Deepfake Detection and Segmentation
Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen

Motivation

Deepfake contents could be used to:
• Breaking authentication systems
• Impersonating people, creating fake news or porn videos.
→ Need to detect them and specify the manipulated regions.

Objective

Solving 3 problems simultaneously:
1. Identifying manipulated images/videos (real or fake → classification)
2. Specifying manipulated regions (segmentation)
3. Detecting unseen attacks (transferability / cross-database detection)
→ Heading toward explainable AI

Methodology

Combining classification, segmentation, and image reconstruction in a single network
→ Sharing mutual information between tasks to improving the overall performance.

Results

<table>
<thead>
<tr>
<th>Type of attack</th>
<th>Classification EER (%)</th>
<th>Segmentation Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match condition of seen attack</td>
<td>8.18</td>
<td>90.27</td>
</tr>
<tr>
<td>Mismatch condition of seen attack</td>
<td>8.07</td>
<td>90.20</td>
</tr>
<tr>
<td>Unseen attack 1 (without fine-tuning)</td>
<td>42.24</td>
<td>70.37</td>
</tr>
<tr>
<td>Unseen attack 2 (without fine-tuning)</td>
<td>34.04</td>
<td>84.67</td>
</tr>
<tr>
<td>Unseen attack 2 (fine-tuning on small data)</td>
<td>15.07</td>
<td>93.01</td>
</tr>
</tbody>
</table>