2018 ChaLearn Looking at People Challenge
- Track 2. Video Decaptioning

DVDNet

Deep Blind Video Decaptioning with 3D-2D Gated Convolutions

Dahun Kim*, Sanghyun Woo*, Joonyoung Lee, In So Kweon
Our Problem

Remove text overlays in video

Need to consider two important points:
1. Video : Sequence of frames)
2. Blind : No inpainting mask)
Model Overview

Two important points:

- Video: Sequence of frames
- Blind: No inpainting mask

- 3D-2D U-net
- Residual learning
  + Gated convolution
Vanilla 2D U-Net*

Two important points:
• **Video**: Sequence of frames
• **Blind**: No inpainting mask

Frame-by-frame operation
• Spatial context

Input: Multiple frames

Scene dynamics
- Aggregate hints from spatio-temporal neighborhoods

→ Object movements

→ Subtitle changes
Vanilla 3D U-Net*

Multiple frame prediction

- Hard problem
- Heavy
- Not uniform prediction

Focus on a single frame

- Aggregate hints from **lagging and leading frames**.

**Lagging frames**
- we both have a lot in common we do yeah you'll get along just fine
- we both have a lot in common we do yeah you'll get along just fine

**Leading frames**
- we both have a lot in common we do yeah you'll get along just fine
- we both have a lot in common we do yeah you'll get along just fine

**3D-2D U-Net**
- Easy problem
- Light-weight
- Temporal view range

**Center frame**

**Output**
3D-2D U-Net architecture

Focus on a single frame

- 3D convolutions to **flatten** the encoder features **into one frame**.

  → to match the shape and concatenate.
Residual Learning

Two important points:
• Video: Sequence of frames
• Blind: No inpainting mask

• Residual learning
  - Not touching good pixels
  - Focus on the corrupted regions

→ Implicitly knows the inpainting mask
+ Attention

Gated Convolution*

\[ \text{Gate} = \sigma(W_g \otimes \text{Input}) \]
\[ \text{Feature} = \phi(W_f \otimes \text{Input}) \]
\[ \text{Out} = \text{Feature} \odot \text{Gate} \]

- 0-1 value (Gating)
- Attentioning

Loss Function

L1 + gradient L1 + SSIM loss

(a) input  (b) L1  (c) L1 + grad.L1  (d) L1+grad.L1+SSIM  (e) target(GT)
## Quantitative Results

<table>
<thead>
<tr>
<th>User</th>
<th>MSE</th>
<th>PSNR</th>
<th>DSSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>KAIST-RCV</td>
<td>0.0011</td>
<td>33.3527</td>
<td>0.0404</td>
</tr>
<tr>
<td>ucs</td>
<td>0.0011</td>
<td>33.0052</td>
<td>0.041</td>
</tr>
<tr>
<td>hcilab</td>
<td>0.0012</td>
<td>33.0228</td>
<td>0.0424</td>
</tr>
<tr>
<td>anubhap93</td>
<td>0.0012</td>
<td>32.0021</td>
<td>0.0499</td>
</tr>
<tr>
<td>arnavk95</td>
<td>0.0012</td>
<td>32.1713</td>
<td>0.0482</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.0022</td>
<td>30.1856</td>
<td>0.0613</td>
</tr>
</tbody>
</table>
Qualitative Results
2018 ChaLearn Looking at People Challenge
- Track 2. Video Decaptioning

DVDNet
Deep Blind Video Decaptioning with 3D-2D Gated Convolutions

Dahun Kim*, Sanghyun Woo*, Joonyoung Lee, In So Kweon