Local Conditional Watermarking for Geometrically Robust Deepfake Proactive Forensics

Xiaoshuai Wu and Xin Liao, *Senior Member, IEEE*

***Abstract*—This letter proposes LocMark, a local feature modulation-based conditional watermarking method, to address the limitation that existing watermarking methods for deepfake proactive forensics are vulnerable to geometric distortions. Specifically, LocMark establishes the spatial synchronization capability of the conditioned encoder and decoder via local feature modulation. It first utilizes the local mask as the input condition of the encoder to guide a passive localization network, and uses the localization result as the synchronization condition for the decoder, thereby achieving watermark synchronization and blind extraction. Furthermore, considering that deepfake manipulations primarily occur in the face region, the background mask is utilized as the prior condition of the encoder, thus enhancing the watermarking robustness against deepfake distortions. Experimental results demonstrate that, when subjected to individual or composite pixel-value-based and geometric distortions, LocMark can still utilize the synchronized condition mask to effectively extract the watermark message, significantly enhancing the geometric robustness of proactive forensics. Quantitatively, under individual geometric distortions, LocMark achieves an average bit error rate below 1%, whereas the baseline methods approach 50%; meanwhile, under composite deepfake and geometric distortions, LocMark outperforms the two SOTA local watermarking methods by approximately 30%.**

***Index Terms*—Data Hiding, Image Watermarking, Deepfake Proactive Forensics, Geometric Robustness**

I. Introduction

ROBUST watermarking has emerged as a crucial tool for proactive forensics, facilitating content provenance and regulation [1]. In this domain, encoder-decoder-based deep watermarking has advanced significantly, owing to the effectiveness of end-to-end training with a flexible noise layer in achieving desirable robustness [2]. Recently, the research focus has progressively shifted from resisting common digital distortions [3] to treating complex deepfake manipulations as a distinct form of distortion [4], [5], [6], [7], [8], [9], [10]. However, both common and deepfake distortions are pixel-value-based in nature, mainly altering pixel values while keeping their global coordinates unchanged. Once the pixel positions of the watermarked image are shifted due to geometric distortions, the spatial synchronization between the watermark encoder and decoder is severely disrupted.

To mitigate the desynchronization induced by geometric distortions, existing solutions predominantly fall into two paradigms. The first attempts to explicitly estimate the geometric deformation parameters, compute the inverse transform, and finally extract the watermark from the corrected image. The second establishes an inherent synchronization mechanism between watermark embedding and extraction, completely bypassing the need for any manual or automated correction of the deformed image prior to extraction. For example, Ma *et al.* [11] leverage the global receptive field of a Transformer-based network to enforce watermark synchronization across global coordinates. Most recently, WAM [12] and MaskWM [13] integrate an auxiliary localization network to synchronize the local watermark, successfully breaking the reliance on global coordinates and converting global watermark extraction into local watermark extraction based on relative positions. Despite such advances, these local watermarking methods, as well as traditional methods that resist individual pixel-value-based or geometric distortions, struggle to cope with composite distortions, a severe scenario where both pixel values and positions are altered.

In this letter, we propose LocMark, a local condition-based watermarking method designed to enhance robustness against composite attacks, specifically sequential deepfake and geometric distortions. To resist desynchronizing geometric distortions, LocMark leverages local feature modulation to establish the spatial synchronization capability of the conditioned encoder and decoder. Unlike previous conditional watermarking [14], which uses a 1D one-hot condition vector to modulate the global features of the encoder-decoder, LocMark utilizes a 2D mask representing local regions as the input condition to modulate the local features. Conceptually, by transforming global watermark extraction into conditional watermark extraction based on local relative positions, LocMark overcomes the potential for shortcut learning that relies heavily on the global absolute spatial coordinates of the image. The main contributions of this letter are summarized as follows:

1) We propose a conditional watermarking method, LocMark, which utilizes the local mask as the input condition of the conditioned encoder and decoder, and establishes their spatial synchronization capability via local feature modulation.

2) We develop a conditional synchronization mechanism for blind extraction, utilizing the background mask as a prior condition for the encoder to guide the passive localization network, thereby providing the decoder with a precisely synchronized condition mask even with deepfake distortions.

3) Experimental results under individual and composite distortions confirm that LocMark yields a lower bit error rate (BER), validating the geometric robustness of conditional watermark extraction based on local relative positions.

