MuST: Robust Image Watermarking for Multi-Source Tracing

Guanjie Wang[1], Zehua Ma[*1], Chang Liu[1], Xi Yang[1]
Han Fang[2], Weiming Zhang[*1], Nenghai Yu[1]

[1]University of Science and Technology of China
[2]National University of Singapore
wangguanjie@mail.ustc.edu.cn

Abstract

In recent years, with the popularity of social media applications, massive digital images are available online, which brings great convenience to image recreation. However, the use of unauthorized image materials in multi-source composite images is still inadequately regulated, which may cause significant loss and discouragement to the copyright owners of the source image materials. Ideally, deep watermarking techniques could provide a solution for protecting these copyrights based on their encoder-noise-decoder training strategy. Yet existing image watermarking schemes, which are mostly designed for single images, cannot well address the copyright protection requirements in this scenario, since the multi-source image composing process commonly includes distortions that are not well investigated in previous methods, e.g., the extreme downsizing.

To meet such demands, we propose **MuST**, a multi-source tracing robust watermarking scheme, whose architecture includes a multi-source image detector and minimum external rectangle operation for multiple watermark resynchronization and extraction. Furthermore, we constructed an image material dataset covering common image categories and designed the simulation model of the multi-source image composing process as the noise layer. Experiments demonstrate the excellent performance of MuST in tracing sources of image materials from the composite images compared with SOTA watermarking methods, which could maintain the extraction accuracy above 98% to trace the sources of at least 3 different image materials while keeping the average PSNR of watermarked image materials higher than 42.51 dB. We released our code on https://github.com/MrCrims/MuST.

Introduction

With the rapid growth of entertainment and business demand, the creation of composite images using different materials is common in all walks of life. Nonetheless, the incorporation of unauthorized image materials in composite images will infringe the copyright of the original authors and even lead to commercial disputes. Therefore, how to trace the copyright of image materials for image libraries and individual creators from composite images has become an urgent problem in need of resolution.

*Corresponding authors.

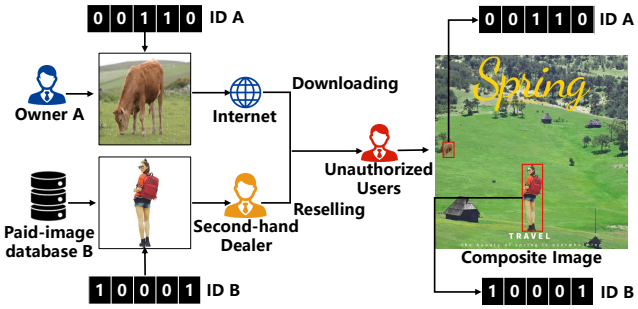


Figure 1: The schematic diagram of the multi-source tracing process. The image watermarking scheme needs to be robust to the multi-source image composing process to verify the multi-source copyright ownership.

Digital watermarking is a commonly used technology to protect the copyright of image owners. The existing watermarking schemes (Zhu et al. 2018; Jia, Fang, and Zhang 2021; Ma et al. 2022) commonly focus on the robustness of an entire image under a series of classical distortions. Yet the composite image is commonly generated by multiple image materials with complex image editing operations, which poses challenges to the existing watermarking schemes.

As shown in Figure 1, the unauthorized user commonly takes the primary component of different image materials to create a composite image with a specific theme. These unauthorized image materials commonly come from multiple sources, including a vast number of personal images on the internet (Kroll 2023) and various image libraries that are being resold without authorization (Li 2019). To make the composite image more natural, attackers will employ complex image editing operations, which are summarized as two main distortions: image pasting distortions and image fusion distortions. Specifically, image pasting distortions (**IPD**) occur when unauthorized users paste image materials onto background images, typically including irregular cropping, resizing, and positional changes. The image fusion distortions (**IFD**) commonly include complex image processing operations, such as feathering and contrast transformation, to better fuse the image material and the background image. For the existing watermarking schemes, the image past-

ing distortions are the main culprit of performance degradation, because irregular cropping, extreme resizing, and positional change operation make it difficult to recover the watermarked image materials to the desirable input of the decoder from the composite image. Therefore, this difficulty leads to destroying the synchronization established in the training process between the encoder and decoder and causes the failure of multi-source image tracing.

In the context mentioned above, other methods for multi-source tracing, such as image retrieval and image matching, face a significant challenge. The images in these scenarios undergo substantial distortion, making it difficult to achieve precise results. Even with slight distortions, there may be more accurate matches; however, due to numerous copies of images spread across the Internet, tracing the source of an image remains challenging as it could not verify if the image copy owner is the actual copyright owner, even if it's an accurate match.

