Detecting Audio-Visual Deepfakes with Fine-Grained Inconsistencies (Supplementary)

Marcella Astrid[1]
marcella.astrid@uni.lu
Enjie Ghorbel[1,2]
enjie.ghorbel@isamm.uma.tn
Djamila Aouada[1]
djamila.aouada@uni.lu

[1] Computer Vision, Imaging & Machine Intelligence Research Group (CVI[2]), Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg

[2] Cristal Laboratory, National School of Computer Sciences, Manouba University, Tunisia

Abstract

This supplementary material accompanies the paper titled "Detecting Audio-Visual Deepfakes with Fine-Grained Inconsistencies" and includes additional illustrations of our method along with further qualitative results.

1 More illustrations on the method

In this section, we provide additional figures to assist readers in understanding the method outlined in the main manuscript. Figure 1 demonstrates the calculation of the distance map and attention map as described in Section 3.1.2 (Spatially-local distance map) and 3.1.3 (Attention module), respectively. Additionally, Figure 2 illustrates the model with residual connection utilized in the ablation study discussed in Section 4.2.1 of the manuscript.

2 More qualitative results

We also present additional results on the DFDC dataset to complement the FakeAVCeleb results presented in the main manuscript. Figures 3, 4, and 5 correspond to Figures 6, 7, and 8 of the main manuscript, respectively. Similar observations to those with the FakeAVCeleb dataset in the main manuscript are noted. However, since the performance difference is not as significant compared to FakeAVCeleb (see Table 1(b), (d), (e) of the manuscript), the distribution difference between settings shown in Figure 4 is not as pronounced as the one shown in Figure 7 of the manuscript.

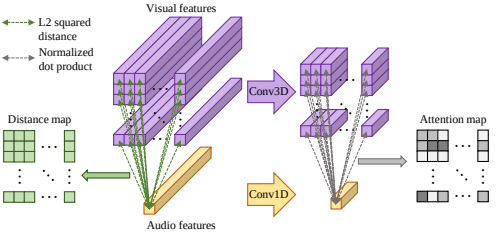

Figure 1: Calculation of the distance map (Section 3.1.2 of the manuscript) and attention map (Section 3.1.3 of the manuscript) based on the visual and audio features.

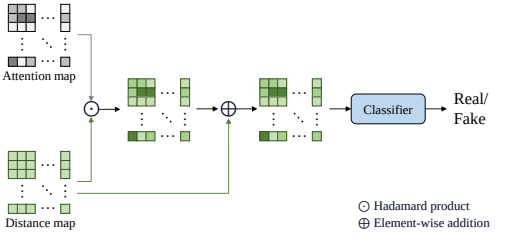

Figure 2: Model with residual connection used in the ablation study.(Section 4.2.1 of the manuscript).

Figure 6 presents visualizations of $\hat{\mathbf{M}}$ with different map sizes, corresponding to the results reported in Figure 5(c) of the main manuscript. Despite the less fine-grained setup, our model is capable of identifying inconsistency-prone regions, resulting in a minimal performance drop (as observed in Figure 5(c) of the manuscript).

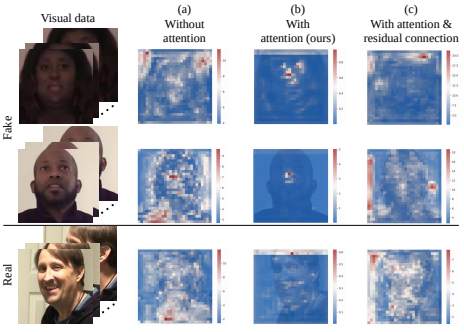

Figure 3: Visualization of $\hat{\mathbf{M}}$ on a few samples from the DFDC test split. This figure corresponds to Figure 6 of the main manuscript.

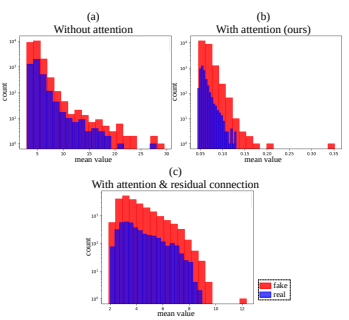

Figure 4: Histograms illustrating the distribution of the mean value of $\hat{\mathbf{M}}$ on the real and fake data of the DFDC test split. This figure corresponds to Figure 7 of the main manuscript.

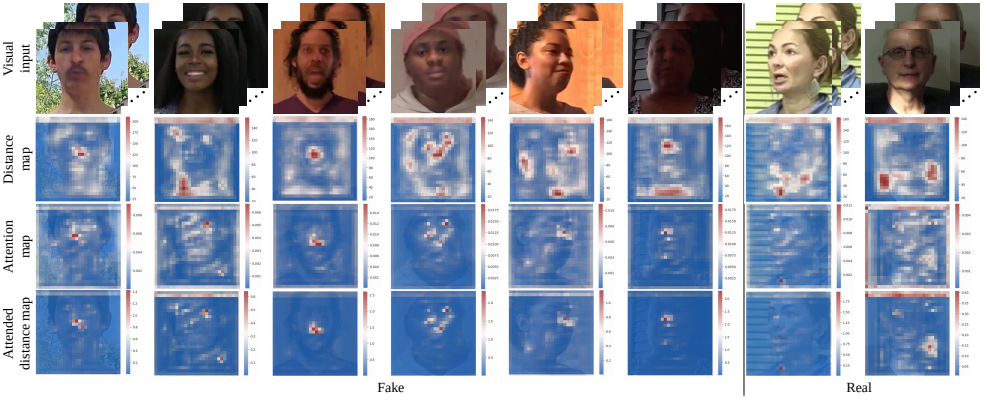

Figure 5: Visualization of the distance map $\mathbf{M}$, attention map $\mathbf{A}$, and the attended distance map $\hat{\mathbf{M}}$ for several examples from the DFDC dataset. This figure corresponds to Figure 8 of the main manuscript.

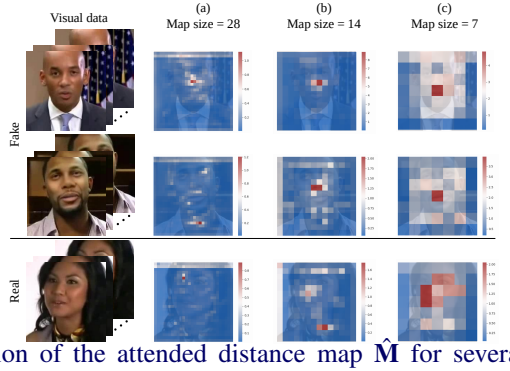

Figure 6: Visualization of the attended distance map $\hat{\mathbf{M}}$ for several examples from the FakeAVCeleb dataset with different map sizes ($H^{v\prime}$ and $W^{v\prime}$).