Rethinking Image Watermarking for Better Embedding

Sigui Chen[a], Jie Zhang[b], Jinyang Huang[*a], Xin Liu[a], Xiang Zhang[a], Jixuan He[a], Zhi Liu[c], Dan Guo[a], and Meng Li[a]

[a]Hefei University of Technology, 485 Danxia Ave., Hefei, China
[b]A* STAR, 2 Fusionopolis Way, Singapore, Singapore
[c]University of Electro-Communications, 1-5-1 Chofugaoka, Tokyo, Japan,
*hjy@hfut.edu.cn

ABSTRACT

Image watermarking plays a critical role in protecting intellectual property. While most state-of-the-art (SOTA) methods based on deep learning enhance watermark performance, they often fail to consider the variability of embedding differences among different cover images. This oversight neglects prior image knowledge, which limits embedding performance. To address this issue, by associating the significant spatial redundancy in natural images that can well depicts the prior image knowledge, we propose a novel local watermarking framework based on *local consistency*, called **LocMark**. Unlike traditional digital watermarking methods that target texture-rich regions, **LocMark** adaptively embeds watermarks into highly consistent smooth regions. Specifically, to identify these consistent regions, we design a simple yet effective region localization mechanism by correlating local consistency with pixel distribution. By normalizing the total gradient to the scale of one region, a variable-length training strategy based on gradient accumulation is further proposed to handle varying regions. Compared to the baseline methods, **LocMark** achieves SOTA performance with an average PSNR increase of 2.96 dB and a significant reduction in detection rate by steganalysis of 32.85%.

Keywords: Image watermarking, Local consistency, Variable-length training strategy

1. INTRODUCTION

The rapid growth of the Internet has brought significant convenience to the public, yet it has also created opportunities for criminal activities. For example, many malicious individuals may illegally exploit images from the Internet for commercial gain, inevitably infringing on the rights of copyright holders, making it extremely necessary to defend these legitimate rights. Digital watermarking technology, known for its advantages in imperceptibility, robustness, and security, is widely used for information tracking and copyright protection.

Traditional digital watermarking suggests that texture-rich regions—areas with complex changes in grayscale-tend to exhibit better visual quality in image watermarking. Therefore, to achieve a balance between imperceptibility and robustness, these methods tend to embed stronger watermarks in texture-rich regions.[1,2] Although traditional watermarking may not always match the performance of deep learning-based methods, the underlying principle of leveraging the inherent characteristics of images to enhance the adaptability for different images remains noteworthy and continues to contribute to the field. In recent years, some deep learning-based watermarking frameworks have emerged.[3,4] These frameworks mainly concentrate on improving the robustness against common distortions.[4,5] Though these efforts have yielded promising results, they often neglect the differences in watermark embedding, leading to a lack of flexibility in some images.

By thoroughly considering these embedding differences, it has been observed that significant spatial redundancy exists in natural images,[6] which can be referred to as *local consistency*. This concept contributes intuitively to the generation of watermarked images. From the perspective of the image itself, objects within an image share similar representations in terms of attributes such as color, shape, and brightness. Once the pixel values within a region are known, it becomes relatively straightforward to infer the pixel values in neighboring regions. From an architectural perspective, given that the central component of deep learning-based methods is the convolutional kernel, the value of each pixel is influenced by the collective impact of its neighboring pixels. Local consistency, therefore, suffers less interference from inconsistent distributions, which enhances the reconstruction of watermarked images.

Based on the above analysis, we propose a novel local image watermarking framework **LocMark** by employing local consistency. **LocMark** adaptively embeds watermarks into regions with high local consistency, which can acquire capacity gains while maintaining better imperceptibility and robustness. As shown in Fig. 1, the overall framework **LocMark** basically consists of five main components: a region locator to identify areas with highly consistent distribution, an encoder to imperceptibly embed watermarks into the identified regions, a noise layer to enhance the robustness of the watermarks, a decoder to reconstruct watermarks from the noisy regions, and a discriminator to enhance the visual quality of watermarked images. Since the number of regions satisfying high local consistency is variable for each image, *i.e.*, the cover region to be watermarked of each batch is unfixed during training. To ensure that the model learns adequately, a variable-length training strategy is employed.