To address the shortcomings of current watermarking algorithms in the aforementioned situations, we propose the **MuST**, a multi-source tracing watermarking scheme with a multi-source image detector and minimum external rectangle (MER) operations to resynchronize the image materials from the composite image. Then, watermark messages are independently extracted from these synchronized image materials for multi-source copyright protection. In summary, we have the following contributions:

- We proposed a novel watermarking scheme, named MuST, for multi-source image tracing. Based on the detector and the MER operation, the proposed MuST can resist the extreme distortions introduced by the multi-source image composing process and achieve impressive performance in various real-world composing scenarios.
- We designed a concise and effective simulation module of the multi-source image composing process, which is leveraged as the noise layer to enhance the robustness of the proposed MuST.
- We selected and constructed a dataset of image materials with the mask of the primary component of the images (Single-Object Image Material Dataset (SOIM)) based on the existing datasets to better simulate the multi-source image composing process.

Related Work

Image Watermarking. Image watermarks can be used for copyright statements. Traditional watermarking algorithms typically embed information using image transformations, such as DCT and DWT. LGDR (Ma et al. 2021) embed symmetric watermarks to obtain robustness against local geometric distortions, global geometric distortions, common image processing operations, and some kinds of combined attacks. In recent years, with the development of deep learning, many works that aim to embed watermarks in images have been developed, such as HiDDeN (Zhu et al. 2018) and StegaStamp (Tancik, Mildenhall, and Ng 2020). Specifically, HiDDeN proposed an autoencoder-like architecture to jointly train an encoder and a decoder for information embedding and extraction. Based on this, StegaStamp further enhanced robustness to distortions resulting from real-world printing and photography. Based on the encoder-decoder framework, MBRS (Jia, Fang, and Zhang 2021), utilizing the Squeeze-and-Excitation blocks (Hu, Shen, and Sun 2018) and proposing a message processor to expand the message in a more appreciated way, aims to enhance the robustness against JPEG compression. LIM (Jia et al. 2022) is to hide information in a sub-image rather than the entire image and include a localization module to correct the shooting distortions in the end-to-end framework. Different from mainstream encoder-decoder architectures, SSLWM (Fernandez et al. 2022) embeds messages in the image features extracted by ResNet-50 (He et al. 2016) and extracts messages with a group of learnable secret keys. The robustness of SSLWM comes from data augmentation. DIPW (Luo et al. 2023) is designed to enhance the robustness of watermarking methods under deliberate plagiarism. Similar to LIM, DIPW also hides copyright evidence in a patch determined by SIFT. Although these schemes achieve good performance on a single image with common distortions, they are not up to the multi-source image materials tracing task for three reasons: (1) lack of the ability to locate multi images in the composite image. (2) fragile synchronization. (3) lack of robustness under extreme distortions.

Method

In this section, we will initially present the architecture of our proposed MuST method and subsequently elaborate on the training strategy for each model.

Framework Overview

Figure 2 shows the architecture of our proposed MuST, including a Watermark Encoder ENC, a Discriminator DIS, a Multi-source Image Composing Simulation Module MIC, a Detector based on U-Net architecture DET, and a Watermark Decoder DEC.

The inputs of encoder ENC include an image material I_{co} of shape $C \times H \times W$ and a watermark message W_m of length L. The output of ENC is an encoded image I_{en} of the same shape as I_{co}. Then, the multi-source image composing process is simulated based on the proposed MIC, in which the primary components of several I_{en} are reused to generate a composite image I_{com}. After that, the detector model DET locates and segments the possible reused image materials I_{dis}, which are distorted by image pasting distortions and image fusion distortions. Finally, DEC tries to extract the corresponding watermark W'_m from the detected multiple I_{dis} to verify the copyright of different owners.

Training Strategy

Encoder. The encoder is trained to encode the watermark message of length L into the input image material, i.e., the cover image I_{co}, while minimizing perceptual disparity between the input and the output to satisfy the requirement of imperceptibly of the watermark algorithm. First, to ensure the robustness of MuST against image pasting distortions, a preprocess is implemented. For each input image

