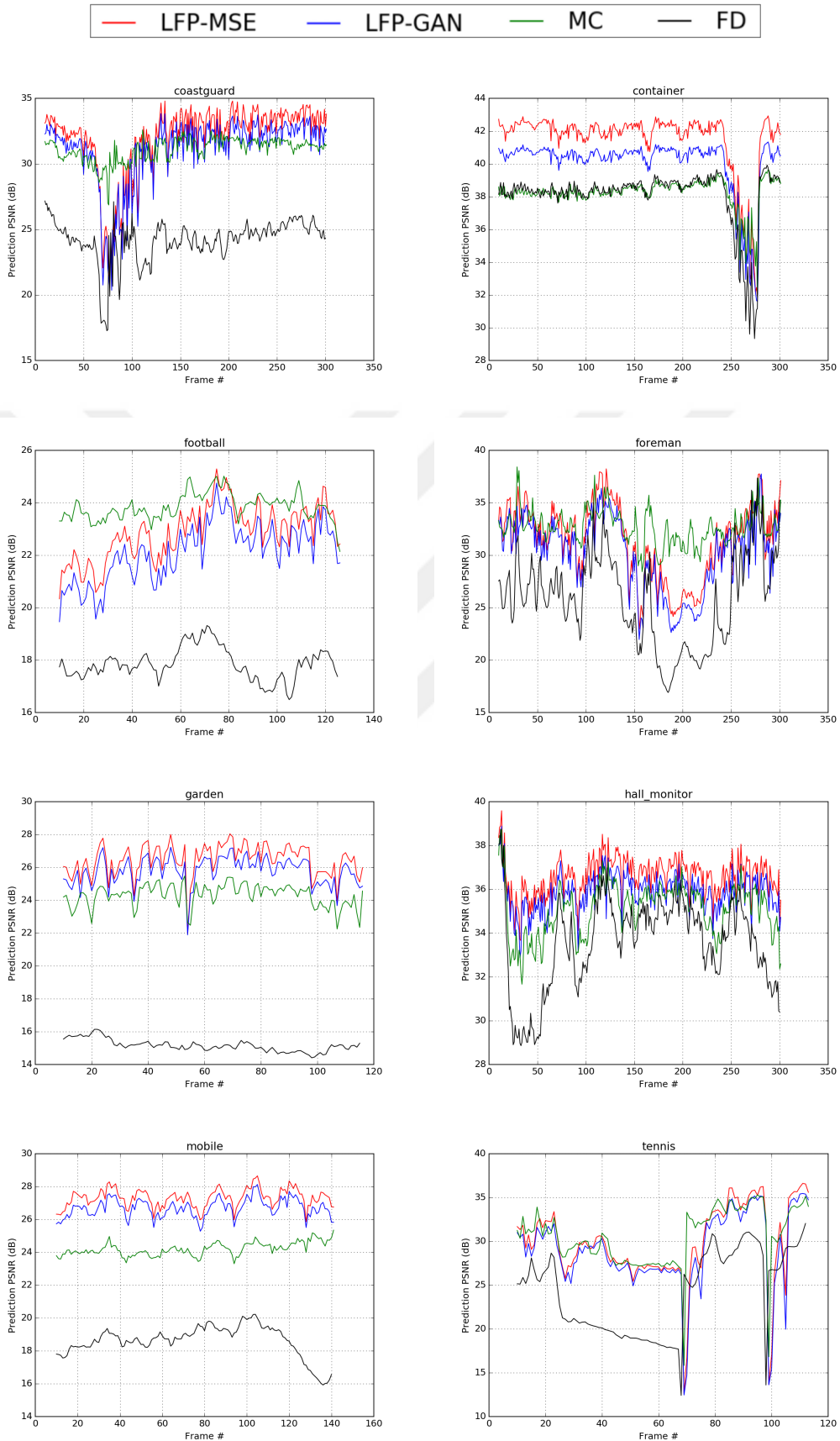


Figure 4.2: PSNR of predicted frames vs. frame number for all test videos.

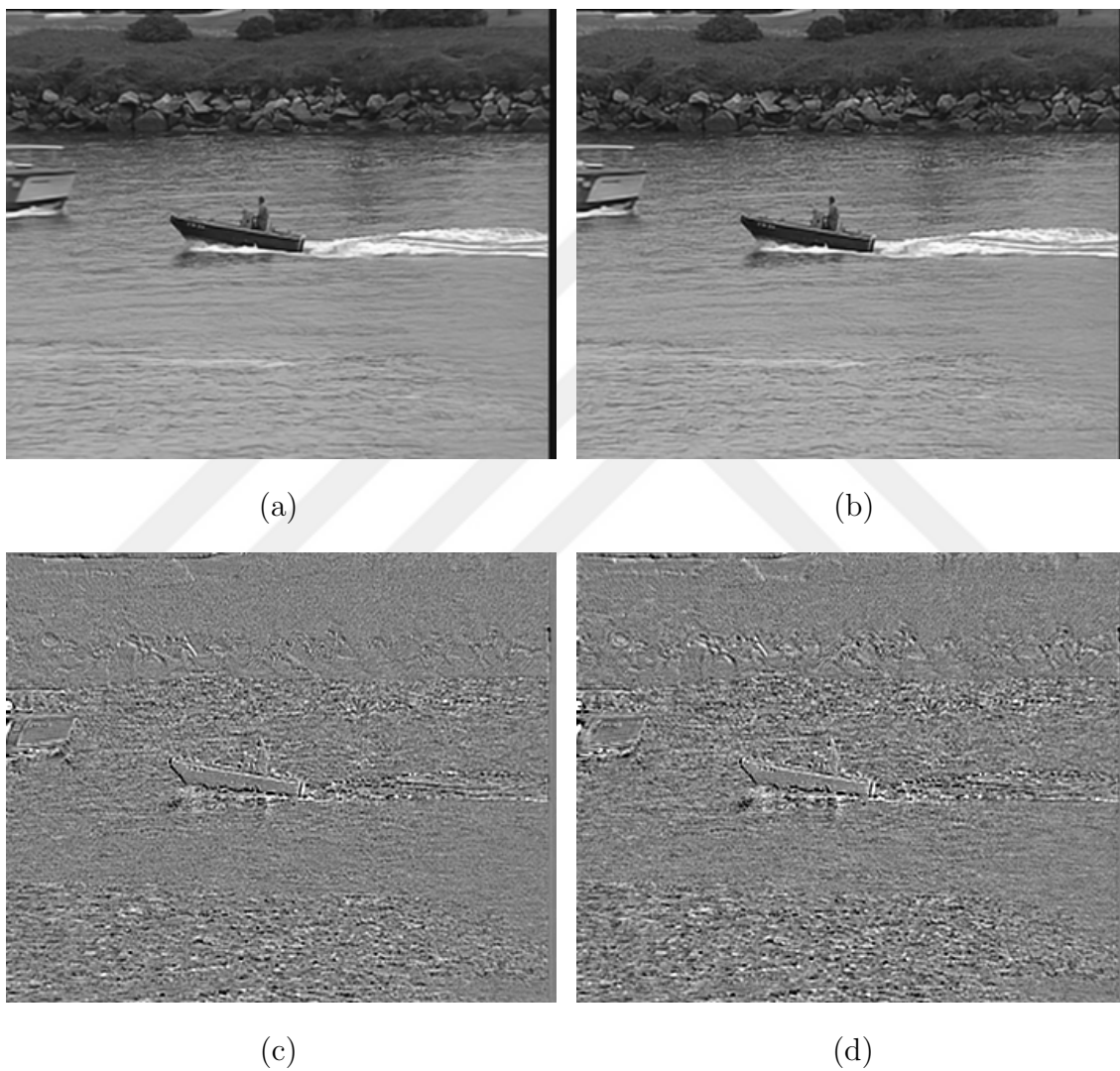


Figure 4.3: Coastguard. (a) Predicted frame using LFP-MSE method, (b) Predicted frame using LFP-GAN method, (c) Frame difference of LFP-MSE, (d) Frame difference of LFP-GAN. Difference values are scaled by 5 and level shifted by 128.