In a nutshell, the main contributions of this paper can be summarized as follows:

- To the best of our knowledge, this paper is the first to leverage images' inherent properties, referred to as *local consistency*, to enhance the embedding adaptability of SOTA deep learning-based methods.
- We propose **LocMark**, a new framework that adaptively embeds watermarks into highly consistent smooth regions. This framework incorporates a region localization mechanism, addressing the lack of flexibility in deep watermarking across various images.
- To reduce the learning difficulty, we propose a novel variable-length training strategy based on gradient accumulation. This strategy optimizes the average loss magnitude and ensures a uniform optimization direction against noise attacks.
- We conduct extensive experiments on many SOTA deep learning-based watermarking methods to demonstrate the generalization of *local consistency*. **LocMark** achieves excellent performance in terms of imperceptibility, robustness, and secrecy, further achieving a $1.5\times\sim3\times$ capacity gain.

2. RELATED WORKS

2.1 Traditional Digital Watermarking

Traditional digital watermarking is an important technology to protect intellectual property, characterized by imperceptibility and robustness. Watermarks are often embedded in the original spatial domain, or other image transform domains.[7] To strike a balance between imperceptibility and robustness, by adjusting the strength factor according to texture values, several works have been proposed.[2,8] For example,[1] proposed a texture-aware local adaptive image watermarking that associates complex texture with low-frequency components. However, all these traditional methods can only hide several bits of messages, and their performance is weaker in terms of visual quality and robustness when compared to deep learning-based methods.

2.2 Deep Learning-based Image Watermarking

In recent years, with the rise of artificial intelligence, there have been some novel deep learning-based watermarking frameworks proposed,[3,5] which show superior performance in imperceptibility and robustness. Subsequent work mainly focused on enhancing the robustness of image watermarking to resist JPEG Compression.[4,9,10] For example,[4] proposed a mini-batch of real and simulated JPEG compression to enhance the JPEG robustness.[5] combines invertible and non-invertible mechanisms to gain better imperceptibility. Besides, some methods also leverage watermarking to protect copyright of deep models.[11–13] Despite the fact that these deep learning-based methods surpass traditional approaches in numerous aspects, they predominantly neglect the embedding difference in various images, resulting in a lack of flexibility.

2.3 Build A Bridge between Traditional and Deep Watermarking

Overall, though existing SOTA deep learning-based watermarking algorithms perform well, they lack adaptiveness toward various images. A natural thought is to introduce prior knowledge in deep watermarking like traditional digital watermarking. Unfortunately, the theoretical basis of traditional digital works, that the embedding performance of complex texture-rich regions is superior to that of smooth regions is not directly applicable to deep learning-based methods. Therefore, to further enhance embedding performance, a novel deep learning-based method that can utilize images' inherent natures to gain flexibility like traditional watermarking is urgently needed to be proposed while maintaining imperceptibility and robustness.

