Learned Video Compression with Conditional Augmented Normalizing Flows

Chih-Hsuan Lin[1] Yung-Han Ho[1] Mu-Jung Chen[1] Wen-Hsiao Peng[1] Hsueh-Ming Hang[2]
meerkat10.cs09g@nctu.edu.tw hectorho0409.cs04g@nctu.edu.tw mujung.cs10@nycu.edu.tw wpeng@cs.nctu.edu.tw hmhang@nctu.edu.tw

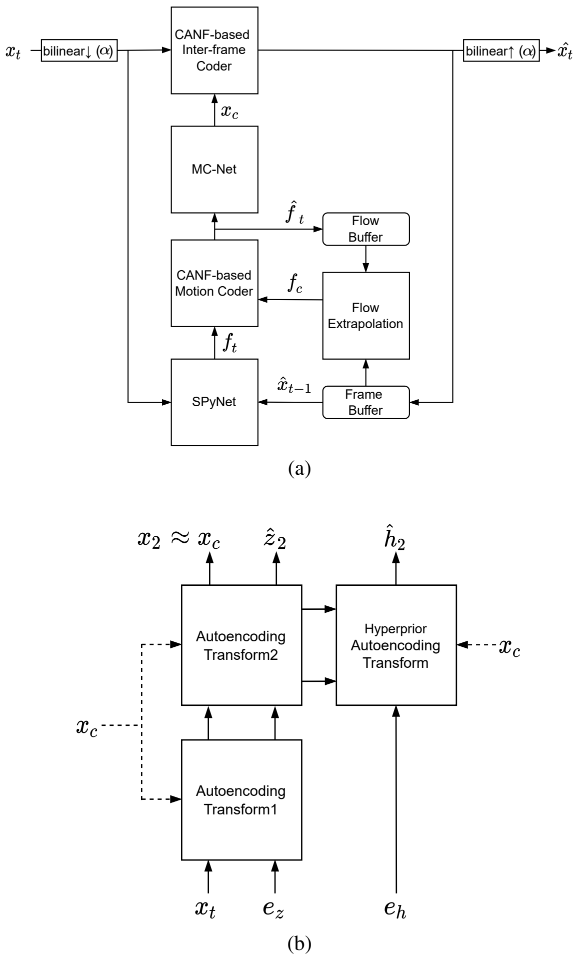
[1]Computer Science Dept., [2]Electronics Engineering Dept., National Yang Ming Chiao Tung University, Taiwan

Abstract

In response to 2022 CLIC Learned Video Compression Challenge, we submit a learned video compression scheme based on conditional augmented normalizing flows (CANF). Motivated by augmented normalizing flow-based image compression (ANFIC), this proposal introduces conditional augmented normalizing flows to encode every p-frame conditionally based on its motion-compensated reference frame. CANF is a conditional coding scheme, which is utilized in place of the conventional residual coding. A separate CANF is deployed for motion coding, where a flow map extrapolation network is adopted to extrapolate a flow map that serves as a condition for motion coding. To address low-rate compression, our CANF-based coding framework includes image downscaling and upscaling as pre- and post-processing steps.

1. System Overview

Fig. 1a depicts our CANF-based video compression system. It includes two major components: (1) the CANF-based inter-frame coder and (2) the CANF-based motion coder. The inter-frame coder encodes a video frame x_t conditionally, given the motion-compensated frame x_c. It departs from the conventional residual coding by maximizing the conditional log-likelihood $p(x_t|x_c)$ with a conditional, multi-step ANF model (see Fig. 1b). The motion coder shares a similar architecture to the inter-frame coder. It extends conditional coding to motion coding, in order to signal the flow map f_t, which characterizes the motion between x_t and its reference frame $\hat{x}_{t-1}$. Next, the flow map f_t is estimated by SPyNet [4]. The compressed flow map $\hat{f}_t$ serves to warp the reference frame $\hat{x}_{t-1}$, with the warped result enhanced further by a motion compensation network (MC-Net) to arrive at x_c. To formulate a condition for conditional motion coding, we introduce a flow extrapolation network to extrapolate a flow map f_c from three previously decoded frames and two decoded flow maps. To achieve extreme low-rate compression, the input video frame may be downscaled bilinearly by a factor of α, the value of which is signaled in the bitstream. Accordingly, the decoded video frame is upscaled to obtain the final reconstructed frame.



