GENERATIVE ADVERSARIAL NETWORKS BASED ERROR CONCEALMENT FOR LOW RESOLUTION VIDEO

Chongyang Xiang *Jiajun Xu** *Chuan Yan* *Qiang Peng* *Xiao Wu*

School of information science and technology, Southwest Jiaotong University
Chengdu 611756, China, (xiangchongyang@my.swjtu.edu.cn)

ABSTRACT

In this paper, a novel deep generative model-based approach for video error concealment is proposed. Our method is comprised of completion network and two critics. The frame completion network is trained to fool the both the local and global critics, which requires completion network to conceal frame distortions with regard to overall consistency as well as in details. Specifically, mask attention convolution layer is proposed, which utilize not only the temporal information of the previous frame, but also the intact pixels of the current distorted frame to mask and re-normalize convolution features. Then, both qualitative and quantitative experiments validate the effectiveness and generality of our approach in advancing the error concealment on low resolution video.

Index Terms— Generative Adversarial Network, HEVC, Error concealment

1. INTRODUCTION

The latest video coding standard HEVC [1] has a very high compression ratio which significantly reduce the network traffic load and bandwidth requirement. However, HEVC bit streams are very sensitive to packet error [2]. Transmission of video over packet-switched wireless networks, especially for real-time applications, is still challenging because of network congestion, delay, limited available bandwidth and error prone nature of wireless channel. When bit errors or packet losses occur, it makes the decoder unable to fully recover the video quality. Comparing to H.264/AVC [3], HEVC introduces the temporal candidate in the set of possible motion vectors increases the dependencies between subsequent frames, which leads to higher quality degradation in presence of errors. To recover the lost region, error concealment is usually utilized at the decoder, using the available information in its spatial and/or temporal neighborhood due to the spatial/temporal redundancy. However, HEVC cannot guarantee end-to-end reproduction quality and does not suggest any error concealment when the bitstream is lossy.

There is little literature on HEVC error concealment. In [4], a motion vector extrapolation based method was proposed for whole frame loss. A motion vector correlation from the co-located largest coding unit (LCU) was calculated for deciding whether to divide a large block into smaller ones or not. [5] uses weighted boundary matching (WBMA) algorithm that finds prediction unit (PU) blocks with the best matched boundaries from the reference frame to conceal the currently corrupted PU blocks. However, all of them do not consider the spatial smoothness of the lost CUs, and may make those estimated motion vectors fail to represent the true motion. Thus, these algorithms cannot reconstruct a satisfying reconstruction of a lost area and may not reduce error propagation effect. Inspired by the great success of Generative Adversarial Network (GAN) [6] and convolutional neural network (CNN) in image inpainting [7, 8], we attempt to apply GAN on video error concealment. To the best of our knowledge, this idea has not been previously explored. A novel framework for error concealment is proposed in this paper,which consist of a completion network and two critic network. Our approach is based on the network architecture proposed by [7, 8] which are dedicated to deal with single image completion. For multiple successive distorted frames recover, these two image inpainting algorithms did not consider the temporal information from previous frames. Therefore, we make a few minor but important changes. A pair of neighbor frames and mask are fed into the completion network which aims to recover lost area in the current frame by not only the undistorted pixels in current image but also referring the previous frame information. In summary, we make the following contributions:

- We propose the *Mask Attention Convolutional Layer* to focus on the pixels around the lost area in current distorted frame and the co-located region in the previous frame for error concealment.
- We use SSIM loss to enforce structural similarity and attach the WGAN-GP [9] loss to both global and local critics to enforce global and local consistency.
- We construct a large mask dataset for training and generating video error concealment model. These masks are extracted from the decoded low resolution videos caused by real transmission loss.