This work is supported by National Major Project for New Generation Artificial Intelligence (Grant No. 2025ZD0123601), National Natural Science Foundation of China (Grant No. U22A2030), Hunan Provincial Funds for Distinguished Young Scholars (Grant No. 2024JJ2025), and Hunan Provincial Key R&D Program (Grant Nos. 2024AQ2027, 2025AQ2022).

Xiaoshuai Wu and Xin Liao are with College of Cyber Science and Technology, Hunan University, Changsha, China. (e-mail: shinewu@hnu.edu.cn; xinliao@hnu.edu.cn).

Corresponding author: Xin Liao

This article has been accepted for publication in IEEE Signal Processing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LSP.2026.3731912

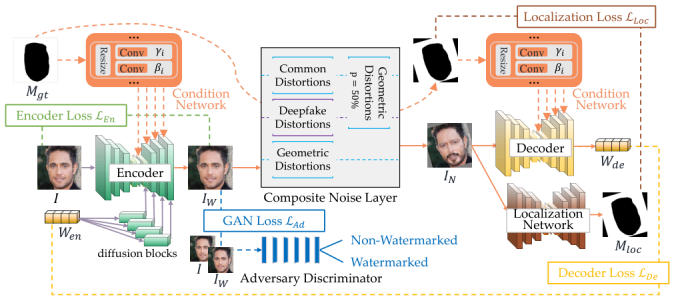

Fig. 1. Overview of the proposed LocMark.

II. Proposed LocMark

As illustrated in Fig. 1, our proposed LocMark consists of five main components: a conditioned encoder En for watermark embedding, a composite noise layer $\mathcal{N}$ for data augmentation, a passive localization network Loc for predicting the synchronized condition mask, a conditioned decoder De for watermark extraction, and an adversary discriminator Ad for adversarial training.

During training, 1) the encoder En receives the original image $I \in \mathbb{R}^{B\times 3\times H\times W}$, the watermark message $W_{en} \in \{-\alpha, \alpha\}^{B\times L}$, and the condition mask $M_{gt} \in \{0,1\}^{B\times 1\times H\times W}$ as inputs to generate the watermarked image I_W; 2) the noise layer $\mathcal{N}$ randomly samples one or more distortions from common distortions δ_{co}, deepfake distortions δ_{df}, and geometric distortions δ_{ge} as pseudo-noise and adds them to the watermarked image I_W, resulting in the single- or composite-distorted image I_N, while simultaneously applying the identical geometric distortions δ_{ge} to the condition mask M_{gt}; 3) the localization network Loc takes the distorted image I_N as input and outputs the localization result M_{loc}, which is jointly affected by the mask M_{gt} and the noise layer $\mathcal{N}$; 4) the decoder De receives the distorted image I_N and the condition mask $\mathcal{N}(\delta_{ge}; M_{gt})$ as inputs to extract the watermark message W_{de}; and 5) finally, the adversary discriminator Ad is trained alternately with the encoder-decoder to improve the visual quality of the watermarked image.

During inference, the decoder De is required to perform fully blind extraction. In this scenario, where both the original image I and the ground-truth synchronized mask $\mathcal{N}(\delta_{ge}; M_{gt})$ are unavailable, the decoder utilizes the mask M_{loc} predicted by the upstream localization network Loc as the synchronization condition, thereby independently achieving conditional watermark extraction based on local relative positions.

A. Conditioned Watermark Encoder

The encoder En adopts a conditional architecture, which is composed of two parts: a base watermark encoder and a condition network. The base watermark encoder employs a U-Net-like structure with several diffusion blocks to increase watermark redundancy; see [14] for details. However, it is worth emphasizing that LocMark differs significantly from prior art [14] in terms of the condition network (e.g., local feature modulation, 2D condition mask, and convolutional structure). The condition network consists of several independent sets of convolutional layers. Each set corresponds to a specific residual block within the upsampling path of the U-Net. Specifically, the i-th set takes the condition mask M_{gt}, downsampled to the corresponding spatial resolution, as input, and generates two spatially varying modulation parameters $\gamma_i, \beta_i \in \mathbb{R}^{B\times C_i\times H_i\times W_i}$ via two parallel convolutional layers. Herein, C_i, H_i, and W_i represent the number of channels, the height, and the width of the feature map F_i output by the i-th residual block, respectively. Subsequently, the Spatial Feature Transform (SFT) [15] is applied to F_i using parameters γ_i and β_i to achieve spatially-aware, element-wise local feature modulation. The new feature map $\widetilde{F}_i$ is calculated as follows:

$$\widetilde{F}_i = F_i \odot \gamma_i + \beta_i \tag{1}$$

where $\odot$ denotes the Hadamard product.