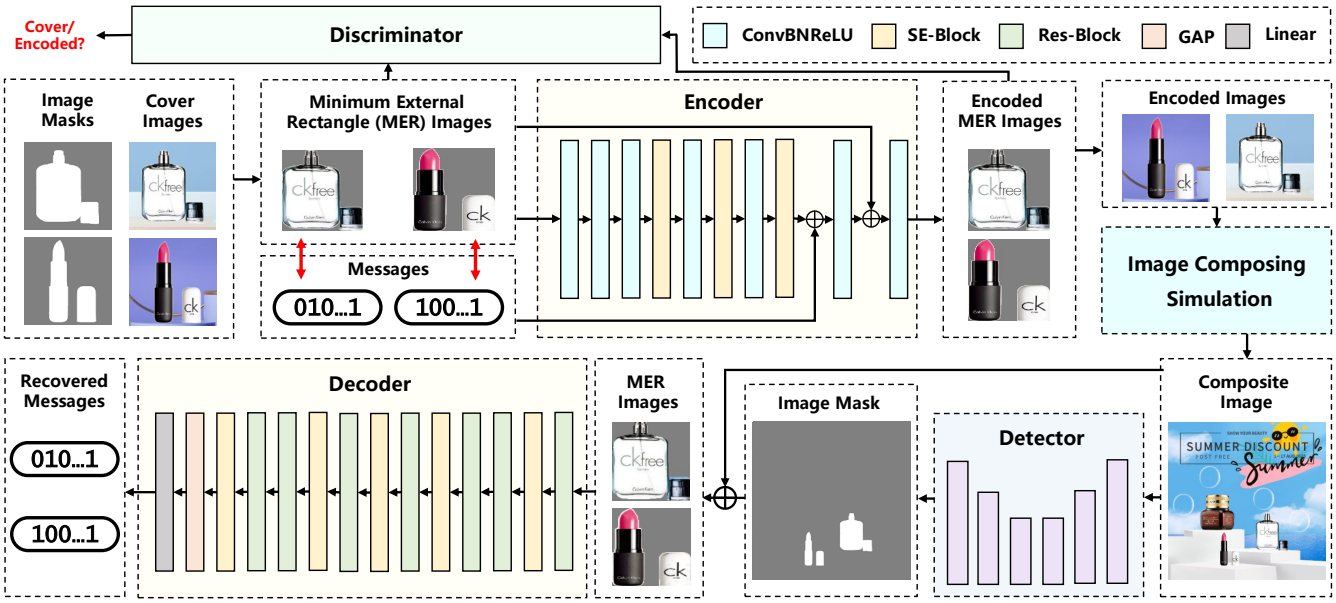


Figure 2: The framework of the proposed MuST. The encoder embeds the watermark messages into the minimum external rectangle (MER) image to generate the encoded images. An adversary discriminator is used to improve the visual quality of the image generated by the encoder. Then, multiple encoded images are fed into the proposed multi-source image composing simulation model to generate the composite image. The detector estimates the image mask from the composite image to segment the MER images. Finally, the decoder extracts the watermark message from each MER image.

denoted as I_{co}, its primary components are extracted using the corresponding mask and subsequently resized to obtain the minimum external image I_{mer}. To simplify the representation, we define $MER(\cdot)$ to refer to this preprocessing, i.e., $MER(I_{co}) = I_{mer}$. Then, these processed image elements are fed into the following convolution layers to obtain an intermediate image feature, in which SE Block (Hu, Shen, and Sun 2018) is adopted to enable the learnable channel-wise weights. After that, the intermediate image feature is concatenated with the secret message expanded to the same dimensions, and fed together into the subsequent convolution layers. Finally, the encoded image element is restored through inverse resizing and then combined with the background of the cover image to generate the watermarked image material I_{en}.

To guarantee the visual quality of the encoded image, we introduce the basic loss $\mathcal{L}_{ENC}$, which calculates the mean squared error (MSE) between $MER(I_{co})$ and $MER(I_{en})$:

$$\mathcal{L}_{ENC} = MSE(MER(I_{co}), MER(I_{en})). \quad (1)$$

Besides, inspired by the design of GAN (Goodfellow et al. 2020), we introduce the adversarial loss $\mathcal{L}_{ENC2}$ to further enhance the visual quality of I_{en}, which make ENC aim at generating indistinguishable $MER(I_{en})$ by the additional adversarial discriminator:

$$\mathcal{L}_{ENC2} = \log(1 - DIS(MER(I_{en})). \quad (2)$$

Meanwhile, the discriminator DIS aims to minimizing the loss L_{DIS}:

$$\begin{aligned}\mathcal{L}_{DIS} = &\log(DIS(MER(I_{en})) \\ &+ \log(1 - DIS(MER(I_{co})).\end{aligned} \quad (3)$$

Image Composing Simulation. By conducting surveys, interviews with design professionals, and studying widely available Photoshop tutorials on the internet, we proposed a succinct and universally applicable multi-source image composing process and designed the corresponding model as the noise layer of MuST, i.e., the Image Composing Simulation module as shown in Figure 2. As mentioned in `Introduction`, the multi-source image composing process includes two main distortions: image pasting distortions and image fusion distortions.