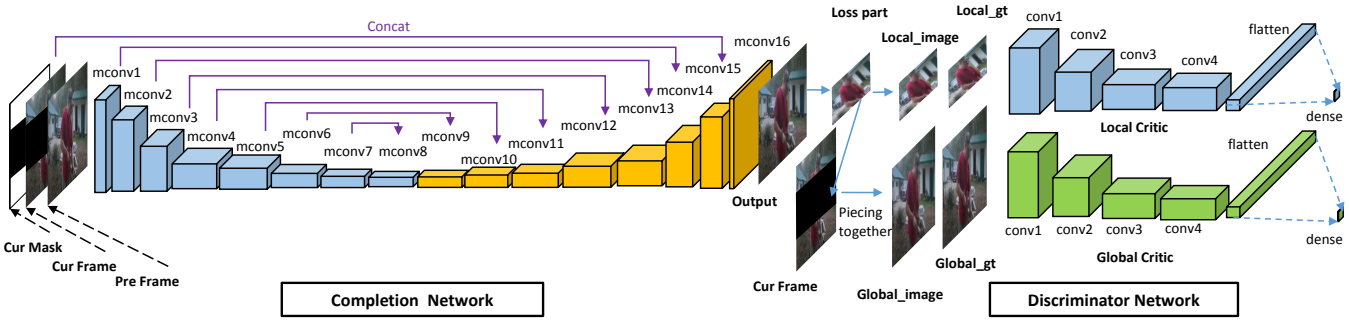

Fig. 1. Overview of our network architecture

The remainder of the paper is organized as follows: in Section 2, the proposed network architecture is described. Experimental results and conclusion are given in Section 3, and Section 4, respectively.

2. APPROACH

In this section, we focus on our model architecture and the mask attention convolutional layer. Our approach is based on deep convolutional neural networks. A single completion network is used for the error frame concealment, a discriminator network similar to the one used in [7], which consists of local critic and global critic. An overview of this approach can be observed in Fig. 1.

2.1. Completion Network

The pixels in the previous frame which co-located to lost pixels in the current frame are copied for initialization. The general architecture follows an encoder-decoder structure, which uses stacked mask attention convolution operations and mask updating steps to perform image completion. We refer to our mask attention partial convolution operation and mask update function jointly as the *Mask Attention Convolutional Layer*. The *Mask Attention Convolutional Layer* is designed based on the partial convolution [8], which focus on valid pixels around the lost area. The partial convolution at every location is expressed as Eq. (1) in [8].

$$x' = \begin{cases} \mathbf{W}^T(\mathbf{X} \odot \mathbf{M})\frac{1}{\mathrm{sum}(\mathbf{M})} + b, & \mathrm{sum}(\mathbf{M}) > 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where W is the convolution filter weights for the convolution filter and b is the corresponding bias. X are the feature values (pixels values) for the current convolution (sliding) window and M is the corresponding binary mask. $\odot$ denotes element-wise multiplication. As can be seen, output values depend only on the unmasked inputs. The scaling factor $1/\mathrm{sum}(\mathbf{M})$ applies appropriate scaling to adjust for the varying amount of valid (unmasked) inputs. After each partial convolution operation, the mask m' is updated. If the convolution was able to condition its output on at least one valid input value, then the mask is removed for that location. This is expressed as Eq. (2)

$$m' = \begin{cases} 1, & \mathrm{sum}(\mathbf{M}) > 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

For video error concealment, the completion network not only needs the valid pixels around the lost area but also refer the temporal information (especially the co-located region) in the previous frame. Therefore, we overwrite the lost area by co-located pixels in the previous frame. We modify the Eq. (1) to Eq. (3) to perform mask attention convolution for referring the temporal information. The mask updating rule is the same as Eq. (2).

$$x' = \begin{cases} \mathbf{W}^T(\mathbf{X} \odot \mathbf{M})\frac{1}{\mathrm{sum}(\overline{\mathbf{M}})}) + b, & \mathrm{sum}(\mathbf{M}) = 0 \\ \mathbf{W}^T(\mathbf{X} \odot \mathbf{M})\frac{1}{\mathrm{sum}(\overline{\mathbf{M}})} + b, & \mathrm{sum}(\overline{\mathbf{M}}) = 0 \\ \mathbf{W}^T(\mathbf{X} \odot (\frac{\mathbf{M}}{\mathrm{sum}(\mathbf{M})} + \frac{\overline{\mathbf{M}}}{\mathrm{sum}(\overline{\mathbf{M}})})) + b, & \text{otherwise} \end{cases} \quad (3)$$

