MULTI-VIEW FRAME RECONSTRUCTION WITH CONDITIONAL GAN

Tahmida Mahmud, Mohammad Billah, Amit K. Roy-Chowdhury
University of California, Riverside

UCR Bourns College of Engineering
Video Computing Group

IEEE GlobalSIP
2018 6th IEEE Global Conference on Signal and Information Processing

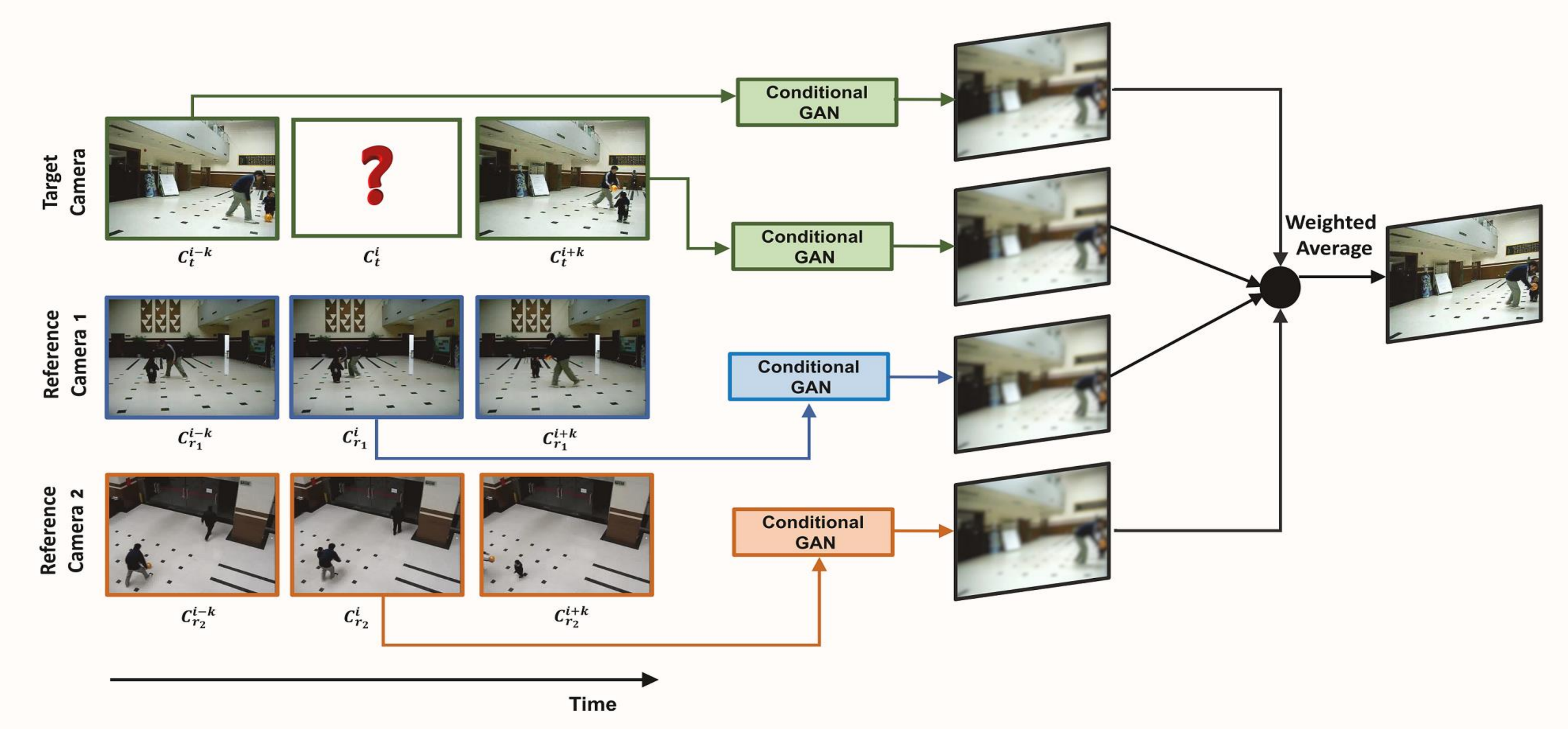
1. Motivation

- Frame reconstruction is critical in applications like retrieving missing frames in surveillance videos, anomaly detection, data compression, video editing, video post-processing, animation, spoofing and so on.
- When multiple frames are missing and adjacent frames within the camera are far apart, realistic coherent frames can still be reconstructed using corresponding frames from other overlapping cameras.