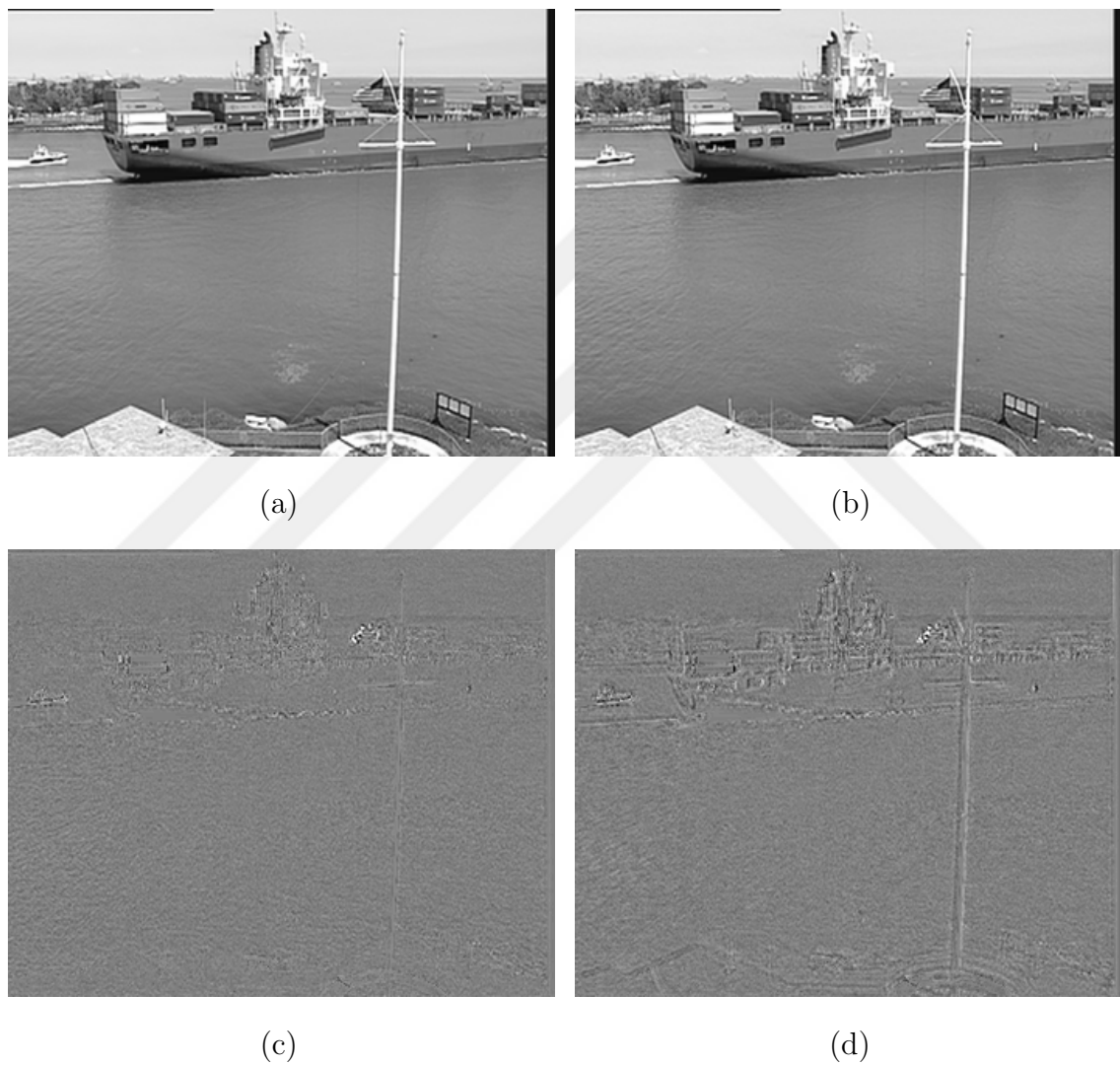


Figure 4.4: Container. (a) Predicted frame using LFP-MSE method, (b) Predicted frame using LFP-GAN method, (c) Frame difference of LFP-MSE, (d) Frame difference of LFP-GAN. Difference values are scaled by 5 and level shifted by 128.