Our UNet-like completion network is similar to the one used in [10], replacing all convolutional layers with mask attention convolution layers. The inputs are: the previous RGB frame, the current distorted RGB frame and the corresponding binary channel mask (0 for a pixel to be completed). Then they are all resized to 512×512. ReLU and LeakyReLU with alpha = 0.2 is used between all encoding, decoding layers respectively. Batch normalization layer is used before ReLU/LeakyReLU except for the first mask attention convolutional layer. The encoder part comprises eight mask attention convolutional layers with stride=2 and kernel sizes of first three layers are 7, 5, 5 and the others are 3. The output sizes of the first three are 64, 128, 256 and the others are 512. The skip links feed the decoder stage concatenate both the feature maps and masks before being input to the next mask convolution layer. The decoder includes eight upsampling layers with a factor of 2 followed by a mask convolutional layer. The output channels for mask attention convolutional layers in the decoder are 512, 512, 512, 512, 256, 128, 64 and 3. The output of last mask attention convolution layer is clipped to -1 and 1.

Table 1. Architectures of the local critic

Local critic			
Type	Kernel	Stride	Outputs
conv1.	5×5	2×2	64
conv2.	5×5	2×2	128
conv3.	5×5	2×2	256
conv4.	5×5	2×2	512
FC	-	-	1

Table 2. Architectures of the global critic

Global critic			
Type	Kernel	Stride	Outputs
conv1.	5×5	2×2	64
conv2.	5×5	2×2	128
conv3.	5×5	2×2	256
conv4.	5×5	2×2	256
FC	-	-	1

2.2. Discriminator Network

A global critic and a local critic are designed to match potentially correct images and train the generator with adversarial gradients. The inputs to the network are image concealed by the completion network and the ground truth of current frame. The two critics are based on four convolutional layers and a single fully-connected layer that compress the images into small feature vectors and outputs two continuous value, one is for local WGAN-GP loss and the other is for global WGAN-GP loss. All the convolutional layers employ a stride of 2×2 pixels to decrease the image resolution. In contrast with the completion network, the kernel sizes of all convolutions are 5. The details of the local and global critics are shown in Table 1 and Table 2.

3. EXPERIMENTS

3.1. Distorted Video Mask Dataset

100 videos with different scenarios are selected from UCF101 [11] and converted to YUV(420) format firstly. Then, each YUV sequences are compressed by HEVC reference software 16.7 (HM16.7), the intra period is every 8 frames with only P frames in between, SliceMode is set 1 and quantization parameter is set 0 (avoid compression distortion). Next, we simulate the real lossy transmission environment and add only one loss pattern, which is random dropping of slices in all the P frames according to the slice loss rate (SLR). SLR values are 1%, 3%, and 5% respectively. Finally, the lossy bitstream is decoded and the distorted videos are obtained. We repeat above steps and obtain 1000 distorted videos. The 2148 binary mask are extracted from these videos.

3.2. Training

We randomly select 200 videos (different to videos for mask generating in 3.1) from UCF101 for comparisons in 3.3. The all other videos for training. Let $C(x, m)$ denotes the completion network in a functional form, $D(x, m)$ denotes the combined context critics in a functional form. x is the input image, m is the region mask. Here we use the SSIM loss rather than MSE. We also use two Wasserstein GAN-GP losses rather than the sigmoid cross-entropy loss used in the original GAN, one critic looks at the global image while the other looks at the local patch of the missing region. WGAN uses the *Earth-Mover* distance (a.k.a. *Wasserstein-1* distance) $W(\mathbb{P}_r, \mathbb{P}_g)$ for comparing the generated and real data distributions. For our completion and context discriminator networks, our objective function becomes:

$$\min_C \max_D \ \mathbb{E}[D(x, m)] - \mathbb{E}[D(C(z, m), m)] \tag{4}$$

our version of gradient penalty can be implemented as follows:

$$\lambda \mathbb{E}[(\|\nabla_{\hat{x}} D(\hat{x}, m)\|_2 - 1)^2] \tag{5}$$

$$\hat{x} = x + t[C(x, m) - x] \tag{6}$$

where λ is set 10, $\hat{x}$ is defined as Eq. (6), t is range from 0-1.

