Video Inpainting by Jointly Learning Temporal Structure and Spatial Details

Chuan Wang[1] Haibin Huang[1] Xiaoguang Han[2] Jue Wang[1]
[1]Megvii (Face++) USA [2]Shenzhen Research Inst. of Big Data, CUHK (Shenzhen)

Introduction

Scan to visit our project page

YouTube Demo

Figure 1: Video inpainting results by our approach. Row 1, 3: input frames with missing regions (shown in gray). Row 2, 4: our results. Note that the filled regions contain rich image details and are temporally coherent.

Motivations:

1. Missing regions may occur in video by object removal, corruption due to storage or file transfer, causing video inpainting has an increasing demand.
2. Directly applying image inpainting solution to video frames will cause flicker artifacts, and there is no clear deep learning solution for this task.

Contributions:

1. The first work to use deep neural network for video inpainting.
2. A novel deep learning architecture:
   - 3D CNN for temporal structure prediction
   - 2D CNN for spatial detail recovering
   - the output temporal structure is fused into the 2D CNN to guide the detail inference.
3. Joint training of the two sub-networks, which further improves the performance of the overall system.

Methodology

Our video inpainting network contains two sub-networks:

- 3DCNN for temporal structure prediction

$$L^{3DCN}(V_{in}^d, M^d, V_c^d) = \frac{\|M^d \odot (G_v(V_{in}^d, M^d) - V_c^d)\|}{\|M^d\|} \quad (1)$$

- 2DCNN for spatial detail recovering, with the output from 3DCNN as guidance

$$L^{CombCN}(V_{in}, M, V_{out}^d, V_c) \quad (2)$$
$$= \sum_{k=1,2,\ldots,F} \frac{\|M^k \odot (G_i(V_{in}^k, M^k, I_{out}^{d,k}) - V_c^k)\|}{F \cdot \|M^k\|}$$

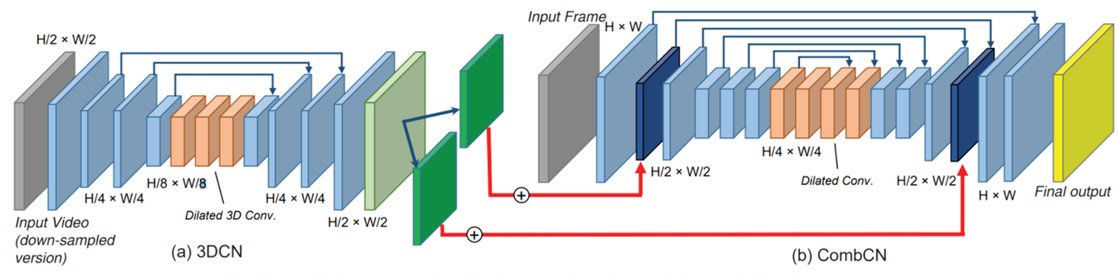
- Jointly train the two networks:

$$L^{total} = L^{3DCN} + \alpha L^{CombCN} \quad (3)$$

Notations:

V_{in} : Input Video M : Mask Video
V_c : Complete Video, i.e. the ground truth
V_{out}^d : Output of 3DCN F : Number of frames
$G_v(\cdot)$: 3D completion network (3DCN) $G_i(\cdot)$: 2D completion network (2DCN)
All notations with superscript d represent downsampled version of the videos

Network Structure:



Note: The gray blocks representing "Input Video" and "Input Frame" contains 4 channels, R/G/B and mask.

Figure 2: Network architecture of our 3D completion network (3DCN) and 3D-2D combined completion network (CombCN). The 3DCN works in low resolution, producing an inpainted video as output. Its individual frames are further convolved and added into the first and last layer of the same size in CombCN. The input video for 3DCN and the input frame for CombCN, shown as gray blocks, are in 4-channel format, containing RGB and the mask indicating the holes to be filled.