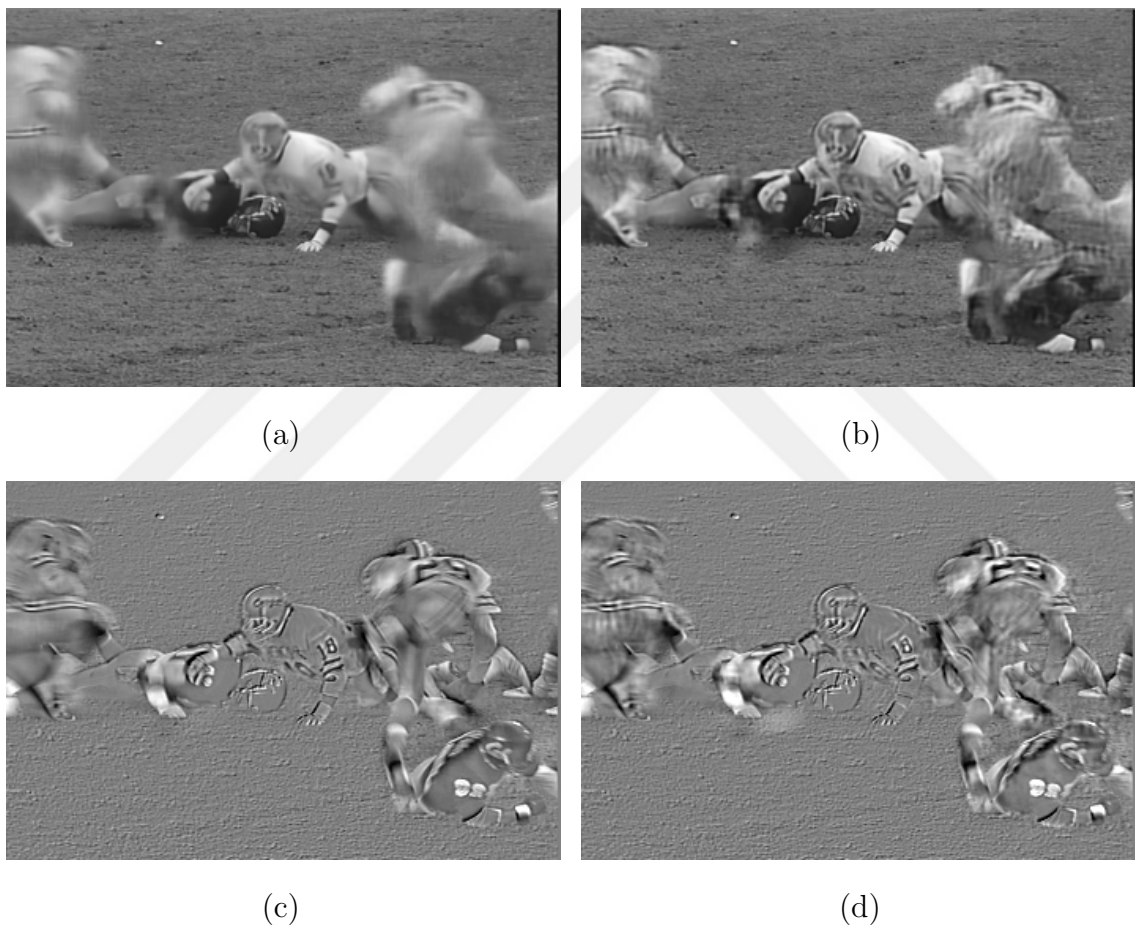


Figure 4.5: Football. (a) Predicted frame using LFP-MSE method, (b) Predicted frame using LFP-GAN method, (c) Frame difference of LFP-MSE, (d) Frame difference of LFP-GAN. Difference values are not scaled. They are level shifted by 128.

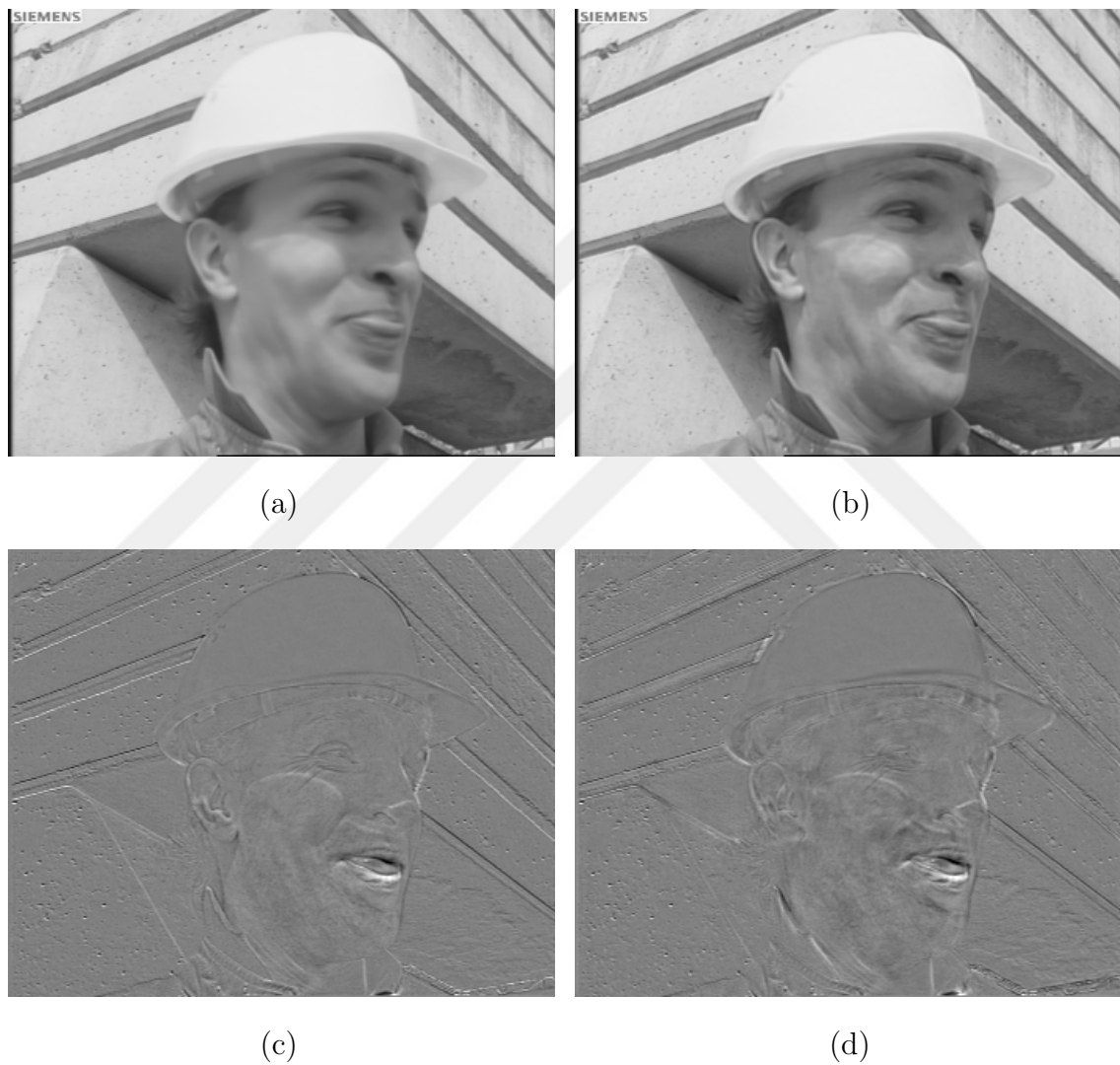


Figure 4.6: Foreman. (a) Predicted frame using LFP-MSE method, (b) Predicted frame using LFP-GAN method, (c) Frame difference of LFP-MSE, (d) Frame difference of LFP-GAN. Difference values are scaled by 2 and level shifted by 128.