Concerning the resizing operations within the image pasting distortion, our analysis of the actual image composition process reveals a tendency among unauthorized users to downsize the image materials for more optimal placement on the background image. This is easy to understand because enlarging image materials will only make them blurry, and finding image materials with the exact right size is extremely rare. Typically, unauthorized users look for image materials with sufficient resolution and downsize them to obtain the proper image components. Therefore, the image pasting distortions in the noise layer include irregular cropping, extreme downsize, and positional change.

After experiencing image pasting distortion, the primary image component is placed on the background image to generate the composite image, whose overall aesthetics are im-

proved by image fusion operations. Specifically, image fusion operations, i.e., image fusion distortion for the watermarking scheme, include edge moving, feathering, smoothing, brightness variation, and contrast adjustment. Noticing the relationship between certain distortions in the IFD and common image processing operations, we design neural network implementations corresponding to these image editing distortions. Specifically, the objective of edge moving is to mitigate the remaining background residue on the primary image component following irregular cropping, which is implemented through slight cropping in the noise layer. The feathering operation is employed to create a smoother transition between the cropped image material and the background. Within the noise layer, we utilize Gaussian blurring with a mask to achieve a comparable effect. Commonly, unauthorized users also employ global smoothing to make each component of the composite image unobtrusive and aesthetically pleasing. Similarly, the Gaussian filtering function is adopted to simulate the smoothing operation.

Considering that it is hard to simulate human aesthetics, we just randomly place several image materials non-overlapping on the background image and apply the two distortions described above to generate a composite image. It should also be noted that according to the zigzag order of image elements in the composite image, the corresponding order of embedded watermark messages will also be rearranged to facilitate the calculation of decoder loss.

Detector. We use the U-Net (Ronneberger, Fischer, and Brox 2015) as our DET to predict the segmentation mask M of reused image elements in the composite image. When training DET, the ground truth Gt is generated by the image composing simulation process and the loss is:

$$\mathcal{L}_{DET} = -Gt \log(M) - (1 - Gt) \log(1 - M). \quad (4)$$

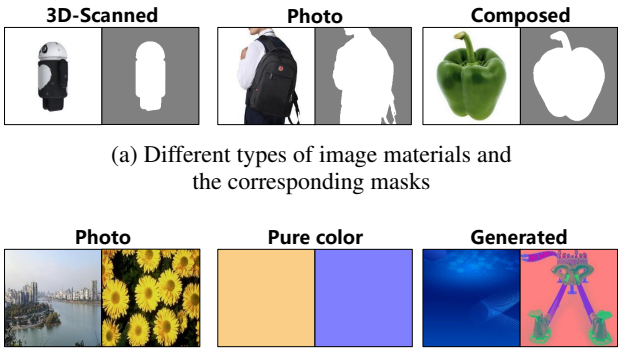
Decoder. According to the segmentation mask calculated by the detector, each distorted image component $MER(I_{com})^i = I^i_{dis}$ is extracted from the composite image based on the connected component labeling algorithm. Then, these extracted components are rearranged in zigzag order and serve as inputs for the decoder. The decoder comprises multiple Res Blocks and SE Blocks, followed by the application of global adaptive average pooling (GAP) to obtain a channel-only feature. Finally, the recovered message W'_m is obtained through a linear mapping layer. The objective of decoder training is to minimize the MSE loss between W_m and W'_m. The loss function L_{DEC} is formulated as follows:

$$\mathcal{L}_{DEC} = MSE(W_m, W'_m). \quad (5)$$

Training. We jointly train ENC, DEC, and DIS, and the training loss of ENC and DEC can be formulated as follows:

$$\mathcal{L} = \lambda_{ENC}\mathcal{L}_{ENC} + \lambda_{ENC2}\mathcal{L}_{ENC2} + \lambda_{DEC}\mathcal{L}_{DEC}. \quad (6)$$

More details of the network architectures and the noise layer are given in the supplemental material.



(a) Different types of image materials and the corresponding masks

(b) Different types of background images

Figure 3: Examples of the image dataset. (a) Examples of SOIM dataset; (b) Examples of background images used in the noise layer.