To ensure that the watermarked image I_W is visually similar to the original image I, the loss function is defined as follows:

$$\mathcal{L}_{En} = L_2(I, I_W) = L_2(I, En(I, W_{en}, M_{gt})) \tag{2}$$

where $L_2(\cdot)$ denotes the mean squared error loss.

B. Composite Noise Layer

The noise layer $\mathcal{N}$ includes common distortions δ_{co}, deepfake distortions δ_{df}, and geometric distortions δ_{ge}, and can be divided into two sequential layers based on the order in which distortions are applied. The first layer consists of various individual distortions and is utilized to generate the single-distorted image $I_N = \mathcal{N}(\delta_{co}, \delta_{df}, \delta_{ge}; I_W)$. In the second layer, if the distorted image I_N has undergone common or deepfake distortions, the geometric distortions δ_{ge} are further applied with a 50% probability, thereby yielding the composite-distorted image $I_N = \mathcal{N}(\delta_{ge}; I_N)$. This noise layer yields not only composite-distorted images but also single-distorted images, which significantly enhances the diversity of the distorted images. Furthermore, while outputting the distorted image I_N, the noise layer applies the identical geometric distortions δ_{ge} to the mask M_{gt}, generating the synchronized mask $\mathcal{N}(\delta_{ge}; M_{gt})$.

C. Passive Localization Network

The localization network Loc similarly adopts a U-Net structure to ensure the output mask has the same spatial resolution as the input image. Its localization result M_{loc} is supervised by the synchronized mask $\mathcal{N}(\delta_{ge}; M_{gt})$ output by the noise layer. The loss function can be formulated as:

$$\mathcal{L}_{Loc} = L_2(\mathcal{N}(\delta_{ge}; M_{gt}), M_{loc}) = L_2(\mathcal{N}(\delta_{ge}; M_{gt}), Loc(I_N)) \tag{3}$$

D. Conditioned Watermark Decoder

To synchronize the watermark embedding and extraction, during the training stage, the decoder De utilizes the exact synchronized mask $\mathcal{N}(\delta_{ge}; M_{gt})$ as the input condition rather than the prediction result M_{loc}. Notably, the feature modulation and structural design of this condition network are completely identical to those of the conditioned encoder.

To ensure that the extracted watermark message W_{de} is consistent with the original embedded message W_{en}, the loss function is formulated as follows:

$$\mathcal{L}_{De} = L_2(W_{en}, W_{de}) = L_2(W_{en}, De(I_N, \mathcal{N}(\delta_{ge}; M_{gt}))) \tag{4}$$

E. Adversary Discriminator

The adversarial discriminator Ad is trained alternately with the encoder-decoder, and its loss function is defined as follows:

$$\begin{aligned}\mathcal{L}_{Adv} =& L_2(Ad(I) - \overline{Ad}(I_W), 1) + L_2(Ad(I_W) - \overline{Ad}(I), -1) \\ &+ 10 \times \mathcal{L}_{GP1} + 10 \times \mathcal{L}_{GP2}\end{aligned} \tag{5}$$

where $\overline{Ad}(\cdot)$ denotes the average output of the discriminator over the current data batch. In the above equation, the first two terms represent the relativistic average discriminator loss [16], while the latter two terms denote $R_1 + R_2$ regularization. Both theoretical and empirical studies demonstrate that this combination effectively guarantees the local convergence and stability of adversarial training [17].

When training the encoder, the parameters of the discriminator are kept frozen:

$$\mathcal{L}_{Ad} = L_2(Ad(I) - \overline{Ad}(I_W), -1) + L_2(Ad(I_W) - \overline{Ad}(I), 1) \tag{6}$$

and the backpropagated gradients are utilized to update the parameters of the encoder.

