A Geometric Distortion Immunized Deep Watermarking Framework with Robustness Generalizability

Linfeng Ma[1], Han Fang[2,*], Tianyi Wei[1], Zijin Yang[1], Zehua Ma[1,*], Weiming Zhang[1], and Nenghai Yu[1]

[1] Anhui Province Key Laboratory of Digital Security, University of Science and Technology of China
{linfengma@mail., bestwty@mail., bsmhmmlf@mail., mzh045@, zhangwm@, ynh@}ustc.edu.cn

[2] National University of Singapore
fanghan@nus.edu.sg

Abstract. Robustness is the most important property of watermarking schemes. In practice, the watermarking mechanism shall be robust to both geometric and non-geometric distortions. In deep learning-based watermarking frameworks, robustness can be ensured by end-to-end training with different noise layers. However, most of the current CNN-based watermarking frameworks, even trained with targeted distortions, cannot well adapt to geometric distortions due to the architectural design. Since the traditional convolutional layer's position structure is relatively fixed, it lacks the flexibility to capture the influence of geometric distortion, making it difficult to train for corresponding robustness. To address such limitations, we propose a Swin Transformer and Deformable Convolutional Network (DCN)-based watermark model backbone. The attention mechanism and the deformable convolutional window effectively improve the feature processing flexibility, greatly enhancing the robustness, especially for geometric distortions. Besides, for non-geometric distortions, aiming at improving the generalizability for more distortions, we also provide a distortion-style-ensembled noise layer, including an image encoder, an image decoder, and distortion-style layers that can effectively simulate styles of different kinds of distortions. Then we can simply train our watermark model with the proposed noise layer for overall robustness. Experiments illustrate the superiority of our method compared to existing state-of-the-art (SOTA) works, such as the 100.00% watermark extraction accuracy under almost all tested geometric distortions.

Keywords: Digital Watermarking · Neural Networks · Style Transfer

1 Introduction

Digital watermarking [10, 29, 33, 56] provides vital support for copyright protection [4, 8, 19, 41, 44]. In recent years, deep watermarking methods with better

* Corresponding authors

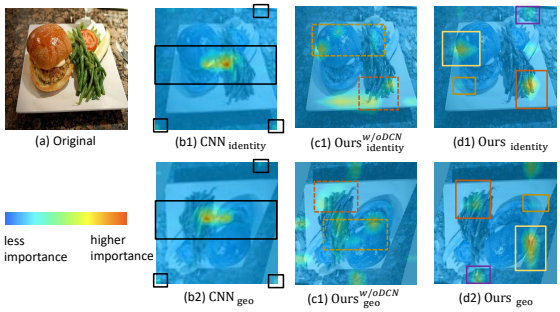


Fig. 1: The processing importance heatmap of the previous CNN-based model [31] (b1, b2) and our model (d1, d2) during watermark extraction without distortion (denoted as identity) or with geometric distortion (rotation degree is 180 and the shear value is 15). (c1) and (c2) are obtained by replacing the DCN in our model with CNN. Heatmaps are drawn with Grad-Cam [47]. As indicated by the marker in the figure, a stronger red color in a location means its higher importance when extracting watermarks, while the blue color means the opposite.

.

performance are proposed, which take being robust against both geometric [21, 30,40,49,55] and non-geometric distortions [2,31,32,37,57] as the paramount demand. The framework mainly includes the watermark encoder, noise layer, and watermark decoder. The watermark encoder embeds the watermark into images, and the watermark decoder makes accurate extraction. Noise layers distort watermarked images, adding disturbance during training to assist in enhancing the model's robustness against related distortions.

The geometric distortions (Rotation, Shear, etc.) will cause the desynchronization problem [26] which may seriously disrupt the originally synchronized correspondence between watermark embedding and extracting, increasing the difficulty for retrieving watermarks [13]. However, even trained with geometric distortions, the existing CNN [35]-based watermark model structure is still vulnerable to them [30, 49]. This has become a stubborn unresolved problem and it stems from their limited internal network mechanisms, such as the geometrically fixed convolution units sampling the image feature at fixed positions, etc., which are not flexible enough when extracting watermarks under geometric distortions [14]. It can be observed from Fig. 1 (b1) and (b2) that positions with high importance weights, as marked by black boxes, are mostly fixed to similar absolute positions (enlarge to get a better view of black boxes close to corners). For example, the marked areas around the center in Fig. 1 (b1) do not shift flexibly with the shear operation in Fig. 1 (b2). So after geometric distortions, the decoder extracts a part of watermarks from asynchronous positions compared to where the watermark is originally embedded, decreasing the extraction accuracy.