In optimization, we use the Adam algorithm [12] which sets a learning rate 0.0001. The exponential decay rate for the first moment estimate is 0.5, for the second moment estimate is 0.9. An overview of the general training procedure can be seen in Algorithm 1.

Algorithm 1: Training procedure of network

```
while iterations t_discriminator < T_train do
    foreach t_completion = 1, ..., T_subtrain do
        Sample a minibatch of images x and mask m
        from training data;
        Construct inputs z ← x ⊙ (1 − m) + x_pre ⊙ m;
        Get predictions
        x̃ ← x ⊙ (1 − m) + C(z, m) ⊙ m;
        Sample t ∼ U[0,1] and x̂ ← (1 − t)x + tx̃;
        Update network C with SSIM loss and two
        adversarial critic losses;
    end
    Get predictions x ← x ⊙ (1 − m) + C(z, m) ⊙ m;
    Update two critics with x, x̃ and x̂;
end
```

3.3. Comparison

Our method is compared with three other algorithms:

1) **copy**: directly copy pixels from the co-located LCUs from the reference frame.

2) **WBMA** [5]: a weighted boundary matching algorithm to sort PUs in a lost LCU.

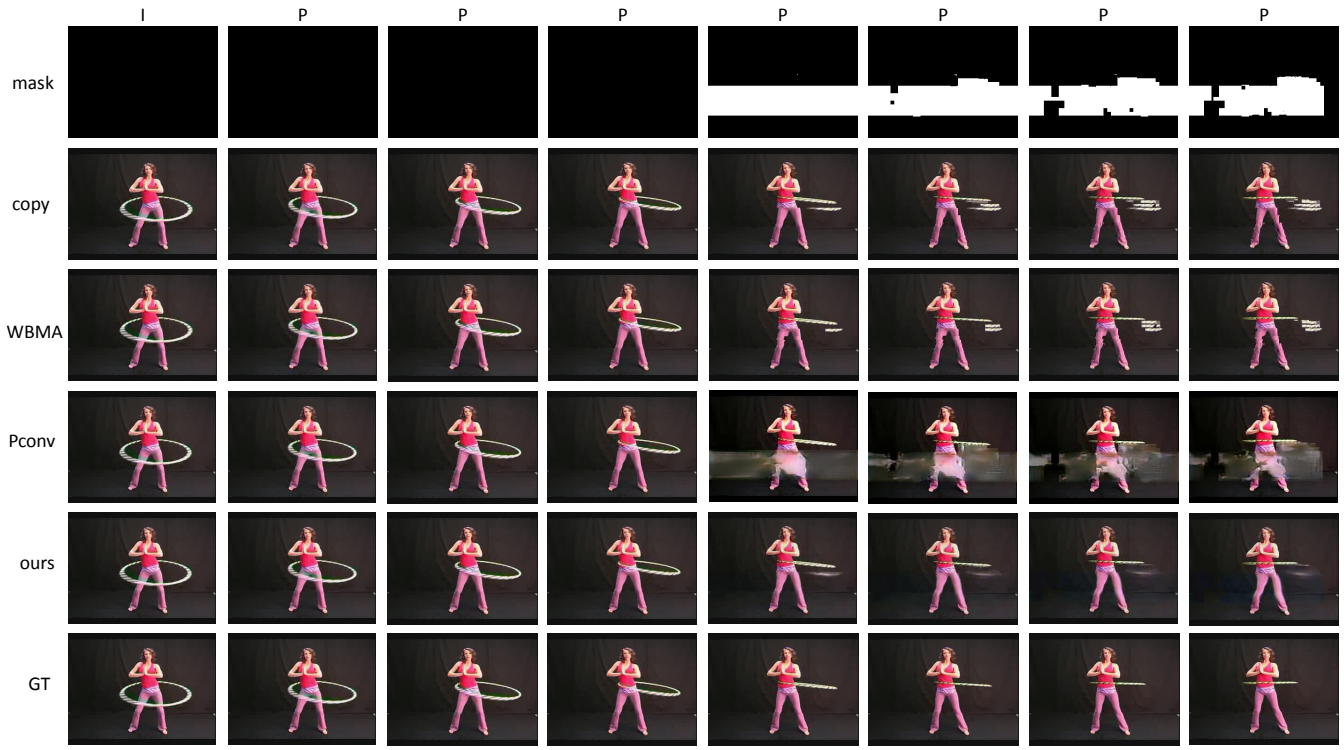

Fig. 2. Reconstructed results of one group frames between two I Frames

3) **Pconv** [8]: partial convolutions for single image inpainting.

Both the objective results and the subjective visual quality of concealment performance are evaluated. The experimental settings are consistent with the setting of generating lossy videos in 3.1. For comparison, we also trained **Pconv** network for concealment, respectively. **copy** and **WBMA** are implemented in HM16.7, with the same loss pattern. Table 3 shows the performance of the different methods, in terms of difference of quality (SSIM). Method of **copy** is used as reference for comparison. The results are the average of the first 40 frames (YUV format) in 200 videos. The result presented in Table 3 shows that when the SLR is 1%, our method achieves quality improvements than other methods. However, with the SLR growing, the performance of the proposed method decreasing. That is because the completed frame different from the ground truth is referred by next lossy P frame and the quality of subsequent completed frames will be influenced by the error accumulation until next I frame.

The visual comparisons of concealment are presented in Fig. 2. Our method successfully preserves the shape of the moving object with smooth edges while the other two methods fail to maintain the structure of the moving object with blockiness, smear and deformed boundary.