In summary, the total loss function can be expressed as:

$$\mathcal{L}_{LocMark} = \lambda_1 \mathcal{L}_{Ad} + \lambda_2 \mathcal{L}_{En} + \lambda_3 \mathcal{L}_{De} + \lambda_4 \mathcal{L}_{Loc} \tag{7}$$

III. Experiments

A. Experimental Setups

The experiments are conducted on CelebA-HQ [18], referencing the official dataset splits. The face masks are obtained from CelebAMask-HQ [19], and an inversion operation is applied to yield the background masks. Regarding pixel-value-based distortions, we adopt the typical settings consistent with prior art [4], [14], which involve various common distortions and three deepfake distortions: SimSwap for face swapping [20], GANimation for facial reenactment [21], and StarGAN for attribute editing [22]. Regarding geometric distortions, the settings are as follows:

- Rotation: The angle parameter is set to $[-30, 30]$.
- Perspective: The distortion scale is set to $[0.1, 0.3]$.
- HorizontalFlip: The image is mirrored horizontally along the vertical axis.

In our experiments, the composite distortion is constructed by sequentially applying a randomly selected geometric distortion on top of a common- or deepfake-distorted image.

1) Implementation Details: The proposed LocMark is implemented using PyTorch and executed on a single NVIDIA RTX 4090 GPU. Training at the 128×128 resolution lasts for 100 epochs, whereas training at the 256×256 resolution is extended to 150 epochs. The training batch size is set to 16. The weights λ_1, λ_2, λ_3, and λ_4 in Eq. (7) are set to 0.1, 1, 10, and 1, respectively. Additionally, the scaling factor α for the watermark message is set to 0.11.

2) Baselines: At the 128×128 resolution, LocMark is compared with MBRS [3] and SepMark [4]. Meanwhile, the model trained at the 256×256 resolution is compared with AdvMark [5] and ConMark [14], as well as the local watermarking methods WAM [12] and MaskWM (the decoder-only masked version) [13], which support global watermark embedding and local watermark extraction. All the baselines are implemented using their official models and pre-trained weights for reproducibility.

Fig. 2. Visualization examples of watermarked images and normalized residuals. Left: 128×128 image. Right: 256×256 image.

TABLE I
QUANTITATIVE EVALUATION OF VISUAL QUALITY.

Method	Image Size	Capacity	PSNR ↑	SSIM ↑	LPIPS ↓
MBRS [3]	128×128	30-bit	33.0456	0.8106	0.0141
SepMark [4]	128×128	30-bit	38.5112	0.9588	0.0028
LocMark (Ours 128)	128×128	30-bit	37.9078	0.9510	0.0048
AdvMark [5]	256×256	128-bit	39.0292	0.9321	0.0115
ConMark [14]	256×256	128-bit	39.7650	0.9454	0.0059
LocMark (Ours 256)	256×256	128-bit	35.3444	0.9083	0.0172

3) Evaluation Metrics: To evaluate visual quality, we use PSNR, SSIM, and LPIPS. For the robustness comparison, BER is utilized as the default metric. Moreover, the prediction accuracy of the localization network is measured using IoU.

B. Visual Quality

As shown in Table I, at the 128×128 resolution, the PSNR, SSIM, and LPIPS values of LocMark are comparable to those of SepMark and superior to those of MBRS, achieving a PSNR of approximately 38 dB. In contrast, at the higher 256×256 resolution, where the desynchronization caused by geometric distortions becomes more severe, the specifically designed LocMark yields lower quantitative values than the baseline methods. Nevertheless, it is worth emphasizing that, benefiting from the advanced adversarial training, LocMark maintains highly competitive subjective visual quality. As demonstrated by the visualization examples in Fig. 2, the watermarked images generated by LocMark are free from obvious visual artifacts. As observed, the watermark residuals are preferentially distributed in highly textured areas, while they are kept low in smooth areas of the image.

C. Robustness Comparison

The results of the individual distortions are presented in Table II. Under common distortions, the average BERs of all methods are below 1% at both resolutions. Under deepfake distortions, LocMark exhibits a lower average BER compared to the baseline methods, achieving 1.3158% and 1.0097% at the two resolutions, respectively. Moreover, under geometric distortions, the average BERs of the baseline methods all surge to approximately 50%. In contrast, LocMark transforms global watermark extraction into conditional watermark extraction based on local relative positions, thereby demonstrating excellent geometric robustness. Specifically, its average BERs are a mere 0.0299% and 0.4066%, respectively. More importantly, due to the lack of basic geometric robustness, these baselines all fail in global watermark extraction and degrade to the level of random guessing under composite attacks.

