

Supplementary Materials: V²A-Mark: Versatile Deep Visual-Audio Watermarking for Manipulation Localization and Copyright Protection

Anonymous Authors

In the supplementary materials, we demonstrate implementation details, time and computational cost, more results of robustness, and additional visualization results. We also construct an HTML file to present more visualization results.

1 DETAILS OF BHM AND BRM

The **bit hiding and bit recovery modules** in our V²A-Mark utilize U-shaped networks, as described in [5, 6]. Specifically, for encoding the copyright watermark w_{cop} into the video frame $I_{med}^{(k)}$, we first transform $w_{cop} \in \{0, 1\}^L$ into $\{-0.5, 0.5\}^L$ and input it into stacked Multilayer Perceptrons (MLPs) to generate message features. Meanwhile, $I_{med}^{(k)}$ is passed to a U-shaped feature enhancement network with N layers. During downsampling, a convolution with stride=2 followed by a "Conv-ReLU" layer is applied to reduce spatial resolution by half and double feature channels. During upsampling, we use nearest neighbor interpolation combined with "Conv-ReLU" layers to increase spatial resolution and reduce feature channels. Subsequently, message features from MLPs are up-scaled via nearest interpolation, and integrated with the downsampled and upsampled image features, facilitating modulation of bit-image information. The fusion operation involves residual blocks with a channel attention module [2] and concatenation operation. Finally, the bit hiding module combines $I_{med}^{(k)}$ and w_{cop} at multiple levels to produce a dual-encoded container video frame $I_{con}^{(k)}$. In the decoding process, the received video frame $I_{rec}^{(k)}$ undergoes processing through a U-shaped sub-network, followed by downsampling to size $L \times L$. The recovered copyright watermark w_{cop}^v is then extracted via an MLP, with a threshold of 0 applied to transform it into $\{0, 1\}^L$.

2 TIME AND COMPUTATIONAL COST

To evaluate the computational efficiency and complexity of our method, we input a $3 \times 256 \times 448$ tensor on an NVIDIA 3090Ti to test inference time and the number of parameters. The comparison tamper localization methods are OSN [4], PSCC-Net [3], HiFi-Net [1] and EditGuard [6]. As reported in Tab. 1, our method can decode one frame in 0.129s, which proves the efficiency of our method in processing long videos. We also observe that our method can achieve the best localization performance (IoU) with comparable or close inference time and the number of parameters with other methods.

3 MORE RESULTS OF ROBUSTNESS

To measure the capabilities of copyright protection of our approach, we evaluate the bit reconstruction performance of our approach on five degradations including Gaussian Noise ($\sigma=10$), H.264 coding compression (QP=10), Poisson noise, Gaussian blur, and Median blur. Tab. 2 reports the results of the V²A-Mark. Note that we test the bit accuracy on the DAVIS dataset without tampering. It can

Table 1: Inference time and the number of parameters comparison of our V²A-Mark and other methods.

Methods	IoU	Inference Time (s)	# Params.(M)
OSN [4]	0.125	0.164	128.82
PSCC-Net [3]	0.186	0.091	3.67
HiFi-Net [1]	0.123	1.512	10.13
EditGuard	0.866	0.069	5.45
V ² A-Mark	0.897	0.129	10.28

be seen that the recovered bit accuracy is generally above 99.5%, which is comparable to other video watermarking methods.

To further evaluate the excellent robustness of our V²A-Mark, we present the visualized localization results of our method under different degradations, including Gaussian noise ($\sigma=5$ and $\sigma=10$) and H.264 coding compression (QP=5 and QP=10). As depicted in Figure 1, our V²A-Mark consistently exhibits accurate localization outcomes across various degradations. Despite escalating levels of degradation, there is merely a marginal reduction in localization accuracy, highlighting the practicality and resilience of our approach.

Table 2: Bit Accuracy of the proposed V²A-Mark

Methods	Gaussian Noise	H.264	Poisson	Gaussian Blur	Median Blur
V ² A-Mark	99.64%	99.51%	99.86%	99.72%	99.83%

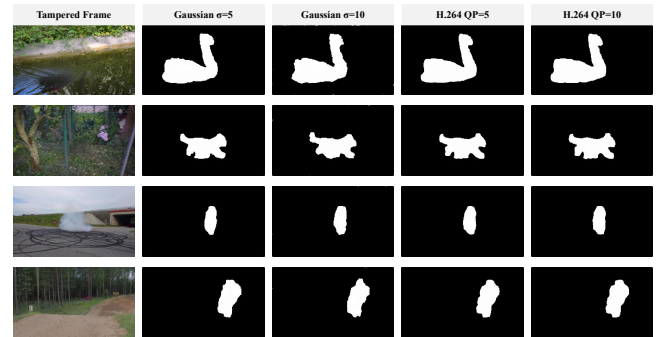


Figure 1: Visualization results of the proposed V²A-Mark under different levels of degradations including Gaussian Noise and H.264 compression.

4 ADDITIONAL VISUALIZATION RESULTS

Relevant results are showcased in the attached HTML file. Please click on the ["./v2a-mark/index.html"](http://v2a-mark/index.html) to check our visual and audio localization performance. Obviously, our V²A-Mark can accurately predict the tampered visual masks and tampered audio period.

REFERENCES

[1] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. 2023. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[2] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. 2022. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 11 (2022), 7505–7517.

[4] Haiwei Wu, Jiantao Zhou, Jinyu Tian, and Jun Liu. 2022. Robust image forgery detection over online social network shared images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Xiaoshuai Wu, Xin Liao, and Bo Ou. 2023. SepMark: Deep Separable Watermarking for Unified Source Tracing and Deepfake Detection. In *Proceedings of the ACM international conference on Multimedia (MM)*.

[6] Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. 2024. EditGuard: Versatile Image Watermarking for Tamper Localization and Copyright Protection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.