Table 3. Comparison of the different EC methods

SLR	Δ SSIM compared to **copy**		
	WBMA	Pconv	ours
1%	0.0007	-0.0431	0.0008
3%	0.0014	-0.0506	0.0005
5%	0.0016	-0.0614	-0.0011

4. CONCLUSION

In this paper, we have proposed a novel deep generative model based approach for video error concealment. In our network, the mask attention convolution layer is designed for focusing on the valid pixels around the lost area and the co-located region in the previous frame. Besides, we also construct a large mask dataset for training and generating video error concealment model. These masks are extracted from the decoded low resolution videos caused by real transmission loss. The test results on UCF101 demonstrate that our method has a better performance in error concealment for HEVC at low SLR rate.

5. ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China (Grant No: 61772436), Foundation for Department of Transportation of Henan Province (2019J-2-2), and the Fundamental Research Funds for the Central Universities.

6. REFERENCES

[1] G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, 2012.

[2] J. Nightingale, Q. Wang, C. Grecos, and S. Goma, "The impact of network impairment on quality of experience (qoe) in h.265/hevc video streaming," *IEEE Trans. Consum. Electron.*, vol. 60, no. 2, pp. 242–250, 2014.

[3] T. Wiegand, G. J. Sullivan, G. Bjntegaard, and A. Luthra, "Overview of the h.264/avc video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, 2003.

[4] C. Liu, R. Ma, and Z. Zhang, "Error concealment for whole frame loss in hevc," in *Advances on Digital Television and Wireless Multimedia Communications*, 2012, pp. 271–277.

[5] Y.-T. Peng and P. C. Cosman, "Weighted boundary matching error concealment for hevc using block partition decisions," in *2014 48th Asilomar Conference on Signals, Systems and Computers*, 2014.

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," vol. 3, 2014, pp. 2672–2680.

[7] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graphics*, vol. 36, no. 4, pp. 1–14, 2017.

[8] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[9] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5767–5777.

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, 2017.

[11] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arxiv*, vol. abs/1212.0402, 2012.

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