To address this limitation, we fill the existing blank in deep watermarking, introducing Swin Transformer [38] and Deformable Convolutional Network (DCN) [14] with smooth adaptions to construct a novel watermark model back-

(a) original **(b)** real_GN **(c)** style_GN **(d)** style_BB **(e)** style_JP

Fig. 2: The visual results of our trained style transfer network.

bone. The more flexible attention mechanism of Swin Transformer improves its adaptive feature processing ability, optimizing the attention for each area based on the whole state of the watermarked image [7, 11, 39]. We notice the advantageous potential of Swin Transformer to fit it into our watermark model backbone, optimizing the watermark extracting process to flexibly focus on watermark-embedded locations under geometric distortions for better accuracy. It can be seen from Fig. 1 (c1) and (c2) that after introducing Swin Transformer, more image parts are utilized for extraction, reducing wasted spaces in watermarked images. Besides, focused areas are not all fixed in certain absolute places anymore, the trend to trace areas originally embedded with watermarks following the geometric distortion can be observed from marked boxes with the same colors, alleviating the desynchronization problem mentioned above. For further enhancing flexibility, we also propose using DCN to replace CNN used by previous watermark backbones. The additional offset to DCN's grid sampling locations with a free form enables DCN to perform better under geometric distortions in an adaptive manner [14]. Fig. 1 (d1) and (d2) demonstrate that almost all of our model's highly focused areas can well correspond following the geometric distortion, as indicated by box pairs with matched colors. This highlights that after adding DCN, the flexibility of our model is further enhanced to extract the watermark precisely in similar areas as where watermarks are embedded, ensuring a consistent synchronization state, thus immunizing the influence of geometric distortions, keeping 100% extraction accuracy in almost all cases.

For non-geometric distortions, the existing watermarking methods propose the combined noise layer [57] for multi-robustness, which contains several distortions, and for each mini-batch during training, one distortion will be randomly chosen as the noise layer. However, it provides deficient knowledge confined to the distribution of the direct implementation of engaged distortions. Thus the correspondingly trained watermark model performs poorly in more varified distortion situations. For better generalizability, we propose a distortion-style-ensembled noise layer, utilizing the style transfer network to simulate various distortion styles. Fig. 2 exhibits the effect of style transfer, in which (b) is the result of adding real Gaussian Noise to the original image (a), while (c), (d), (e) shows the output of our style transfer network simulating the effect of Gaussian Noise, Box Blur, and JPEG Compression, respectively. Detailed visual results about more distortions after style transfer can be seen in the Appendix. As shown by the comparison between ground truth (Fig. 2 (b)) and simulation result (Fig. 2 (c)), our style transfer network makes effective expression about distortion's char-

acteristics, and simultaneously, introduces moderate variance and enriches the expressed distortion features. Thus we can obtain a more diversified distribution representation of distortions, offering our model a broader range of distortion-related knowledge. Consequently, the overall robustness can be further boosted by training with our proposed noise layer, improving our model's generalizability.

Integrating the above-mentioned designs, we propose a geometric distortion immunized deep watermarking framework with robustness generalizability, in which our model backbone and distortion-style-ensembled noise layer can smoothly replace their previous counterparts in classical end-to-end training mode, keeping standard back-propagation. Extensive comparative experiments demonstrate the comprehensive superiority of our method, keeping exactly correct watermark extraction under most geometric distortions. Besides, ablation experiments well justify the effectiveness of our method's components.

Our main contributions can be summarized as follows:

- We propose a watermark model backbone built with Swin Transformer and DCN, delivering superior robustness against geometric distortions.
- A novel distortion-style-ensembled noise layer is provided to effectively enhance our model's generalizability across various non-geometric distortions.
- Extensive experiments show that our model achieves the best results in terms of robustness, generalization ability, and visual quality of encoded images compared with the state-of-the-art deep watermarking methods.