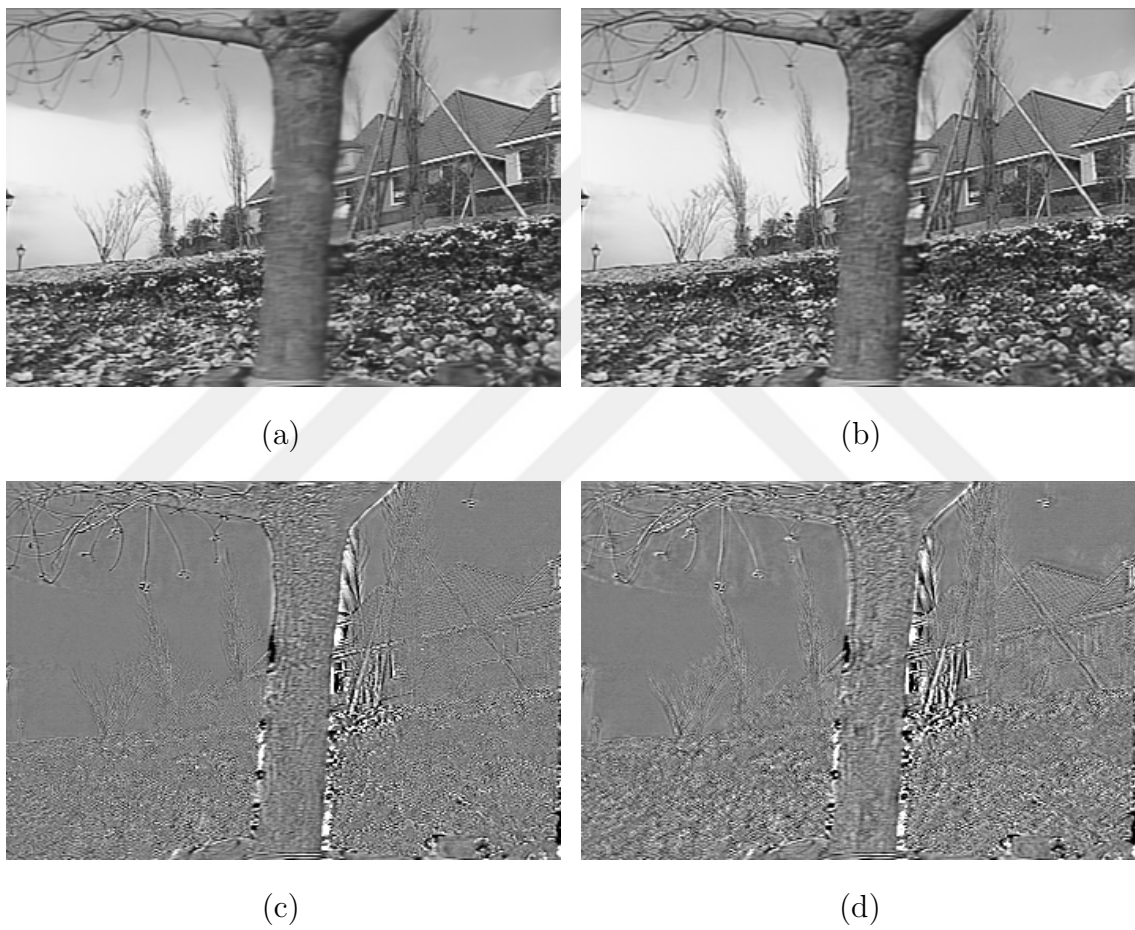


Figure 4.7: Garden. (a) Predicted frame using LFP-MSE method, (b) Predicted frame using LFP-GAN method, (c) Frame difference of LFP-MSE, (d) Frame difference of LFP-GAN. Difference values are scaled by 3 and level shifted by 128.

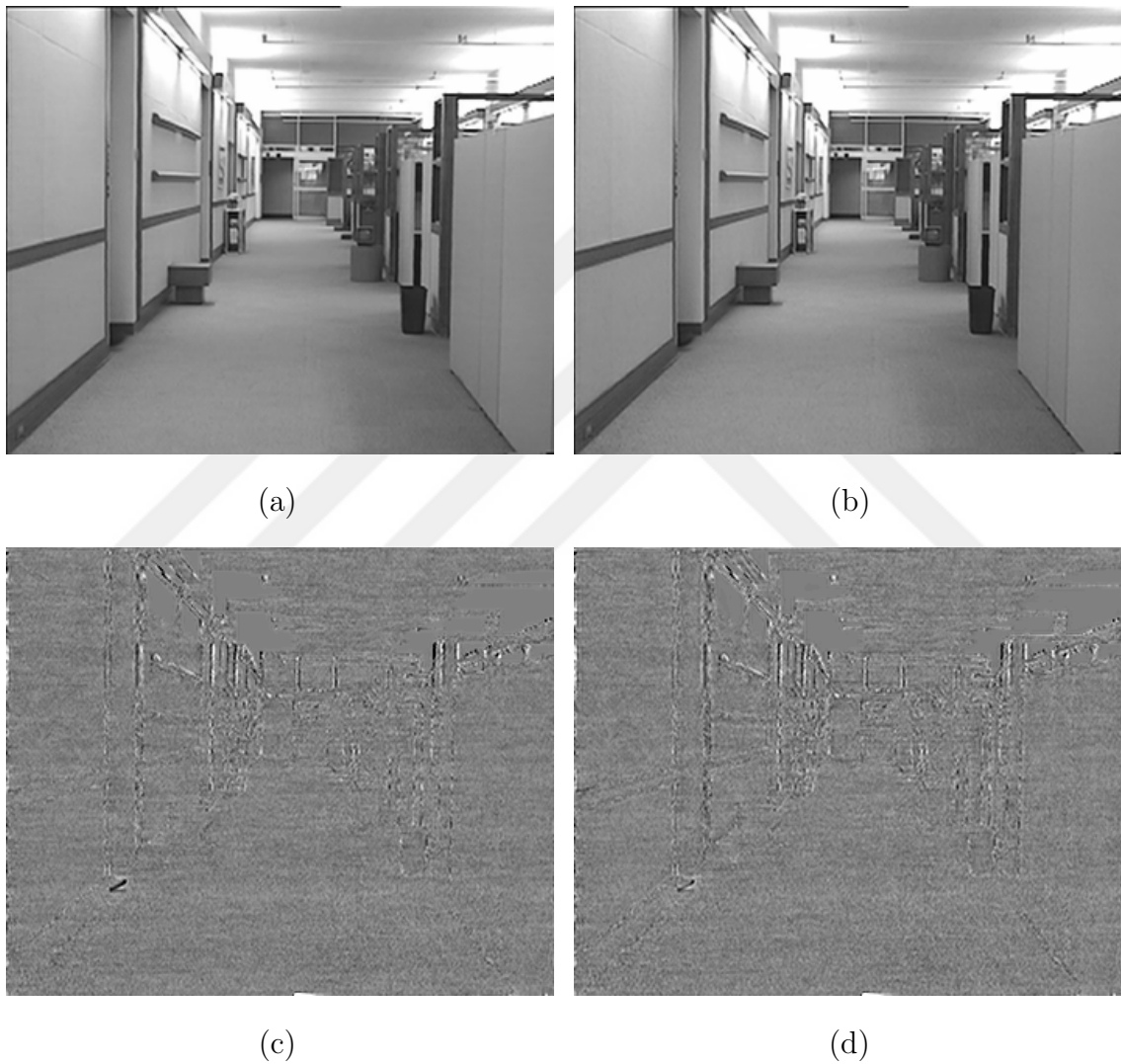


Figure 4.8: Hall Monitor. (a) Predicted frame using LFP-MSE method, (b) Predicted frame using LFP-GAN method, (c) Frame difference of LFP-MSE, (d) Frame difference of LFP-GAN. Difference values are scaled by 5 and level shifted by 128.