Experiments

Experiment Settings

Datasets. To facilitate the network training of our proposed scheme, we constructed a new dataset, namely the Single-Object Image Material Dataset (SOIM). It comprises 6.5k image materials with white background, which were selected from JD Product-10k (Bai et al. 2020), Google Scanned-Objects (Downs et al. 2022), and GroceryStore-Dataset (Klasson, Zhang, and Kjellström 2019). Furthermore, each image material in SOIM is paired with a corresponding mask of the primary component, as shown in the first line in Figure 3. These masks are adopted to segment the primary component, which is placed on the background image to generate the composite image in the proposed multi-source image composing simulation model. Some examples of background images are shown in the second line of Figure 3. In our experiments, 6k image materials were used for training and 0.5k were dedicated to testing. Besides, to validate the generalization capability of MuST across different types of image materials, the CASIA V2.0 dataset (Dong, Wang, and Tan 2013; Pham et al. 2019) and Stanford Background dataset (ICCV09) (Gould, Fulton, and Koller 2009) are used as extra testing datasets.

Implementation Details. The whole framework is implemented by PyTorch (Paszke et al. 2019) and executed on NVIDIA RTX A6000. We utilize AdamW (Loshchilov and Hutter 2017) as the optimizer of our models. The watermark messages are randomly generated sequences of 30 bits. In the proposed multi-source image composing simulation model (MIC) implemented by Kornia(E. Riba and Bradski 2020), the input image materials are of size $3 \times 640 \times 640$ pixels and the background images are of size $3 \times 1000 \times 1000$. The parameter settings in MIC are as follows:

- Gaussian blur: $\sigma \in [0.1, 1]$
- Resize: $Scale\ Rate \in [0.3, 0.4]$
- Brightness adjustment: $[-0.2, 0.2]$
- Contrast adjustment: $[0.8, 1.2]$

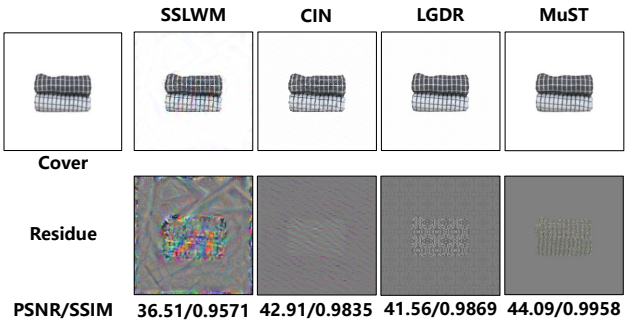
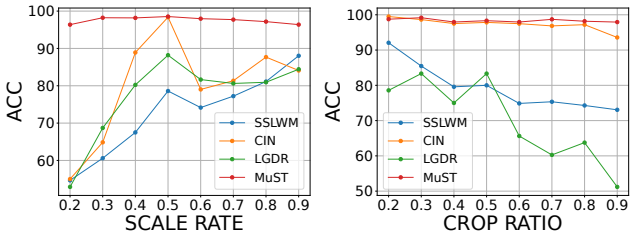


Figure 4: Visual comparisons of watermarked images of MuST and the comparison methods SSLWM, CIN, and LGDR. The upper row showcases the cover image and the watermarked images, while the lower row illustrates the watermark residue. At the same distortion setting, the extraction accuracies of these watermarked images are 83.33%, 86.67%, 83.33%, and 100%, respectively.

Methods	SSLWM	CIN	LGDR	MuST
PSNR	35.12	42.56	42.10	**42.89**
SSIM	0.9571	0.9792	0.9854	**0.9883**

Table 1: The average PSNR (dB) and SSIM (%) of each method tested on 556 images from SOIM, CASIA V2.0, and ICCV09.



(a) Impact of downsizing. (b) Impact of irregular cropping.

Figure 5: The performance of MuST and comparative watermark schemes under main distortions in image pasting distortions. CROP RATIO refers to the removed part.

$\overline{ACC}$(%)	Brightness		Contrast		Gaussian Blur (σ)		
Parameters	-50%	+50%	-50%	+50%	0.1	1	2
SSLWM	95.70	95.13	95.53	95.56	99.66	97.39	94.02
CIN	98.09	**98.76**	98.34	98.08	**99.71**	98.17	97.73
LGDR	**99.68**	84.83	**99.62**	90.33	99.62	94.35	61.20
MuST	98.33	98.41	98.79	**98.28**	98.42	**98.38**	**98.13**

Table 2: For each single type of noise in IFD, we tested under different ratios independently and exhibited some representative results. Bold results indicate the best outcome among the comparisons, while underlined results signify the second-best outcome.

For the loss function in the Eq. (6) , we choose $\lambda_{ENC} = 0.7$, $\lambda_{ENC2} = 0.001$, $\lambda_{DEC} = 2.0$. The batch size in the training is set to 3, and the MuST models are trained for 500 epochs with an initial learning rate $= 0.0001$.