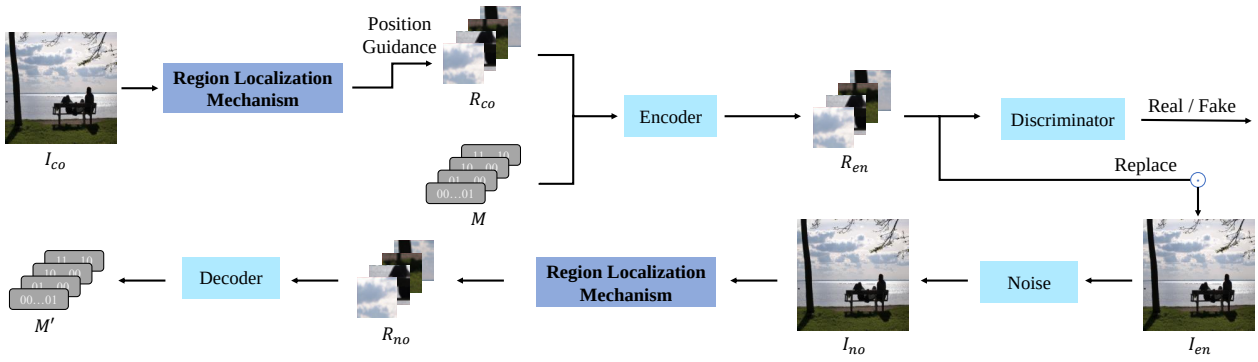

Figure 1. The diagram of the adaptative image watermarking via local consistency, which usually consists of five components: a region locator based on region localization mechanism, an encoder, a decoder, a noise layer, and a discriminator network. Though some other basic watermarking frameworks may differ from ours in implementation, the key idea of them remains the same.

3. METHODS

In this section, we initially present the overview of our proposed local watermarking framework based on local consistency called **LocMark**, followed by a detailed description of the region localization mechanism and the variable-length training strategy. Then, the differences between our work and other similar works are directly demonstrated.

3.1 Framework Overview

Based on *local consistency*, we propose the local image watermarking framework **LocMark**, whose workflow is shown in Fig. 1. **LocMark** aims to fully utilize images' inherent consistency to enhance the model's embedding adaptability by adaptively embedding watermarks into regions satisfying this characteristic. Totally, five main components are included in **LocMark**, *i.e.*, a region locator to position regions sharing highly consistent distribution from cover images and noisy images, an encoder to embed watermarks to cover regions in an unseen way, a noise layer to enhance the robustness of watermarking to resist distortion, a decoder to reconstruct watermarks from noisy regions, and a discriminator to improve the visual quality of watermarked regions.

Given a cover image $\boldsymbol{I}_{co} \in \mathbb{R}^{C\times H\times W}$ where C refers to the numbers of image channels, the region locator identifies several regions of high local consistency $\boldsymbol{R}_{co} \in \mathbb{R}^{N\times C\times h\times w}$, where N means the number of regions. Besides, H and W denote the height and width of each image, while h and w represent the height and width of each region respectively. Then, these regions $\boldsymbol{R}_{co}$ are watermarked by the encoder with corresponding watermarks M, getting watermarked regions $\boldsymbol{R}_{en}$. As the basic unit of information transfer and noise distortion, the watermarked image $\boldsymbol{I}_{en} \in \mathbb{R}^{C\times H\times W}$ is obtained by replacing corresponding regions in $\boldsymbol{I}_{co}$ by $\boldsymbol{R}_{en}$. After the noise attack, we acquire the noisy watermarked regions $\boldsymbol{R}_{no} \in \mathbb{R}^{N\times C\times h\times w}$ from the noisy image $\boldsymbol{I}_{no}$ through the region locator, either. Finally, the decoded watermarks M' are reconstructed from noisy regions $\boldsymbol{R}_{no}$ by the decoder. In addition, $\boldsymbol{I}_{co}$ and $\boldsymbol{I}_{en}$ are put into the discriminator to improve the visual quality of watermarked images.

3.2 Region Localization Mechanism

To locate highly consistent regions, we design a simple yet effective region localization mechanism containing two cascade modules. As depicted in Fig. 2, such a design significantly helps the model identify highly consistent regions, due to the fact that the redundant semantic information is highly correlated with the local consistency.

Distribution Extraction Module. Significant spatial redundancy in natural images makes the value of a pixel be inferred from its neighborhood regions, which can be referred to as local consistency. Similar to image reconstruction, this nature helps to improve the visual quality of deep watermarking. Typically, this redundancy is represented in the same object sharing similar representations in terms of attributes such as color, shape, and