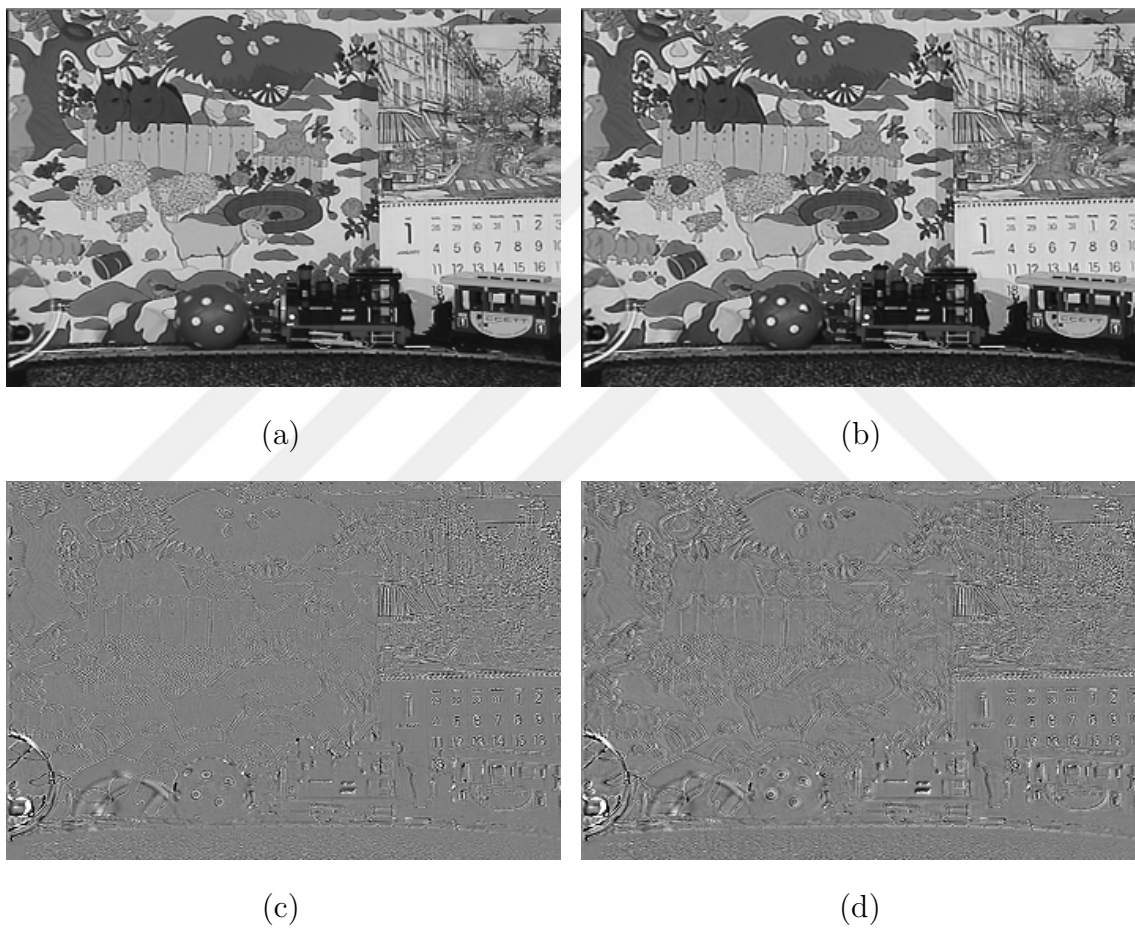


Figure 4.9: Mobile. (a) Predicted frame using LFP-MSE method, (b) Predicted frame using LFP-GAN method, (c) Frame difference of LFP-MSE, (d) Frame difference of LFP-GAN. Difference values are scaled by 2 and level shifted by 128.

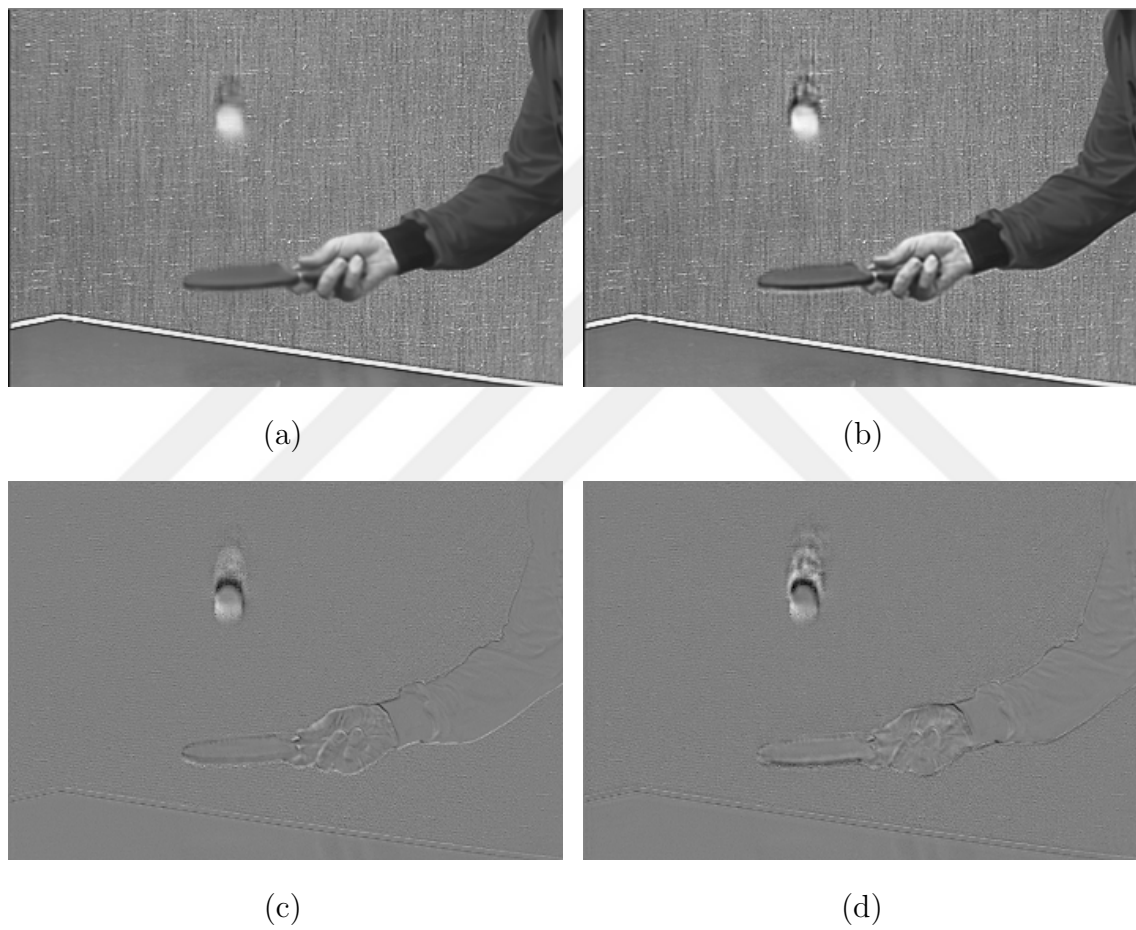


Figure 4.10: Tennis. (a) Predicted frame using LFP-MSE method, (b) Predicted frame using LFP-GAN method, (c) Frame difference of LFP-MSE, (d) Frame difference of LFP-GAN. Difference values are not scaled. They are level shifted by 128.

