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

\[ L^{3DCN}(V^d_{\text{in}}, M^d, V^d_{\text{c}}) = \| M^d \odot (G_v(V^d_{\text{in}}, M^d) - V^d_{\text{c}}) \| + \| M^d \| \]  \( (1) \)

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

\[ L^{CombCN}(V_{\text{in}}, M, V_{\text{out}}, V_{\text{c}}) = \sum_{k=1}^{K} \| M^k \odot (G_v(V^k_{\text{in}}, M^k, V^k_{\text{out}} - V^k_{\text{c}})) \| + \| M^k \| \]  \( (2) \)

- Jointly train the two networks:

\[ L^{total} = L^{3DCN} + \alpha L^{CombCN} \]  \( (3) \)

Notations:
- \( V_{\text{in}} \): Input Video
- \( M \): Mask Video
- \( V_{\text{c}} \): Complete Video, i.e. the ground truth
- \( V_{\text{out}} \): Output of 3DCN
- \( F \): Number of frames
- \( G_v(\cdot) \): 3D completion network (3DCN)
- \( G_v(\cdot) \): 2D completion network (2DCN)

All notations with superscript \( d \) represent downsampled version of the video

Network Structure:

Comparisons with existing methods:

<table>
<thead>
<tr>
<th>Reconstruced Frame Details</th>
<th>3DCN</th>
<th>2DCN</th>
<th>CombCN (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smooth Transition (no flicker artifacts)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

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.

Table 2: Final \( L_1 \) losses. Top: the losses of 3DCN, 2DCN and CombCN of datasets FaceForensics and Calltech. 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. KJTD2015033114415450
- "The Pearl River Talent Recruitment Program Innovative and Entrepreneurial Teams in 2017" under grant No. 20172T07X152
- 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.