Baselines. Our baselines for comparison are SSLWM (Fernandez et al. 2022), and, CIN (Ma et al. 2022), and LGDR (Ma et al. 2021). All the methods except LGDR are deep-learning-based. Although we attempted to conduct experiments using CIN, we were unable to replicate their reported best performance under conditions aligned to MuST. Therefore, we used the well-trained model they had released for comparison. For a fair comparison, we assisted CIN by providing it with the encoded images segmented from the composite images, as CIN lacks the ability to detect or segment the watermarked image components.

Metrics. The peak signal-to-noise ratio (PSNR (Almohammad and Ghinea 2010)) and structural similarity (SSIM (Wang et al. 2004)) are used to evaluate the visual quality of the encoded image materials. The average extraction accuracy ($\overline{ACC}$), i.e., the correct percentage of the extracted watermark message, is adopted to evaluate the robustness under the multi-source image composing process.

Experimental Results

Visual Quality. Table 1 shows the average objective metrics of the visual quality of the comparative methods and the proposed MuST. The proposed MuST has the highest PSNR and SSIM of 42.89 dB and 0.9883, respectively. Figure 4 presents a representative example of the encoded image and the watermark residue of each method. More examples can be found in the supplementary material.

Robustness under Different Distortions of MIC. It should be noted that the distortions introduced by the proposed MIC pose serious challenges for most existing watermarking schemes, resulting in poor extraction performance. Therefore, in this subsection, we evaluate the robustness of comparative watermarking schemes under several independent distortions of MIC, aiming to derive meaningful experimental results for further analysis and discussion.

First, we evaluated the robustness of these watermarking schemes under two main distortions belonging to image pasting distortions: downsizing and irregular cropping, meaning cropping along the edges of the primary component of the image. As shown in Figure 5, MuST demonstrated its SOTA performance with various distortion parameters, indicating the excellent robustness of MuST under image pasting distortions. Besides, we notice that both LGDR and CIN exhibited unusually good performance when their watermarked images were downsized to 50%. The watermark generation process of LGDR leverages one doubly upsampling 2D matrix to represent watermark bits for better robustness to image processing operations, which also results in the watermark component being sharper at 50% downsizing. For CIN, we speculate that its watermark component has similar characteristics, resulting in similar results. Be-

$\overline{ACC}$ (%)	Number of Materials					
Scale Rate	1	2	3	4	5	6
0.2	96.37	95.92	95.71	95.23	95.32	95.07
0.3	98.22	98.17	98.25	98.03	97.85	-
0.4	98.34	98.02	98.25	97.87	-	-
0.5	98.54	98.27	98.07	-	-	-
0.6	97.97	97.23	-	-	-	-
0.7	97.72	98.02	-	-	-	-
0.8	97.18	96.93	-	-	-	-
0.9	96.38	96.29	-	-	-	-

Table 3: The average extraction accuracy ($\overline{ACC}$ (%)) on different numbers of materials. '-' means the background image cannot hold all the image materials.

Type of Material	SSLWM	CIN	LGDR	MuST
SOIM	75.21	85.83	89.28	98.06
CASIA V2.0	75.42	86.15	89.38	96.23
ICCV09	74.79	85.31	88.54	97.46
Average	75.25	85.89	89.28	97.92

Table 4: The average extraction accuracy ($\overline{ACC}$ (%)) of each type of materials under MIC.

sides, for all compared methods, as they lack the detector and MER strategies in MuST, the input of their extractor is an image cropped from the composite image, with the same size as the cover image and centered on the primary component. Therefore, the input image contains part of the background image, which may cause performance degradation.

Then, we mainly evaluated the impact of basic distortions belonging to image fusion distortions on the watermark performance. As shown in Table 2, in most tested parameters, the proposed MuST has a similar performance with CIN and is better than the rest of the compared schemes.

Overall, the proposed MuST demonstrates the best performance among all the independent distortions within MIC, indicating the excellent robustness of MuST to the multi-source image composing process.

Robustness under Different Numbers of Image Materials. Given that the baselines cannot automatically segment different image materials from composite images, we have not included them in this evaluation. Here, we focused on testing the performance of our method under different numbers of image materials in one background image. The results displayed in Table 3 are the average values of $\overline{ACC}$ extracted from all image materials in composite images, indicating that the quantity of image materials does not negatively impact the performance of our watermark extraction.

Robustness under Different Types of Image Materials. To test our method with image materials of varying kinds and styles, we selected image materials not only from our SOIM dataset but also from the CASIA V2.0 and ICCV09 datasets. Samples from these datasets are presented in the first row of Figure 6. For these experiments, we set the noise

Figure 6: The first line exhibits examples of cover images from three different datasets. The second line exhibits the corresponding watermarked images of MuST.