4.3 Baseline Video Codec Configurations

We use the x264 codec supported by the FFmpeg library [35] as baseline for comparison with our compression results. FFmpeg provides the following options:

- i : Input file
- c:v : Video codec
- qp : Quantization parameter
- preset : Encoding preset (slower is better)
- r : Frame rate
- f : Output format
- g : Maximum keyframe interval
- keyint_min : Minimum keyframe interval
- bf : Number of B-frames
- b:v : Video bitrate
- maxrate : Maximum bitrate
- minrate : Minimum bitrate
- bufsize : Rate control buffer size
- x264-params nal-hrd=cbr : Set network abstraction layer's (NAL) hypothetical reference decoder (HRD) to constant bitrate (CBR). Note that the output format needs to be ts, since mp4 format does not support this option [36].
- x264-params force-cfr=1 : Force constant framerate timestamp generation
- pix_fmt : Pixel format

4.3.1 x264 Hierarchical Coding

The standard (default) usage of the x264 codec, which usually yields the best results, is the hierarchical motion compensation (prediction) with variable bitrate (VBR) compression. The command line we used to encode the sequence "Coastguard" with QP value of 25 in HIE-VBR mode is provided as an example.

```
ffmpeg -i coastguard/%3d.png -c:v libx264rgb -qp 25  
-preset veryslow -r 25 -f mp4 coastguard.mp4
```

Alternatively, we can perform hierarchical prediction in constant bitrate (CBR) mode at a prespecified bitrate. The command line for x264 default CBR (HIE-CBR) coding with the bitrate of 1000k and buffer size of 2000k is provided as an example.

```
ffmpeg -i coastguard/%3d.png -c:v libx264rgb
-x264-params "nal-hrd=cbr:force-cfr=1" -b:v 1000k
-maxrate 1000k -minrate 1000k -bufsize 2000k
-preset veryslow -r 25 coastguard.ts
```

Note that FFmpeg documentation recommends a buffer size of 1-2 times the maximum bitrate [37].

4.3.2 x264 Sequential Coding

Because our learned models work sequentially, we also used x264 coding with sequential motion prediction, i.e., using a single I frame followed by P frames only.

The command line for coding using x264 sequential VBR (SEQ-VBR) mode is

```
ffmpeg -i coastguard/%3d.png -c:v libx264rgb -g 1000
-keyint_min 1000 -bf 0 -qp 25 -preset veryslow -r 25
-f mp4 coastguard.mp4
```

Since the longest test video consists of 300 frames, by selecting 1000 as the minimum and maximum keyframe interval and zero as the B-frame number, we guarantee sequential motion estimation. This is verified through FFmpeg's output log.

Finally, the command line for x264 sequential CBR (SEQ-CBR) is as follows:

```
ffmpeg -i coastguard/%3d.png -c:v libx264rgb -g 1000
-x264-params "nal-hrd=cbr:force-cfr=1" -b:v 1000k
-maxrate 1000k -minrate 1000k -bufsize 2000k -r 25
-preset veryslow -keyint_min 1000 -bf 0 coastguard.ts
```

For all coding modes, after each command line, the following command line is used to decode the video into grayscale png frames:

```
ffmpeg -i coastguard.mp4 -pix_fmt gray
coastguard_decoded/%3d.png
```

4.4 Predictive Video Compression Results

This section evaluates the use of learned prediction models (instead of standard motion compensation) for predictive video compression. Naturally, in this setting all frame prediction results are obtained by using compressed/decompressed previous frames, since frame prediction must also be duplicated at the receiver.

Using our neural networks, we present our compression performance compared to the x264 codec, MC and FD. In MC and FD, BPG is used to encode the residual image, just like in LFP methods. Only for x264, the motion vectors are extracted from the encoded videos, using the software *mpegflow* [38]. Bitrate of the motion vectors is obtained using their entropy. Note that there are no motion vectors for compression with neural networks. In Figure 4.11 rate-PSNR curves can be seen on the left. On the right we display the bitrate of the motion vectors of x264 codec versus total bitrate.

It is seen that in videos with high motion content, methods using BPG with LFP or 16x16 MC outperform x264. One of the reasons can be that, in these videos, residual images have more content, meaning higher entropy, and BPG is doing such a better job compressing these residual images that advanced motion compensation method of x264 is not sufficient to outperform the methods using BPG.

In videos with slow motion, the entropies of the residual images are low, so the superior motion compensation capability of x264 becomes an advantage over 16x16 MC. Also, for these videos, LFP-GAN performs much worse, as explained in Section 4.2. However, for all videos, our method of LFP-MSE remains competitive.

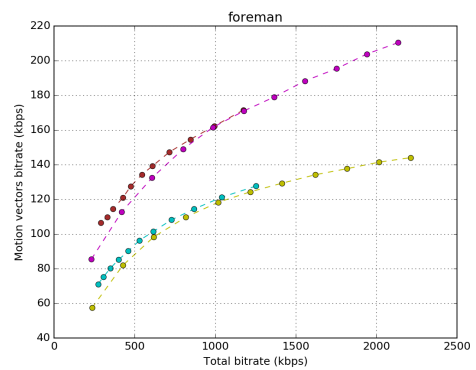
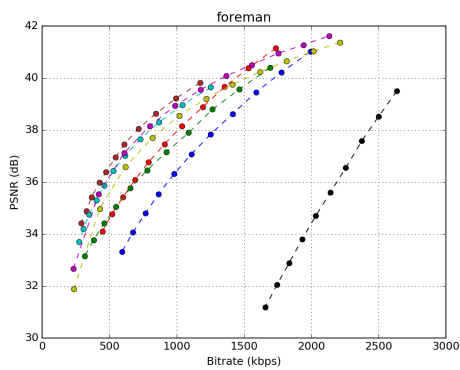
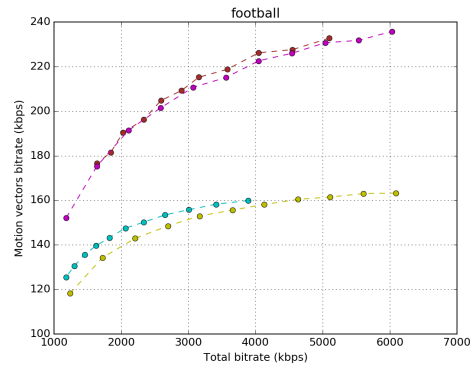
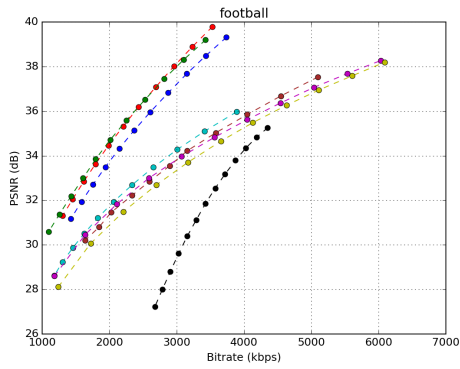
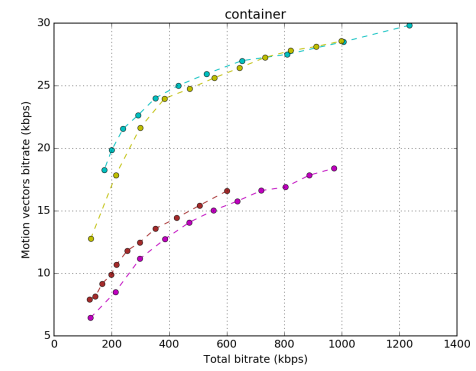
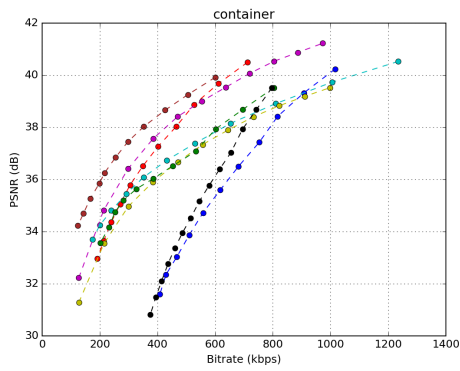
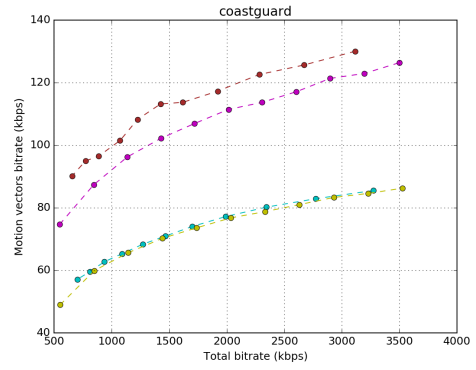
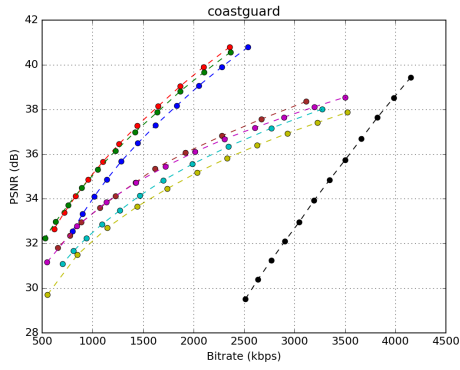
In almost all videos, FD performs the worst. Here, the efficiency of BPG is not sufficient to compensate for the high entropy residual images.

