Appendix of Paper: Uncertainty-guided Learning for Improving Image Manipulation Detection

Kaixiang Ji[1] Feng Chen[1] Xin Guo[1] Yadong Xu[2] Jian Wang[1*] Jingdong Chen[1]
[1] Ant Group [2] Tsinghua University
{kaixiang.jkx, bangzhu.gx, bobblair.wj, jingdongchen.cjd}@antgroup.com
chenfeng1271@gmail.com xuyd17@mails.tsinghua.edu.cn

1. Feature Extractor

It is worth noting that our proposed UEN is applicable to other segmentation-based image manipulation detection methods. Without loss of generality, we present a feature extractor with a general design. Specifically, we adopt HRNetV2 [2] as a basic RGB branch backbone network. The main body of HRNetV2 comprises multiple blocks and is shown in Figure 1. We denote f_{sr} as one block in the s-th stage and r is the resolution index. The RGB branch consists of 4 representations: $\{f_{sr}^{RGB}; s \in \{1,2,3,4\}, r \in \{1,2,3,4\}\}$. The input resolution is $\frac{1}{4}$ of the original image resolution because of a preceding stem ahead of f_{11} which comprises two 3×3 convolutions with stride 2. The resolution of index r is $\frac{1}{2^{r+1}}$ of the original resolution. SRM [3] and resampling feature [1] have shown incredible performance, so we add an SRM branch and a resampling branch parallelly. For the SRM branch, we first pass the input image through an SRM layer, accompanied by two consecutive 3×3 convolution layers with stride 2. Therefore, the SRM branch starts with $\frac{1}{4}$ resolution. The SRM branch consists of 4 representations: $\{f_{sr}^{SRM}; s \in \{1,2,3,4\}, r \in \{1,2,3,4\}\}$. For the resampling branch, following [1], we extract the resampling feature from 32×32 non-overlapping patches. Thus, resampling branch starts with $\frac{1}{32}$ resolution, and consists of 1 representation: $\{f_{sr}^{Resampling}; s \in \{4\}, r \in \{4\}\}$. The output embeddings of each branch in the 4-th stage are all concatenated at the corresponding resolution, and the representation head remains consistent with HRNetV2 except for the number of channels.

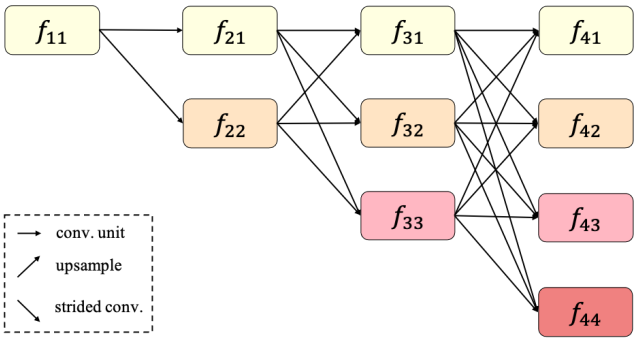

Figure 1: Structure of the RGB stream backbone network.

References

[1] Jawadul H. Bappy, Cody Simons, Lakshmanan Nataraj, B. S. Manjunath, and Amit K. Roy-Chowdhury. Hybrid lstm and encoder-decoder architecture for detection of image forgeries. *IEEE TIP*, 28(7):3286–3300, 2019.

[2] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, 43(10):3349–3364, 2021.

[3] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Learning rich features for image manipulation detection. In *CVPR*, pages 1053–1061, 2018.

*Corresponding author: bobblair.wj@antgroup.com