2 Related Work

2.1 Deep Watermarking Technology

As a common solution for copyright protection, digital watermarking [50] gets wide applications [5, 6, 12, 23, 27, 34, 43, 52]. Recently, deep watermarking obtains better performance than traditional methods. Kandi et al. [32] first apply CNN [35] to digital watermarking and achieve better performance compared to previous traditional methods. Zhu et al. [57] propose HiDDeN, a pioneering end-to-end deep watermarking method, getting wide influence in many subsequent works. Different from HiDDeN, Liu et al. [37] propose a novel two-stage deep watermarking framework to ingeniously avoid the simulation of undifferentiable noises, and it achieves satisfactory performance. Jia et al. [31] propose a watermarking method named MBRS. To improve the robustness against JPEG compression, for each mini-batch during training, the noise layer is randomly chosen from JPEG compression, JPEG Mask [57], and a noise-free layer. Ma et al. [40] introduce the invertible neural network (INN) [15] into the watermarking task for the first time, which achieves good performance. However, previous methods are mainly CNN-based, which lack internal mechanisms to handle the geometric distortions [14], making them less practical for wider applications. Besides, although the combined noise layer [57] is proposed for getting multi-robustness, the obtained progress can still be further enhanced. We build a novel model backbone with superior robustness against geometric distortions, and the style transfer [9,18] is utilized to design our noise layer for better generalizability.

2.2 Vision Transformer

Transformer [51] was originally designed for natural language processing (NLP), which demonstrates promising performance with its powerful attention mechanism. Exploring transplanting Transformer into Computer Vision (CV), Dosovitskiy et al. [16] propose the Vision Transformer (ViT), achieving better performance than previous CNN-based models on various popular datasets. Inspired by ViT, lots of similar works [7, 11, 25, 38, 39] appear recently. Liu et al. [38] propose the Swin Transformer to decrease the computation complexity. It uses a sliding window to speed up the inference of the model and well guarantee the modeling power. Swin Transformer has achieved SOTA performance in many CV tasks, while its application in deep watermarking is still not sufficiently explored. Wang et al. [53] utilize Swin Transformer for a similar application of watermarking: hiding images with images. However, the robustness of this task is relatively weak. Han et al. [24] propose a zero-watermarking scheme based on the Swin Transformer, while the inability to process watermarks with more bits limits its further application. Our method flexibly employs Swin Transformer to build a novel multi-bit watermarking model backbone and performs better than previous methods, filling the existing blank in this field.

3 Method

3.1 Motivation and Overview

For watermarking methods, the key is keeping excellent robustness against various distortions, including: 1) geometric distortions such as Rotation and Shear, which will lead to the serious desynchronization problem, decreasing watermark extraction accuracy significantly [13, 26]; 2) non-geometric distortions including a variety of complex classes, taking Noise, Blur, Compression as representations, which raise the challenge about getting robustness generalizability for them. To perform better for the above two points, CNN-based deep watermarking frameworks are proposed and become mainstream in recent years. Unfortunately, the achieved effect is still limited and needs further improvement in both aspects.

To solve the above limitations, for 1), Swin Transformer [38] and Deformable Convolutional Network (DCN) [14] with powerful feature processing ability to better handle geometric distortions are employed as our model backbone. With its more powerful and adaptive attention mechanism, Swin Transformer can flexibly schedule more attention towards the most vital areas for optimal results, with reference to the image's whole state [7, 11, 39]. Capturing such an advantage, our model's watermark extraction process under geometric distortions can be optimized to consistently allocate most of the attention to needed locations for keeping correctness, which should always match the original watermark embedding areas under various geometric states; see Fig. 1 (d1), (d2). This helps maintain the favorable synchronization state for good accuracy, in which watermarks can be embedded and exactly extracted in similar positions. In addition, by adding offsets to the previous fixed grid sampling location, DCN obtains the