Background	photos	pure color	composed	Average
MuST	98.39	98.03	97.83	98.25

Table 5: The extraction accuracy ($\overline{ACC}$ (%)) under different types of background images.

factor to correspond with each method's best performance, as determined by previous experimental results. It could be found in Table 4 that all the methods consistently exhibited similar performance across varying image types. And among them, the proposed MuST has the best extraction performance. It is necessary to explain that results in Table 4 were obtained under complete MIC unlike experiments in Table 2, thus leading decreased performance.

Robustness under Different Types of Background. As shown in Figure 3 (b), we selected three types of different background images, including photos, pure color background, and composed. We followed the same noise setting in previous experiments. Because we have excluded the influence of background images for SSLWM, CIN, and LGDR for fare comparison in the previous experiments, we only test our proposed MuST. The results are shown in Table 5 demonstrating that the background has little influence to the performance of MuST.

Real-World Performance Assessment with Commercial Image Editors. We selected three commercial image editing software tools, namely Photoshop [1], Canva [2], and Pixso [3], to test our method under real-world scenarios. We used these tools to combine several image materials, which are embedded with different watermark messages, into the background images to generate the composite image. To better simulate the real-world image composing process, the distortions in the composing process depend on image editors rather than the proposed noise layer. Figure 7 presents some examples along with their corresponding extraction accuracy. Additional examples are provided in the supplementary materials. The results underscore the capacity of our proposed MuST to effectively resist distortions of multi-source composing process in the real world.

[1] https://www.adobe.com/products/photoshop.html
[2] https://www.canva.com
[3] https://pixso.cn/

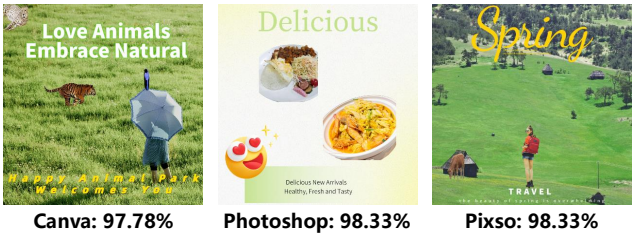


Figure 7: The composite images using watermarked image materials with different image editing software. Under each composite image is the corresponding watermark extracted accuracy of MuST.

Number	1	2	3	4	5	6
TPR (%)	100	100	100	100	98.90	99.16
IoU (%)	99.89	99.67	99.78	99.91	99.89	99.69

Table 6: The true positive rate (TPR) and Intersection over Union (IoU) of the proposed detector with different number of image materials in the composite image.

Detection Accuracy under MIC. In this subsection, we evaluated the true positive rate (TPR) and intersection over union (IoU) of the proposed detector under the MIC. In this paper, TPR refers to the proportion of watermark-containing image materials correctly detected by the proposed detector out of the total actual watermark-containing materials. As shown in Table 6, the proposed detector can accurately identify watermark-containing materials and effectively segment them with accuracy over 98%. It also can be observed from Table 6 that as the number of image materials in the background image increases, there is a slight decrease in TPR. One possible reason is that in order to place more image materials in a limited-size background image, the image materials are downsized further leading to difficult detection.

Ablation Study

Importance of the Noise Layer. As shown in Figure 8, the introduction of the noise layer steers the network towards embedding the watermark within the content area of the image material, rather than across the entire image, which ensures that the watermark can still be extracted when the image material is used in image composing process. The experimental results further prove the above conclusion. As shown in Table 7, under the same distortions, MuST trained with noise layer can achieve better extraction performance.

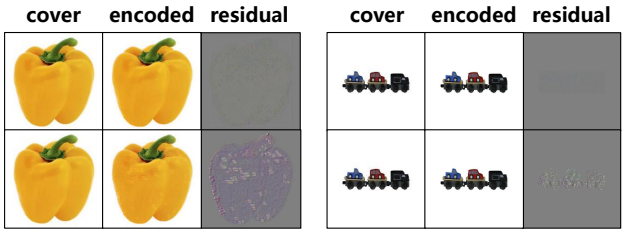


Figure 8: Image materials encoded by the encoder trained with and without noise layer.

Noise Layer	$MER(\cdot)$	$\overline{ACC}$ (%)
✗	✗	50.84
✗	✓	54.43
✓	✗	86.37
✓	✓	98.25

Table 7: The ablation study of noise layer and $MER(\cdot)$. $\overline{ACC}$ (%) is tested on the SOIM dataset under the noise setting shown in the `Implementation Details` part.