2. Contributions

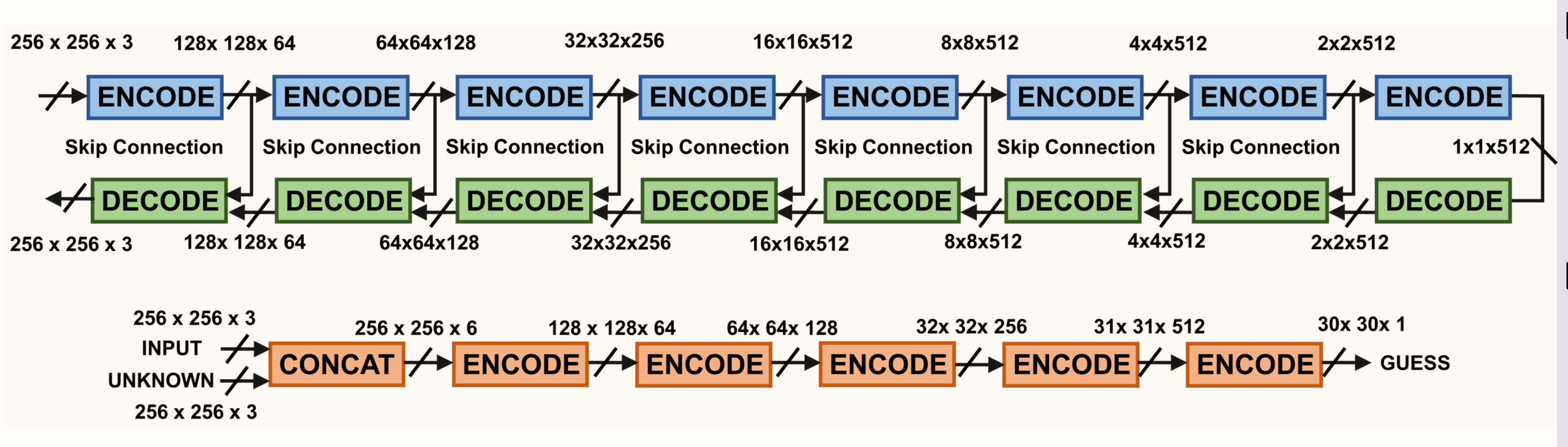
- We tackle a novel problem of frame reconstruction in multi-camera scenario using an adversarial approach.
- We perform extensive experiments on a challenging multi-camera video dataset to show the effectiveness of our method and on a single-camera video dataset to provide quantitative comparison with the state-of-the-art.

