1. Additional Experimental Results

We conduct a set of additional ablation study with MIL-FCN [10] as the baseline.

**Robustness comparison with fully-supervised methods.** We apply JPEG compression and Gaussian blur separately on CASIAv1 [3] to evaluate the robustness of our method. As shown in Fig. 1a and Fig. 1b, our method effectively defends against JPEG compression, especially under the OOD evaluation, where our method significantly outperforms all competing methods. As for the Gaussian blur, our method resists mild Gaussian blur, but is vulnerable to the blur with large kernel sizes, as shown in Fig. 1c and Fig. 1d. While this limitation highlights the need for further research and optimization to improve our method’s resistance to Gaussian blur, it also offers valuable insights into the challenges of robustness in image manipulation detection.

**Ablation study on the early fusion architecture.** In our method, we use a late fusion architecture to fuse multi-source information. We further evaluate our method on the early fusion architecture, where different sources are concatenated in the channel dimension at input. The results are listed in Tab. 1. Note that multi-source consistency learning and ensemble-supervision inter-patch consistency learning do not apply to the early fusion architecture, and thus are excluded. The results show both adaptive pooling and self-supervision inter-patch consistency learning individually improve the performance of the early fusion architecture, and their combination leads to the best performance, demonstrating the effectiveness of our method under the early fusion architecture. Furthermore, under most settings, the performances in early fusion underperform their counterparts in late fusion. Especially, early fusion performances even underperform several single stream performances in late fusion. Such results show early fusion architecture cannot fully utilize each single source under the weakly-supervised setting, and a late fusion design is needed for the weakly-supervised image manipulation detection and localization.

### Table 1. Ablation study on early fusion architecture on IMD2020 [9], where the concatenation of RGB image, Bayar noise map and SRM noise map is fed into a single model. AP is an abbreviation for adaptive pooling.

<table>
<thead>
<tr>
<th></th>
<th>AP</th>
<th>IPC</th>
<th>AUC</th>
<th>Image-Level F1</th>
<th>P-F1</th>
<th>C-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✓</td>
<td>-</td>
<td>0.597</td>
<td>0.700</td>
<td>0.225</td>
<td>0.126</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>✓</td>
<td>0.611</td>
<td>0.750</td>
<td>0.261</td>
<td>0.142</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>0.641</td>
<td>0.733</td>
<td>0.271</td>
<td>0.158</td>
</tr>
<tr>
<td>✓ self</td>
<td>✓</td>
<td>-</td>
<td>0.682</td>
<td>0.771</td>
<td>0.341</td>
<td>0.177</td>
</tr>
</tbody>
</table>

Figure 1. Robustness evaluation against JPEG compression and Gaussian blur. All methods are trained on CASIAv1 [3]. CASIAv2 [3, 4] is used for IND testing, and the average results on Columbia [6], Coverage [11] and IMD2020 [9] are used for OOD testing. Our method is robust against JPEG compression, and mild Gaussian blur.
2. Additional Implementation Details

For unsupervised methods, we use implementations provided by the MKLab\(^1\), and block size of 2 are used for both CFA1 [5] and NOI1 [8] algorithms.

For the results of fully-supervised methods (HP-FCN [7], Mantra-Net [12], CR-CNN [13], and GSR-Net [14]) on NIST16, Columbia, CASIAv1 and Coverage, we use the reproduced results provided by [1, 2]. And the results of the rest of the data are reproduced by us.

References


\(^1\)https://github.com/MKLab-ITI/image-forensics