Finally, to summarize our results, we show the Bjontegaard delta values calculated against the anchor method, in Table 4.2. Since our models work sequentially with no limitation on the bitrate, we have chosen sequential x264 codec with variable bitrate (SEQ-VBR), as the anchor method for a fair comparison.

It is observed that even though it generates sharper and more realistic images,

GAN training worsens compression performance for all videos. Our method LFP-MSE, outperforms the anchor in 5 videos out of 8. In 2 of the 3 videos that LFP-MSE is outperformed by the anchor, the margin is very small. Furthermore in 4 videos, it outperforms HIE-VBR, the best possible configuration in x264 codec.





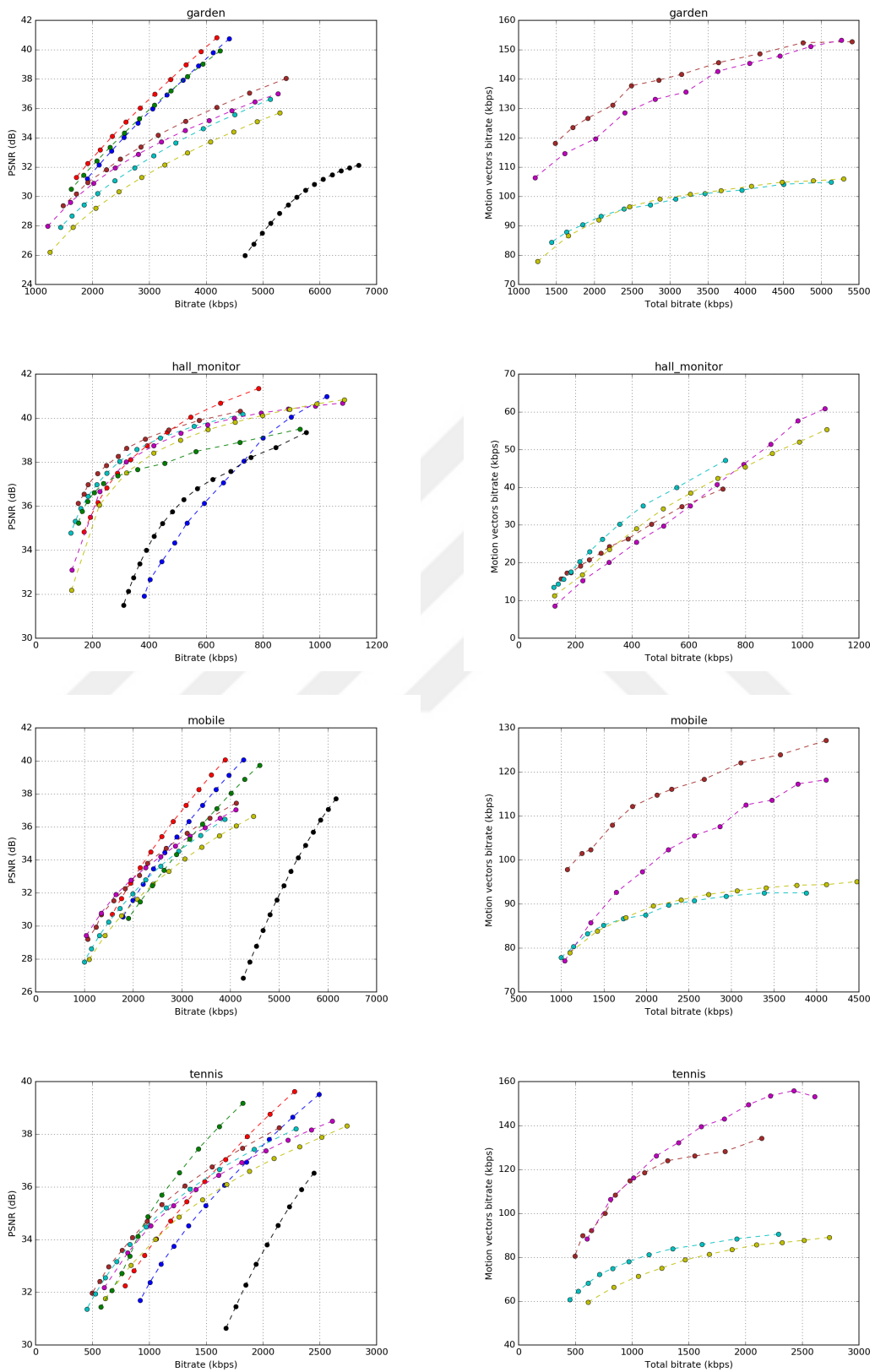


Figure 4.11: Compression performance on test videos. Left: PSNR vs. bitrate. Right: Motion vector bitrate vs. total bitrate

Table 4.2: Bjontegaard delta PSNR metrics with respect to x264 SEQ-VBR (anchor). Positive values indicate higher performance compared to the anchor.