3. Solution Overview



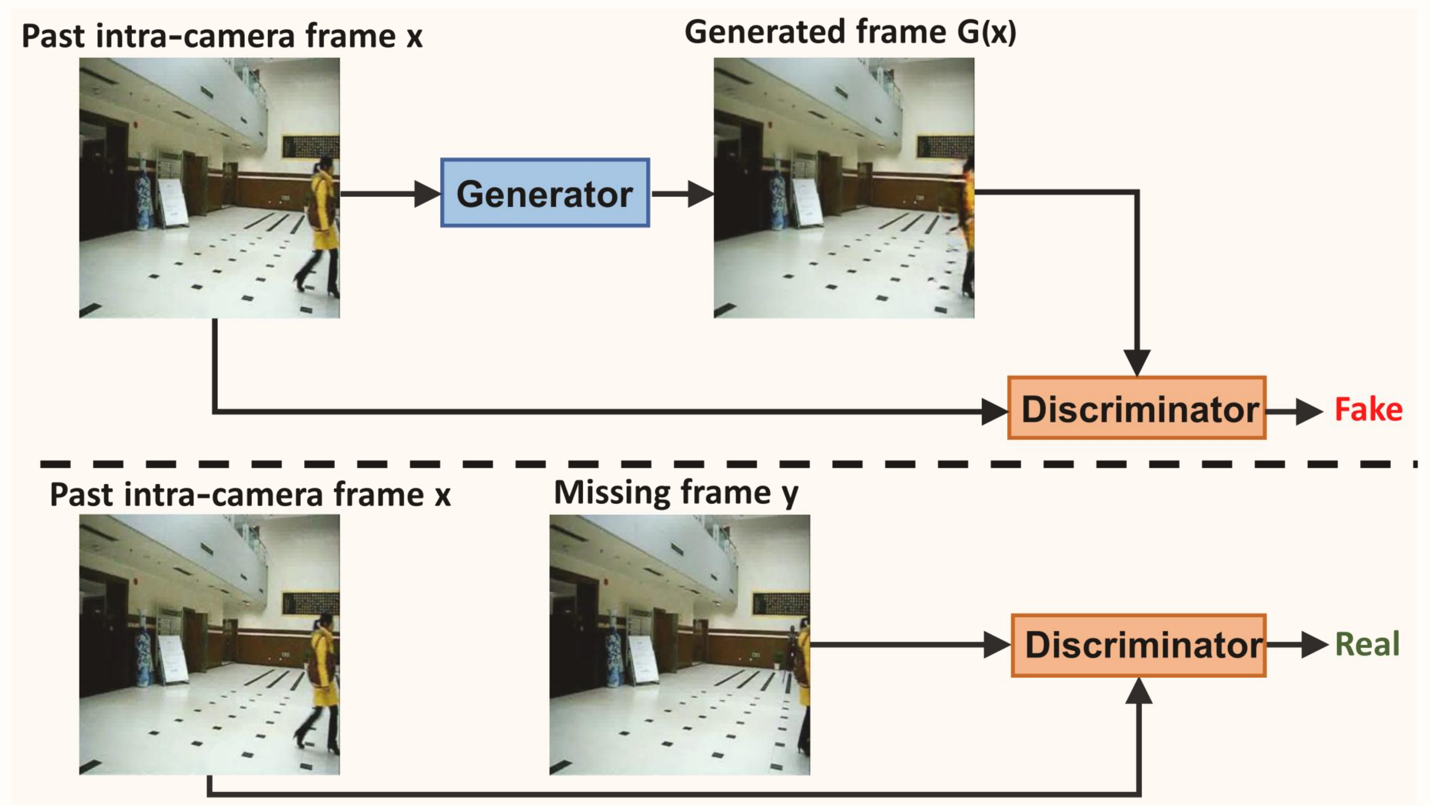
- We learn the representations of the missing frame conditioned on the preceding and following frames within the camera and on the corresponding frames in other overlapping cameras using cGAN.
- These representations are merged together using a weighted average where the weights are chosen by maximizing the average PSNR on a smaller validation set.

4. Network Architecture



- "U-Net"-based architecture of the generator with skip connections which directly connect encoder layers to decoder layers.
- The discriminator tries to differentiate at patch-level and runs convolutationally across the image to generate an averaged output.

5. Model Training Approach



$$G^* = E_{x,y}[\log D(x,y)] + E_{x,z}[\log(1-D(x,G(x,z)))] + \lambda E_{x,y,z}[\|y-G(x,z)\|_1]$$

- We use a combination of $L1$ loss and adversarial loss in the objective function.
- We alternate between a gradient descent step upon D and one upon G and the training maximizes $\log D(x, G(x,z))$.
- To optimize the network, we use a minibatch stochastic gradient descent with an adaptive subgradient method (Adam) and a learning rate of 0.0002.

6. Datasets and Experimental Results

- **KTH Human Action Dataset**: Single-view dataset with 6 types of human activities

Method	PSNR	SSIM
Proposed Method	35.03	0.93
LSTM-Based Method [24]	35.40	0.96

Table 1. Single-view Reconstruction Performance Comparisons for KTH Human Action Dataset.