Results

Comparisons with existing methods:

	3DCN	2DCN	CombCN (ours)
Reconstructed Frame Details		✓	✓
Smooth Transition (no flicker artifacts)	✓		✓

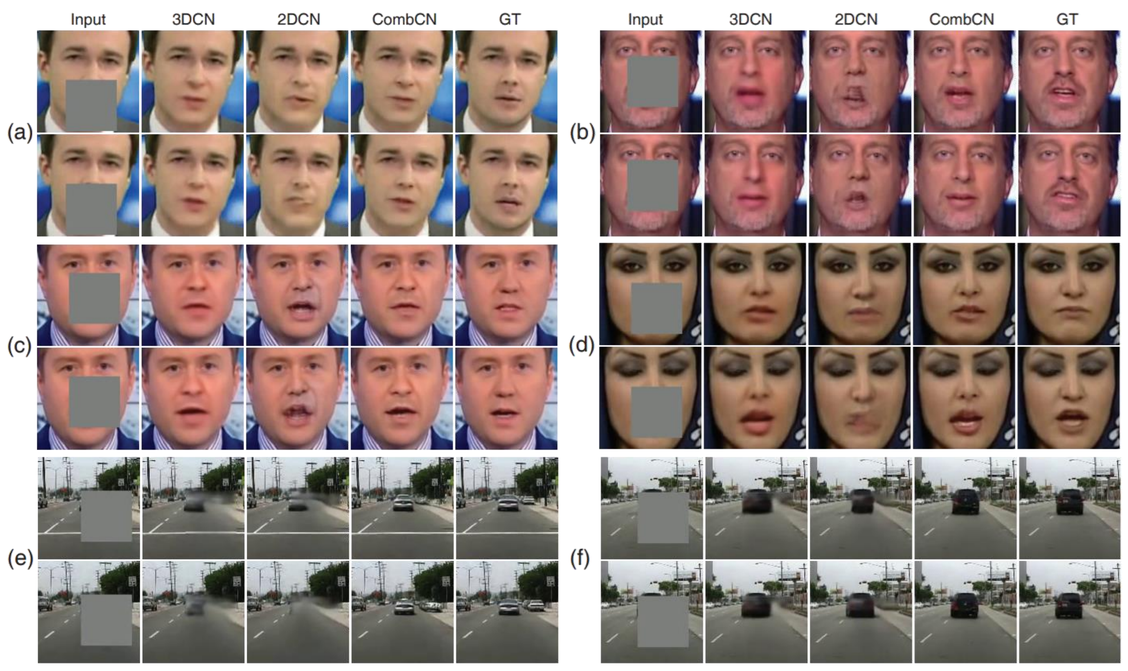


Figure 3: Inpainted frames on datasets FaceForensics (a∼d) and Caltech (e, f). In each panel, the two rows represent two frames of a video, and the five columns from left to right are input, results by 3DCN, 2DCN and CombCN, as well as the target ground truth. Better visual experience can be obtained in our accompanying supplemental materials.

Ablation Studies:

- V-1. Feed 3DCN with videos in lower resolution
- V-2. Involve down-sampling in time axis in 3DCN.
- T-1. Pre-train 3DCN, then train CombCN without finetuning it.
- T-2. Train 3DCN and CombCN jointly from scratch.

	3DCN	2DCN	CombCN (ours)
FaceForensics	7.18	6.77	**6.27**
Caltech	11.91	11.16	**9.56**

	V-1	V-2	T-1	T-2	our method
3DCN	9.51	11.56	6.30	9.28	**4.45**
CombCN	6.39	8.13	5.18	6.31	**4.20**

Table 2: Final l_1 losses. Top: the losses of 3DCN, 2DCN and CombCN of datasets FaceForensics and Caltech. Bottom: the losses of 3DCN and CombCN in 300VW dataset, based on variants of 3DCN (V-1, V-2) and training strategy (T-1, T-2), in comparison with our method.

Conclusion

- An end-to-end framework for video inpainting through a joint 2D-3D CNN which contains a temporal structure inference network and spatial detail recovering network
- These results show that our method significantly improves the performance of existing methods

Acknowledgements

- Shenzhen Fundamental Research Fund under Grant No. KQTD2015033114415450
- "The Pearl River Talent Recruitment Program Innovative and Entrepreneurial Teams in 2017" under grant No. 2017ZT07X152.
- Anonymous Reviewers, and Ms. Chang Li from University of Washington, Mr. Zhangyang Xiong from CUHK (Shenzhen) for their constructive comments and criticism of the manuscript.