	LFP MSE	LFP GAN	HIE VBR	HIE CBR	SEQ CBR	MC	FD
Coastguard	3.103	2.410	0.745	0.642	-0.429	2.982	-5.439
Container	0.567	-2.070	1.897	1.115	-0.461	-0.175	-2.356
Football	3.035	2.371	-0.432	-0.265	-0.852	3.004	-4.041
Foreman	-1.197	-2.732	0.435	0.079	-0.526	-1.372	-11.118
Garden	3.739	3.150	1.168	0.711	-0.996	3.021	-9.173
Hall monitor	-0.200	-4.434	0.318	-0.501	-0.998	-0.763	-3.919
Mobile	1.664	0.656	0.823	0.840	-0.624	-0.216	-11.428
Tennis	-0.220	-1.094	0.231	-0.204	-0.776	0.361	-4.342

Chapter 5

CONCLUSION

In this thesis we have firstly introduced a deep convolutional neural network for the task of video frame prediction. We have trained this model to minimize mean square loss and obtained accurate frame predictions.

We have also employed adversarial training and generated sharper and more realistic images. We have shown that, though GANs perform better qualitatively, the PSNR values of their outputs are lower. We have further demonstrated that adversarial training deteriorates compression performance, hence it is redundant for video compression.

Using our neural networks, we have built encoders and decoders for video compression. These encoders do not require the calculation or transmission of motion vectors, so the size of the transferred data is less compared to the standard codecs. Moreover, we have compared our method to x264 codec, a benchmark for video compression. We have outperformed the anchor method of sequential x264 with variable bitrate, in standard quantitative metrics, for more than half of the test videos. For half of the videos, we have outperformed the x264 codec with best possible configuration, namely hierarchical prediction with variable bitrate.

While the usage of deep neural networks is a relatively recent practice, it sets new standards in common engineering problems every day. There isn't a standard codec that employs a neural network yet, but it is very probable in the near future. To obtain better compression performance using neural networks, bidirectional prediction is a topic that should be investigated upon. Also the compression of the neural network itself to obtain a smaller network with shorter runtime can enable faster video compression. The creation of a standalone video compression library that uses

neural networks can be the final step to create widespread usage in video compression with neural networks.



BIBLIOGRAPHY

- [1] D. Blinder, T. Bruylants, H. Ottevaere, A. Munteanu, and P. Schelkens, “Jpeg 2000-based compression of fringe patterns for digital holographic microscopy,” *Optical Engineering*, vol. 53, no. 12, p. 123102, 2014.
- [2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [3] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [5] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, vol. 1, no. 2, 2017, p. 4.
- [6] R. Timofte, E. Agustsson, L. Van Gool *et al.*, “NTIRE 2017 challenge on single image super-resolution: Methods and results,” in *IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1110–1121.
- [7] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.

- [8] J. Selva Castelló, “A comprehensive survey on deep future frame video prediction,” Master’s thesis, Universitat Politècnica de Catalunya, 2018.
- [9] V. Ge, “A survey on deep video prediction,” *learning*, vol. 59, p. 68, 2017.
- [10] N. Srivastava, E. Mansimov, and R. Salakhudinov, “Unsupervised learning of video representations using LSTMs,” in *Int. Conf. on Machine Learning*, 2015, pp. 843–852.
- [11] M. Mathieu, C. Couprie, and Y. LeCun, “Deep multi-scale video prediction beyond mean square error,” *arXiv:1511.05440*, 2015.
- [12] N. Kalchbrenner, A. v. d. Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu, “Video pixel networks,” *arXiv:1610.00527*, 2016.
- [13] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, “Pixel recurrent neural networks,” *arXiv preprint arXiv:1601.06759*, 2016.
- [14] F. Cricri, X. Ni, M. Honkala, E. Aksu, and M. Gabbouj, “Video ladder networks,” *arXiv:1612.01756*, 2016.
- [15] J. Van Amersfoort, A. Kannan, M. Ranzato, A. Szlam, D. Tran, and S. Chintala, “Transformation-based models of video sequences,” *arXiv:1701.08435*, 2017.
- [16] C. Vondrick and A. Torralba, “Generating the future with adversarial transformers,” in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, no. 2, 2017, p. 3.
- [17] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee, “Learning to generate long-term future via hierarchical prediction,” *arXiv:1704.05831*, 2017.

- [18] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.
- [19] S. E. Reed, Y. Zhang, Y. Zhang, and H. Lee, “Deep visual analogy-making,” in *Advances in neural information processing systems*, 2015, pp. 1252–1260.
- [20] A. Dosovitskiy and T. Brox, “Generating images with perceptual similarity metrics based on deep networks,” in *Advances in Neural Information Processing Systems*, 2016, pp. 658–666.
- [21] N. Wichers, R. Villegas, D. Erhan, and H. Lee, “Hierarchical long-term video prediction without supervision,” *arXiv:1806.04768*, 2018.
- [22] C. Finn, I. Goodfellow, and S. Levine, “Unsupervised learning for physical interaction through video prediction,” in *Advances in neural information processing systems*, 2016, pp. 64–72.
- [23] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine, “Stochastic variational video prediction,” *arXiv:1710.11252*, 2017.
- [24] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine, “Stochastic adversarial video prediction,” *arXiv:1804.01523*, 2018.
- [25] E. Denton and R. Fergus, “Stochastic video generation with a learned prior,” *arXiv:1802.07687*, 2018.
- [26] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning.” in *AAAI*, vol. 4, 2017, p. 12.
- [27] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, “Photo-realistic single image super-

- resolution using a generative adversarial network.” in *CVPR*, vol. 2, no. 3, 2017, p. 4.
- [28] S. Chintala, E. Denton, M. Arjovsky, and M. Mathieu, “How to train a gan? tips and tricks to make gans work,” Dec 2016. [Online]. Available: <https://github.com/soumith/ganhacks>
- [29] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv preprint arXiv:1511.06434*, 2015.
- [30] K. Soomro, A. R. Zamir, and M. Shah, “Ucf101: A dataset of 101 human actions classes from videos in the wild,” *arXiv preprint arXiv:1212.0402*, 2012.
- [31] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [32] F. Bellard, “Better portable graphics,” *BPG*. [Last accessed: November 2018]. <https://www.bellard.org/bpg>, 2014.
- [33] L. Merritt and R. Vanam, “x264: A high performance h. 264/avc encoder,” [online] http://neuron2.net/library/avc/overview_x264_v8_5.pdf, 2006.
- [34] G. Bjontegaard, “Calculation of average psnr differences between rd-curves,” *VCEG-M33*, 2001.
- [35] F. Bellard, “Ffmpeg multimedia system,” *FFmpeg*. [Last accessed: November 2018]. <https://www.ffmpeg.org/about.html>, 2005.
- [36] W. Robitza, “Understanding rate control modes,” Mar 2017. [Online]. Available: <https://slhck.info/video/2017/03/01/rate-control.html>
- [37] FFmpeg, “Encoding for streaming sites.” [Online]. Available: <https://trac.ffmpeg.org/wiki/EncodingForStreamingSites>

- [38] V. Kantorov and I. Laptev, “Efficient feature extraction, encoding and classification for action recognition,” in *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE, 2014*, 2014.