Importance of the MER. The main idea of MER is to reconstruct the synchronization of image materials between the encoder and decoder from composite images and improve the robustness under image composing process. As shown in Table 7, with the implementation of MER, all the networks have an improvement of robustness under MIC.

Limitation

This section describes the limitations of MuST. When multiple interconnected image materials containing watermarks are involved, it poses a significant challenge to MuST's performance because of its difficulty in automating the localization and segmentation of watermark-containing materials. Due to limitations on simulating the designer's aesthetics, it's hard to use automated scripts to simulate complex real-world scenarios, making real-world experiments relatively basic. We'll continue to enhance this work in the future and introduce more types of noises, like JpeG compression and change of colors, under which MuST could still maintain good robustness, to evaluate the performance.

Conclusion

In this paper, we first introduced multi-source image materials tracing and analyzed the limitations of existing watermark-based methods against it. In response to these challenges, we introduced an end-to-end framework, MuST, incorporating a unique noise layer specifically tailored to withstand real-world image composing processes. The detection network and MER operation help to reconstruct the synchronization for image materials from composite images, thus enhancing the robustness under distortions. Compared to existing methods, our framework exhibits robust performance under simulated and real-world scenarios.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant 62121002, 62072421, U2336206, 62002334, 62102386, and U20B2047, and the Students' Innovation and Entrepreneurship Foundation of USTC under Grant XY2022X01CY.

References

Almohammad, A.; and Ghinea, G. 2010. Stego image quality and the reliability of PSNR. In *2010 2nd International Conference on Image Processing Theory, Tools and Applications*, 215–220. IEEE.

Bai, Y.; Chen, Y.; Yu, W.; Wang, L.; and Zhang, W. 2020. Products-10k: A large-scale product recognition dataset. *arXiv preprint arXiv:2008.10545*.

Dong, J.; Wang, W.; and Tan, T. 2013. CASIA Image Tampering Detection Evaluation Database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*.

Downs, L.; Francis, A.; Koenig, N.; Kinman, B.; Hickman, R.; Reymann, K.; McHugh, T. B.; and Vanhoucke, V. 2022. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, 2553–2560. IEEE.

E. Riba, D. P. E. R., D. Mishkin; and Bradski, G. 2020. Kornia: an Open Source Differentiable Computer Vision Library for PyTorch. In *Winter Conference on Applications of Computer Vision*.

Fernandez, P.; Sablayrolles, A.; Furon, T.; Jégou, H.; and Douze, M. 2022. Watermarking images in self-supervised latent spaces. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3054–3058. IEEE.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 139–144.

Gould, S.; Fulton, R.; and Koller, D. 2009. Decomposing a scene into geometric and semantically consistent regions. In *2009 IEEE 12th international conference on computer vision*, 1–8. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Jia, J.; Gao, Z.; Zhu, D.; Min, X.; Zhai, G.; and Yang, X. 2022. Learning invisible markers for hidden codes in offline-to-online photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2273–2282.

Jia, Z.; Fang, H.; and Zhang, W. 2021. 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.

Klasson, M.; Zhang, C.; and Kjellström, H. 2019. A Hierarchical Grocery Store Image Dataset with Visual and Semantic Labels. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Kroll, T. 2023. A Single Picture Database Is Worth a Thousand Statutory Damages Awards.

Li, Y. 2019. Vision China Copyright Dispute.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Luo, Y.; Zhou, T.; Cui, S.; Ye, Y.; Liu, F.; and Cai, Z. 2023. Fixing the Double Agent Vulnerability of Deep Watermarking: A Patch-Level Solution against Artwork Plagiarism. *IEEE Transactions on Circuits and Systems for Video Technology*.

Ma, R.; Guo, M.; Hou, Y.; Yang, F.; Li, Y.; Jia, H.; and Xie, X. 2022. Towards Blind Watermarking: Combining Invertible and Non-invertible Mechanisms. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1532–1542.

Ma, Z.; Zhang, W.; Fang, H.; Dong, X.; Geng, L.; and Yu, N. 2021. Local geometric distortions resilient watermarking scheme based on symmetry. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(12): 4826–4839.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pham, N. T.; Lee, J.-W.; Kwon, G.-R.; and Park, C.-S. 2019. Hybrid Image-Retrieval Method for Image-Splicing Validation. *Symmetry*, 83.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Tancik, M.; Mildenhall, B.; and Ng, R. 2020. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2117–2126.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Zhu, J.; Kaplan, R.; Johnson, J.; and Fei-Fei, L. 2018. HiDDeN: Hiding Data With Deep Networks. In *Proceedings of the 15th European Conference on Computer Vision*, 682–697.