(a)

(b)

Figure 1. Illustration of (a) the CANF-based video compression framework and (b) the CANF-based inter-frame coder.

1.1. CANF-based Inter-frame Coder

Motivated by [3], our CANF-based inter-frame coder is a hybrid of the two-step and the hierarchical ANF's, as de-

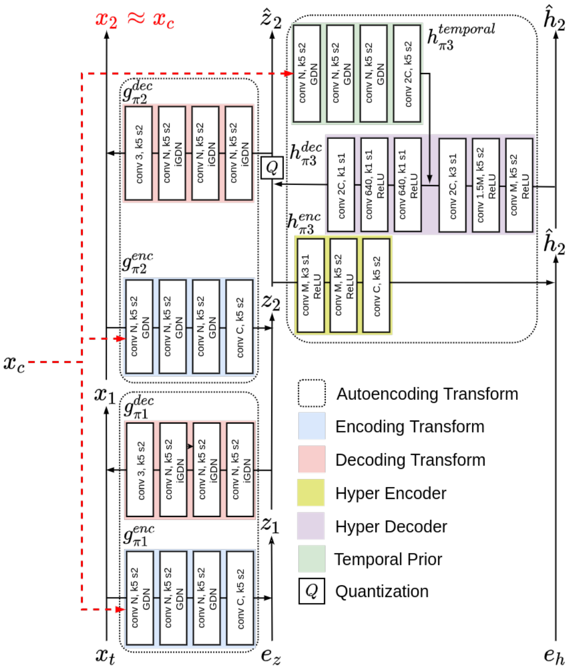


Figure 2. Network details of the CANF-based coder.

picted in Fig. 2. It takes a coding frame x_t as the main input, with e_z, e_h serving as the augmented inputs, which take on values 0 at inference time. The encoding of x_t is done by converting it into its motion-compensated reference frame x_c through the autoencoding transforms, each of which comprises an encoding and a decoding transform organized in a coupling-layer structure. The resulting latent $\hat{z}_2$ captures the information needed to signal the conversion, and $\hat{h}_2$ represents the corresponding hyperprior for entropy encoding the $\hat{z}_2$. To ease the conversion, we additionally feed x_c as a condition to the autoencoding transforms. The decoding of x_t is accomplished by updating x_t through the inverse autoencoding transforms. The CANF-based motion coder shares the same coding structure as the inter-frame coder, where the coding frame x_t is replaced with the optical flow map f_t and the motion-compensated frame x_c with the extrapolated flow map f_c. The flow map extrapolation network is a U-Net-based network.

To achieve variable-rate compression, we introduce conditional convolution (CConv) [1] to both the inter-frame and motion coders. By adjusting the hyperparameter λ of the CConv layers, our p-frame coder is able to support a moderate range of rate points. Still, two distinct p-frame coders are trained to fulfill the two rate requirements of CLIC. One operates in the rate range around 1mbps, whereas the other in the rate range at 0.1mbps. The training objective is the conventional rate-distortion cost, where we use MS-SSIM as the distoriton metric.

1.2. Training details

We train our model on Vimeo-90k [5] dataset, which contains 91,701 7-frame sequences with resolution 448×256. We randomly crop these video clips into 256×256 for training. We adopt the Adam [2] optimizer with the learning rate 10^{-4} and the batch size 32. Two separate models are trained to optimize first the Mean-square Error (MSE). We then fine-tune these models for Multi-scale Structural Similarity Index (MS-SSIM).

In particular, each of these models is able to perform variable-rate compression over a range of bit rates. This is achieved with the use of CConv layers. The end-to-end training with CConv layers is enabled after the MS-SSIM-based fine tuning.

2. Acknowledgement

This work is partially supported by the MOST, Taiwan under Grant MOST 110-2221-E-A49-065-MY3 and National Center for High-performance Computing.

References

[1] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Variable rate deep image compression with a conditional autoencoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3146–3154, 2019. 2

[2] Jimmy Ba Diederik P. Kingma. Adam: A method for stochastic optimization. *International Conference for Learning Representations*, 2015. 2

[3] Yung-Han Ho, Chih-Chun Chan, Wen-Hsiao Peng, Hsueh-Ming Hang, and Marek Domański. Anfic: Image compression using augmented normalizing flows. *IEEE Open Journal of Circuits and Systems*, 2:613–626, 2021. 1

[4] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017. 1

[5] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019. 2