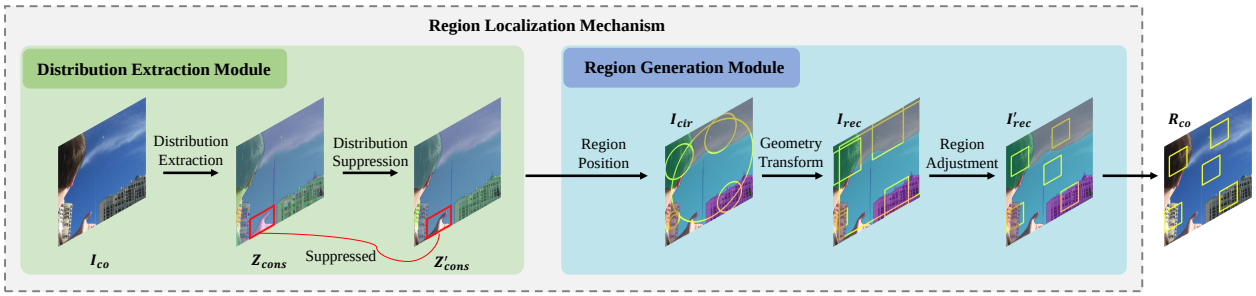

Figure 2. The procedure of region localization mechanism based on local consistency, which consists of two cascade modules: Distribution Extraction Module (DEX) and Region Generation Module (REG). In consistency maps, we assign the identical color for regions sharing consistent representation.

brightness. To effectively identify these regions, we adopt the pretrained Segment Anything Model (SAM)[14] because of its strong semantic segmentation in the DEX module. Through the SAM, the consistency map Z_{cons} is obtained after semantic segmentation, in which each segment object shares a consistent representation.

Since the size of the stego regions is constant, to ensure the local consistency of the located regions, those tiny regions with a pixel area less than a certain amount are regarded as unable to gain enough consistency, which is radically suppressed by setting a fixed threshold. In this way, the suppressed consistency map $Z^{'}_{cons}$ is acquired.

Region Generation Module. The target of the REG module is to identify highly consistent regions based on the suppressed consistency map $Z^{'}_{cons}$ generated in the DEX module. However, the irregular shape of objects leads to difficulty in identifying the regions that completely belong to one object. To mitigate this issue, the REG module first identifies areas that are as highly consistent as possible and then performs region adjustments to accommodate the watermark.

In general, the pixels of an object usually converge in the center region. The outer rectangular region of each irregularly segmented region is first chosen by the REG module as a candidate region to capture the corresponding region with the largest segmentation area. Then, considering that circle regions usually contain more pixels, to filter out the disturbance from other objects, this rectangular region is further used to generate its corresponding maximum inscribed circles I_{cir} with the same center.

Then, to accurately find the initial watermark area, the REG module locates the minimum externally connected rectangle regions I_{rec} of the generated circle regions I_{cir} as the preliminary regions to be watermarked. Regions generated in this way are optimal solutions in the case of objects with irregular shapes since it follows the objective law of pixel center clustering.

Finally, to get the ultimate regions $I^{'}_{rec}$ as the basic unit of watermarking, it is necessary to adjust regions for their possible overlap and different sizes. To maintain a consistent distribution, we propose to adopt the principle of central offset. Those regions whose size is less than the predefined standard region size are expanded towards the center of the images, while the others are shrunk towards their own centers. Besides, to avoid identifying overlapping regions, we remove subsequent regions with overlapping parts in order of traversal.

3.3 Variable-Length Training Strategy

Due to our proposed region localization mechanism, the number of regions identified in different images is not fixed. Consequently, the number of watermarked regions in each batch varies, which can pose challenges for model training, such as "cuda out of memory".

To meet such demands, we propose a variable-length training strategy based on gradient accumulation, which enables the model to fully learn effective representatives from varying regions. Regardless of the number of regions to be watermarked in a batch, the total loss is normalized to the scale of one region. Additionally, to avoid optimization confusion, the optimization direction towards noise attack in a batch should be the same. We further study the effectiveness of the proposed training strategy in the ablation study.