Hence, Table III presents the comparison with the local watermarking methods WAM and MaskWM under composite distortions at the 256×256 resolution. In the experiments,

TABLE II
ROBUSTNESS COMPARISON UNDER INDIVIDUAL DISTORTIONS (BER: %).

Distortion	MBRS [3]	SepMark [4]	LocMark (Ours 128)	AdvMark [5]	ConMark [14]	LocMark (Ours 256)
Identity	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
JPEG	0.2597	0.2136	1.0966	0.0122	0.0307	0.0426
Resize	0.0000	0.0059	0.0000	0.0000	0.0006	0.0008
GaussianBlur	0.0000	0.0024	0.0012	0.0000	0.0000	0.0000
MedianBlur	0.0000	0.0012	0.0000	0.0000	0.0000	0.0000
Brightness	0.0000	0.0059	0.0177	0.0025	0.0028	0.0017
Contrast	0.0000	0.0012	0.0059	0.0000	0.0000	0.0000
Saturation	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Hue	0.0000	0.0000	0.0012	0.0000	0.0000	0.0006
Dropout	0.0000	0.0000	0.0000	0.0000	0.0000	0.0008
SaltPepper	0.0000	0.0413	0.0106	0.0000	0.0011	0.0022
GaussianNoise	0.0000	0.7460	0.5099	0.0575	0.2257	0.0824
Average	0.0216	0.0848	0.1369	**0.0060**	0.0217	0.0109
SimSwap	19.3744	13.8255	6.7186	24.9394	10.3214	7.0622
GANimation	0.0000	0.0000	0.0012	0.0000	0.0003	0.0008
StarGAN(Male)	18.3133	0.1157	0.4237	0.0075	0.0006	0.0014
StarGAN(Young)	17.0562	0.1074	0.4698	0.0072	0.0008	0.0008
StarGAN(BlackHair)	19.2233	0.1416	0.5288	0.0116	0.0008	0.0017
StarGAN(BlondHair)	18.4478	0.1712	0.6409	0.0083	0.0014	0.0006
StarGAN(BrownHair)	17.6381	0.0980	0.4273	0.0086	0.0003	0.0006
Average	15.7219	2.0656	1.3158	3.5689	1.4751	**1.0097**
Rotation	44.5774	49.8277	0.0708	48.0779	47.3652	0.0523
Perspective	48.2413	46.3999	0.0189	49.7704	49.7743	1.1674
HorizontalFlip	53.4018	53.3239	0.0000	51.7672	51.5382	0.0000
Average	48.7402	49.8505	**0.0299**	49.8718	49.5592	0.4066

TABLE III
ROBUSTNESS COMPARISON UNDER COMPOSITE DISTORTIONS (BER: %).

Distortion	WAM [12]	MaskWM [13]			LocMark	
	(32-bit)	(32-bit)	(64-bit)	(128-bit)	(Ours 256)	(GT mask)
Identity	8.1910	0.0055	0.3679	3.7486	0.4219	0.4188
JPEG	43.2509	42.6877	48.6085	41.3930	6.4816	6.4586
Resize	8.8660	0.0077	0.5638	16.7883	1.8818	1.8782
GaussianBlur	9.4801	0.0044	0.6064	17.7112	0.4866	0.4844
MedianBlur	9.0806	0.0066	0.5455	17.8495	0.5270	0.5242
Brightness	7.6421	0.0598	0.5710	4.3945	0.9143	0.9118
Contrast	7.6067	0.0376	0.4891	4.2427	0.8156	0.8103
Saturation	7.5425	0.0066	0.3370	3.7015	0.4642	0.4601
Hue	7.6786	0.0310	0.4194	3.8360	0.4960	0.4946
Dropout	8.1268	0.0077	0.4963	6.0494	1.4549	1.4535
SaltPepper	7.8523	0.0255	0.6568	5.7426	3.6274	3.6172
GaussianNoise	8.0526	0.3895	4.6449	21.8902	6.6960	6.6921
Average	11.1142	3.6085	4.8589	10.8030	**2.0223**	2.0170
SimSwap	41.7006	49.6448	49.7615	46.3156	22.7484	22.7614
GANimation	8.4421	0.0166	0.4980	4.2515	5.0101	5.0079
StarGAN(Male)	41.4372	47.7127	49.3919	44.6718	4.2034	4.2067
StarGAN(Young)	41.3310	47.4095	49.4943	43.8872	4.1931	4.1940
StarGAN(BlackHair)	41.2148	47.5567	49.4201	45.6680	3.9840	3.9663
StarGAN(BlondHair)	41.6818	47.7802	49.5059	42.4459	4.1035	4.0935
StarGAN(BrownHair)	41.8135	47.3995	49.5341	45.7020	3.9865	3.9851
Average	36.8030	41.0743	42.5151	38.9917	**6.8899**	6.8878