- **Office Lobby Dataset**: Multi-view dataset with 3 video clips captured by 3 cameras

Gap (frames)	1	3	5	7	15	30
PSNR	32.06	29.28	28.10	27.19	25.56	25.17
SSIM	0.95	0.92	0.91	0.90	0.88	0.87

Table 2. Multi-view Reconstruction Performance for Office Lobby Dataset.

Gap (frames)	1	3	5	7	15	30
Single	32.06	29.24	28.02	27.02	24.17	23.97
Multi	32.06	29.28	28.10	27.19	25.56	25.17

Table 3. Ablation Study for Frame Reconstruction in Office Lobby Dataset considering Single-View vs. Multi-View.

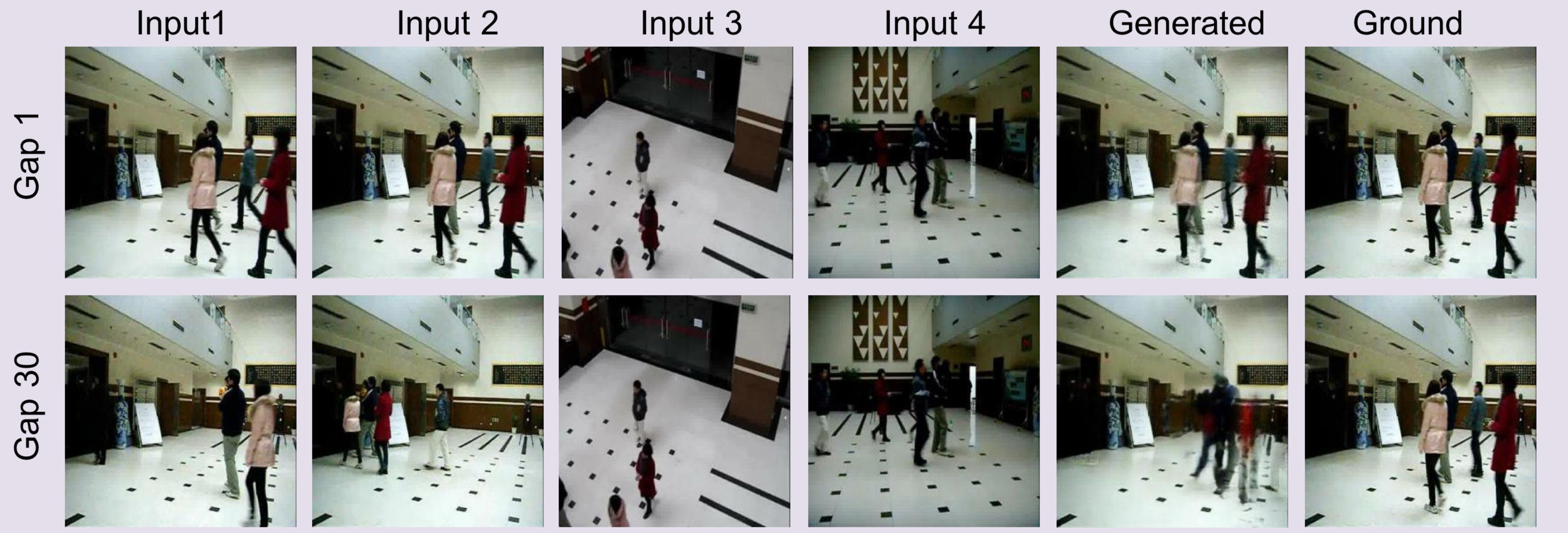


Fig 1. Two examples from Office Lobby Dataset where Input 1, Input 2, Input 3, and Input 4 are the preceding and the following frames of camera 1, and the corresponding frames of camera 2 and 3 respectively. As we increase the gap between the preceding and following frames with the missing frame, frames of camera 2 and camera 3 become more important. For example, due to the large number of missing frames in gap 30, the women in red dress is not visible yet in input 1 and her position is far away in input 2. Still, a person wearing a red dress is visible in the correct position of the generated frame incorporating information from the other two cameras.

7. Acknowledgements

This work was partially supported by NSF grant 1544969 from the Cyber-Physical Systems program.