Table 1. The average visual quality of current SOTA watermarking methods integrated with **LocMark**

Model	PSNR	SSIM	Capacity Gains
HiDDeN	37.22	0.9465	3.013 ×
+ **LocMark**	**42.01(+4.79)**	**0.9711(+0.0246)**	
MBRS	48.54	0.9842	1.702 ×
+ **LocMark**	**50.85(+2.31)**	**0.9943(+0.0101)**	
CIN	62.13	0.9998	1.510 ×
+ **LocMark**	**63.92(+1.79)**	0.9998	

Table 2. Performance of current SOTA watermarking methods integrated with **LocMark** under various distortions

Noise	Factor	HiDDeN	+**LocMark**	MBRS	+**LocMark**	CIN	+**LocMark**
Gaussian Blur	k = 7	34.81	**40.00**	38.91	**42.45**	44.05	**46.86**
		0.9250	**0.9678**	0.9560	**0.9668**	0.9946	**0.9958**
		10.73%	**2.36%**	**0.00%**	3.44%	**0.01%**	0.74%
JPEG	Q = 50	32.92	**41.21**	36.26	**37.73**	53.79	**56.29**
		0.8818	**0.9747**	**0.9269**	0.9166	0.9993	**0.9994**
		0.36%	2.08%	**0.00%**	4.03%	**0.00%**	0.67%
Resize	p = 50%	34.29	**36.81**	45.71	**45.99**	41.02	**44.59**
		0.9142	**0.9327**	0.9832	**0.9854**	0.9929	**0.9950**
		13.67%	**6.69%**	**0.00%**	7.11%	**0.01%**	7.15%
DropOut	p=0.3	33.07	**43.07**	42.95	**49.91**	56.58	**59.59**
		0.8919	**0.9831**	0.9694	**0.9948**	**0.9994**	0.9993
		8.38%	**1.01%**	**0.00%**	3.43%	**0.00%**	2.36%

4. EXPERIMENTS

4.1 Experiment Settings

Implementation Details. To ensure the fairness of the experiments, a series of measures are employed. Following the same experiment setting, we train all baseline models and corresponding models integrated **LocMark**, with the only difference being the embedding regions determined by our methods. All models are trained on a 10,000 images train set and evaluated on a 5000 images validation set from the COCO dataset. The whole experiments are conducted on a single NVIDIA GeForce RTX 4090D.

Metrics. We evaluate the embedding performance in visual quality and extracting watermarking. PSNR and SSIM are used as the default visual quality metric, while BER is adopted to evaluate the extracting performance.

Baselines. Our baseline for comparison are HiDDeN,[3] MBRS[4] and CIN.[5] All the methods are based on deep learning and have been open-sourced. We also try to conduct experiments in CIN, but the results can't replicate the best performance they have reported. Therefore, we uniformly adopt the model we train for a fair comparison.

4.2 Experiment Results

Visual Quality. Our proposed local consistency increases the semantic redundancy of images while filtering out some background interference, which helps the reconstruction of watermarked images. Tab. 1 shows the average objective visual quality metrics of current SOTA watermarking methods integrated with LocMark under free distortion. Watermarks embedded in these highly consistent regions exhibit better visual quality. By embedding watermarks in highly consistent regions to acquire better visual quality, we can embed more watermarks with the same overall PSNR, which achieves a capacity gain. The capacity gain refers to the capacity improvement of ours compared to the baseline methods when the average PSNR of watermark regions is identical, which can be calculated by the inverse ratio of MSE. In addition, we further explore the factors influencing the embedding performance of images and present examples of highly consistent regions in the supplementary material.

Robustness under Different Types of Noise. We test the robustness of current SOTA watermarking methods integrated with LocMark against various types of distortions. Our noise settings are essentially consistent with the baseline model. In Tab. 2, for each distortion, we show PSNR, SSIM and BER in turn. Results show that, for most perturbations, corresponding models integrated with LocMark exhibit better visual quality. However, the decoding rate slightly decreases, since some poor watermarking leads to relatively large changes compared to cover images, which further results in inconsistent segmentation and the misalignment of region localization.

Secrecy against Steganalysis. We measure the secrecy of current SOTA watermarking methods integrated with LocMark by training the Siamese CNN[15] to distinguish between cover and stego images. The Siamese CNN can effectively capture the unnatural noise influenced by watermarks. In Tab. 3, "weights known/unknown"

Table 3. Secrecy of current SOTA watermarking methods integrated with **LocMark**

Detection rate (%)	HiDDeN	+ **LocMark**	MBRS	+ **LocMark**	CIN	+ **LocMark**
weights unknown	99.30	**50.14**	65.64	**50.02**	50.28	**50.04**
weights known	99.90	**50.89**	99.60	**50.09**	50.06	**50.02**

denotes whether the steganalyzer can access the specific model weights. Results illustrate that no matter the steganalyzer can access specific model weights, our variable regions can dilute unnatural noise and disrupt watermarking detection, which significantly reduces the detection rate by steganalysis.