the PSNR of watermarked images is adjusted to around 35 dB to maintain the same level of visual quality. As the experimental results indicate, although WAM and MaskWM possess the capability to resist individual geometric distortions, when they are subjected to composite distortions, particularly JPEG compression or severe deepfake distortions followed by geometric distortions, their BERs increase sharply. Our analysis reveals that in these cases, their localization networks cannot accurately distinguish between the watermarked and non-watermarked regions, leading to the failure of watermark synchronization and subsequent local extraction. In contrast, leveraging the robust conditional synchronization mechanism, the proposed LocMark achieves high localization accuracy for the predicted mask with an average IoU score above 0.97, and thus exhibits a relatively low BER. However, because of the combined effects of pixel-value-based and geometric distortions, the BER decreases only marginally further, even if the ground-truth synchronized mask is available. Nonetheless, we empirically validated the trade-off between message capacity and robustness: at reduced capacities of 64 and 32 bits, the average BER under deepfake composite distortions decreases from 6.8899% to 2.7096% and 0.8366%, respectively.

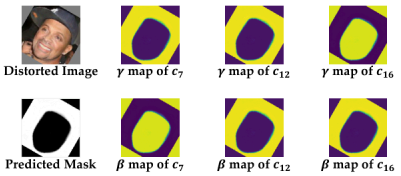

Fig. 3. Visualization of the SFT modulation parameters γ and β. c_i denotes the i-th channel of the modulation parameters.

TABLE IV
COMPARISON UNDER COMPOSITE DISTORTIONS ACROSS DIFFERENT GEOMETRIC INTENSITIES AND MODULATION LOCATIONS (BER: %, IOU).

Distortion	Config 1	Config 2	Config 3	Config 4
SimSwap	11.4884/0.9666	18.1114/0.9594	23.3983/0.9590	20.4237/0.9145
GANimation	2.2816/0.9847	3.2897/0.9816	4.4700/0.9804	5.2089/0.9570
StarGAN(Male)	3.6426/0.9671	9.8761/0.9631	13.3440/0.9642	14.2410/0.9207
StarGAN(Young)	3.6898/0.9673	10.3128/0.9636	13.4714/0.9643	14.4063/0.9222
StarGAN(BlackHair)	3.6520/0.9625	10.2136/0.9608	13.2743/0.9639	14.4370/0.9184
StarGAN(BlondHair)	4.9256/0.9653	11.5628/0.9606	14.4924/0.9606	15.2809/0.9140
StarGAN(BrownHair)	3.6249/0.9658	9.5904/0.9611	12.9143/0.9627	13.5068/0.9155
Average	**4.7578/0.9816**	10.4224/0.9643	13.6235/0.9650	13.9292/0.9232

D. Additional Results and Analysis

This set of experiments uses 128×128 resolution images. In Table IV, we report the results under deepfake composite distortions across different configurations, maintaining an image quality adjustment of around 37 dB.

1) Sensitivity Analysis: We expanded the ranges of Rotation and Perspective from $[-30, 30]$ and $[0.1, 0.3]$ (Config 1) to $[-60, 60]$ and $[0.1, 0.4]$ (Config 2), and further to more severe ranges of $[-90, 90]$ and $[0.1, 0.5]$ (Config 3). The results confirm the expected degradation trend while demonstrating that our method retains the effectiveness of conditional watermark extraction even with extreme geometric distortions.

2) Ablation Study: We compared modulating the background region versus the face region (Config 3 vs. Config 4), and found that the background yields a slightly lower average BER, while achieving better localization accuracy (IoU: 0.9650 vs. 0.9232). The results confirm that the background prior provides a more relevant condition for the watermark encoder and decoder against deepfake manipulations.

3) Visualization Analysis: We visualized the modulation parameters using heatmaps, as shown in Fig. 3. It can be observed that the modulation parameters contain highly distinct spatial information, clearly distinguishing different parts of the image, especially the background and face regions.