5. CONCLUSION

Starting from the motivation to enhance models' embedding adaptability in diverse cover images, we identify that leveraging prior knowledge, termed *local consistency*, significantly contributes to the generation of watermarked images. To fully exploit this nature, by adaptively embedding watermarks to highly consistent regions, we propose a new local framework, **LocMark**, with a region localization mechanism designed. Experiments demonstrate that the proposed **LocMark** based on local consistency can enhance the embedding performance, further improving capacity while maintaining visual quality.

REFERENCES

[1] Huang, Y., Guan, H., Liu, J., and et.al., "Robust texture-aware local adaptive image watermarking with perceptual guarantee," *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[2] PariZanganeh, A., Ghorbanzadeh, G., ShahreBabak, Z. N., Karimi, N., and Samavi, S., "Adaptive blind watermarking using psychovisual image features," in [*2024 13th Iranian/3rd International Machine Vision and Image Processing Conference (MVIP)*], 1–5, IEEE (2024).

[3] Zhu, J., Kaplan, R., Johnson, J., and Fei-Fei, L., "Hidden: Hiding data with deep networks," in [*Proceedings of the European conference on computer vision (ECCV)*], 657–672 (2018).

[4] Jia, Z., Fang, H., and Zhang, W., "Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression," in [*Proceedings of the 29th ACM international conference on multimedia*], 41–49 (2021).

[5] Ma, R., Guo, M., Hou, Y., Yang, F., Li, Y., Jia, H., and Xie, X., "Towards blind watermarking: Combining invertible and non-invertible mechanisms," in [*Proceedings of the 30th ACM International Conference on Multimedia*], 1532–1542 (2022).

[6] He, K., Chen, X., Xie, S., Li, Y., and et.al., "Masked autoencoders are scalable vision learners," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 16000–16009 (2022).

[7] Kahlessenane, F., Khaldi, A., Kafi, R., and Euschi, S., "A dwt based watermarking approach for medical image protection," *Journal of Ambient Intelligence and Humanized Computing* **12**(2), 2931–2938 (2021).

[8] Su, Q., Liu, D., and Sun, Y., "A robust adaptive blind color image watermarking for resisting geometric attacks," *Information Sciences* **606**, 194–212 (2022).

[9] Liu, Y., Guo, M., and et.al., "A novel two-stage separable deep learning framework for practical blind watermarking," in [*Proceedings of the 27th ACM International conference on multimedia*], 1509–1517 (2019).

[10] Ying, Q., Zhou, H., Zeng, X., Xu, H., Qian, Z., and Zhang, X., "Hiding images into images with real-world robustness," in [*2022 IEEE International Conference on Image Processing (ICIP)*], 111–115, IEEE (2022).

[11] Zhang, J., Chen, D., Liao, J., Fang, H., Zhang, W., and et.al., "Model watermarking for image processing networks," in [*Proceedings of the AAAI conference on artificial intelligence*], **34**, 12805–12812 (2020).

[12] Zhang, J., Chen, D., Liao, J., Zhang, W., Feng, H., Hua, G., and Yu, N., "Deep model intellectual property protection via deep watermarking," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(8), 4005–4020 (2021).

[13] Zhang, J., Chen, D., Liao, J., Ma, Z., Fang, H., Zhang, W., Feng, H., Hua, G., and Yu, N., "Robust model watermarking for image processing networks via structure consistency," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[14] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al., "Segment anything," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 4015–4026 (2023).

[15] You, W., Zhang, H., and Zhao, X., "A siamese cnn for image steganalysis," *IEEE Transactions on Information Forensics and Security* **16**, 291–306 (2020).