IV. CONCLUSION

This letter proposes LocMark, a local feature modulation-based conditional watermarking method, to establish a robust mechanism for watermark synchronization and blind extraction, thereby overcoming the potential for shortcut learning that relies heavily on the global absolute spatial coordinates of the image. Experimental results demonstrate that LocMark exhibits a lower bit error rate in conditional watermark extraction based on local relative positions, significantly enhancing the geometric robustness of proactive forensics.

REFERENCES

[1] V. Asnani, X. Yin, and X. Liu, "Proactive schemes: A survey of adversarial attacks for social good," *International Journal of Computer Vision*, vol. 134, no. 4, p. 186, 2026.

[2] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "Hidden: Hiding data with deep networks," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 657–672.

[3] Z. Jia, H. Fang, and W. Zhang, "Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression," in *Proceedings of the ACM International Conference on Multimedia*, 2021, pp. 41–49.

[4] X. Wu, X. Liao, and B. Ou, "Sepmark: Deep separable watermarking for unified source tracing and deepfake detection," in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 1190–1201.

[5] X. Wu, X. Liao, B. Ou, Y. Liu, and Z. Qin, "Are watermarks bugs for deepfake detectors? rethinking proactive forensics," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2024, pp. 6089–6097.

[6] T. Wang, M. Huang, H. Cheng, X. Zhang, and Z. Shen, "Lampmark: Proactive deepfake detection via training-free landmark perceptual watermarks," in *Proceedings of the ACM International Conference on Multimedia*, 2024, pp. 10 515–10 524.

[7] C. Wang, C. Shi, S. Wang, Z. Xia, and B. Ma, "Dual-task mutual learning with qphfm watermarking for deepfake detection," *IEEE Signal Processing Letters*, vol. 31, pp. 2740–2744, 2024.

[8] M. Yang, B. Qi, R. Ma, Y. Xian, and B. Ma, "Hashshield: a robust deepfake forensic framework with separable perceptual hashing," *IEEE Signal Processing Letters*, vol. 32, pp. 1186–1190, 2025.

[9] Z. He, Z. Guo, L. Wang, G. Yang, Y. Diao, and D. Ma, "Waveguard: Robust deepfake detection and source tracing via dual-tree complex wavelet and graph neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 36, no. 4, pp. 4757–4770, 2026.

[10] C. Sun, H. Sun, Z. Guo, Y. Diao, L. Wang, D. Ma, G. Yang, and K. Li, "Diffmark: Diffusion-based robust watermark against deepfakes," *Information Fusion*, vol. 127, p. 103801, 2026.

[11] L. Ma, H. Fang, T. Wei, Z. Yang, Z. Ma, W. Zhang, and N. Yu, "A geometric distortion immunized deep watermarking framework with robustness generalizability," in *Proceedings of the European Conference on Computer Vision*, 2025, pp. 268–285.

[12] T. Sander, P. Fernandez, A. Durmus, T. Furon, and M. Douze, "Watermark anything with localized messages," in *Proceedings of the International Conference on Learning Representations*, 2025, pp. 1–17.

[13] R. Hu, J. Zhang, S. Zhao, N. Lukas, J. Li, Q. Guo, H. Qiu, and T. Zhang, "Mask image watermarking," in *Proceedings of the Advances in Neural Information Processing Systems*, 2025, pp. 1–12. [Online]. Available: https://openreview.net/forum?id=da1u6Cy7Mq

[14] X. Wu, X. Liao, J. Zhang, M. Chen, Y. Wu, and J. Guo, "Versatile and harmless deepfake proactive forensics via conditional watermarking," *Information Sciences*, vol. 742, p. 123030, 2026.

[15] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 606–615.

[16] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard gan," in *Proceedings of the International Conference on Learning Representations*, 2019, pp. 1–10.

[17] N. Huang, A. Gokaslan, V. Kuleshov, and J. Tompkin, "The gan is dead; long live the gan! a modern gan baseline," in *Proceedings of the Advances in Neural Information Processing Systems*, 2024, pp. 44 177–44 215.

[18] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *Proceedings of the International Conference on Learning Representations*, 2018, pp. 1–12.

[19] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5549–5558.

[20] R. Chen, X. Chen, B. Ni, and Y. Ge, "Simswap: An efficient framework for high fidelity face swapping," in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 2003–2011.

[21] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 818–833.

[22] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797.