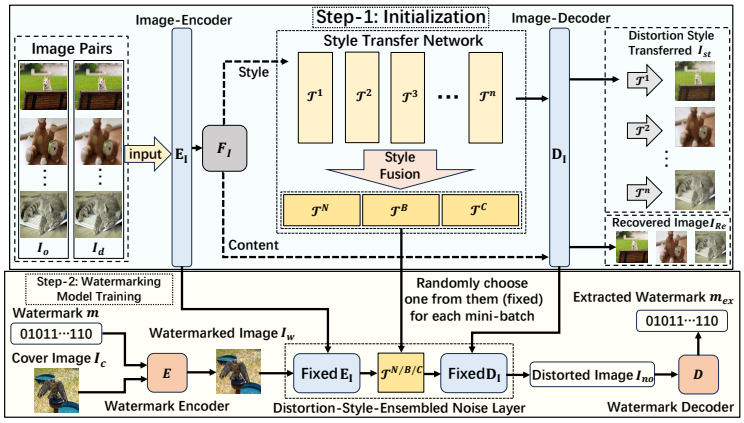


Fig. 3: The overview of our method. In step-1, original and distorted image pairs (I_o, I_d) are collected for training. Image-encoder E_I and image-decoder D_I are trained to extract image features (F_I) and recovered images (I_{Re}). Distortion-style layers $\mathcal{T}^i$ are trained to exert distortion styles. Styles in the same distortion class are fused to simulate the style of whole distortion classes ($\mathcal{T}^N$, $\mathcal{T}^B$, $\mathcal{T}^C$ for Noise, Blur, Compression, respectively). In step-2, our watermarking model is trained end-to-end. Fixed E_I, D_I, $\mathcal{T}^N$, $\mathcal{T}^B$, and $\mathcal{T}^C$ are utilized to form our distortion-style-ensembled noise layer.

free form of its sampling grid, performing more adaptive behavior when processing image features after the geometric distortion [14]. Exploiting the natural characteristics of these two structures for improving the robustness against geometric distortions, with some smooth adaptions, we design a watermark model backbone integrating the merits of Swin Transformer and DCN, achieving immunization against geometric distortions with 100% accuracy in almost all cases.

As for 2), enriching the distortions' distribution representation provided by the noise layer can deliver more distortion-related knowledge and assist in training a watermark model adaptive to a wider range of distortions. Style transfer [9, 18, 22, 42, 46] is effective for simulating various styles, which is accurate enough to report most characteristics of the target distortion style, while moderate variations will also be introduced to form a richer expression than direct implementation. Leveraging such relatively ample expressed distortion features, simulated distortion styles can be fused [9] to simulate entire distortion classes. We consider 3 representative classes in non-geometric distortions, i.e., Noise, Blur, and Compression. Simulated styles are fused following different distortion classes, and then utilized to replace direct distortion implementations in the previous noise layer. With the aforementioned benefit of style transfer, such fused indirect simulations will provide a wider range of distribution representation relating to more distortions during the watermarking model training. Thus, our model adapts to richer distortion coverage with better generalizability.

Based on the above-mentioned analysis, we design a geometric distortion immunized deep watermarking framework with robustness generalizability, which

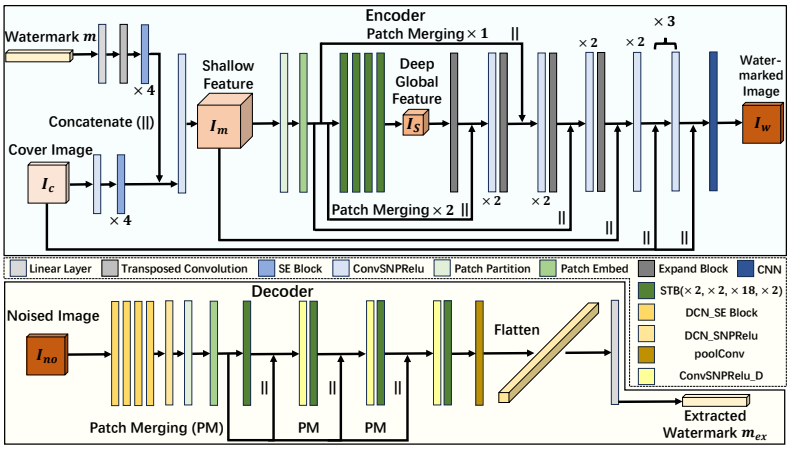


Fig. 4: The structure of our watermarking encoder and watermarking decoder, in which STB is the Swin Transformer Block with Patch Merging [38]. We integrate the Swin Transformer [38] and DCN [14] to build a more flexible watermark model, achieving superior robustness than previous methods, especially against geometric distortions.

consists of a Swin Transformer and DCN-based backbone and a distortion-style-ensembled noise layer, covering a two-step training approach, as shown in Fig. 3. For step-1 initialization, we initialize a style transfer network including an image-encoder, an image-decoder, and distortion-style layers. Image pairs before and after distorting are collected as the ground truth for training, distortion-style layers and the image-encoder/decoder are trained to effectively decouple and then process image style and image content respectively. As for step-2, we construct a novel combined noise layer comprising styles of distortion classes obtained by style fusion in step-1 for the end-to-end training of our watermarking model.

3.2 Structures of Watermark Encoder and Watermark Decoder

Structures of our proposed watermark encoder and watermark decoder are illustrated in Fig. 4. For the watermark encoder, it is mainly consisted of the 'ConvSNPRelu' layer, SE block [28], and the Swin Transformer Blocks (STB) [38], etc. Features of the inputs, including the cover image and watermark information, are firstly processed in the shallow dimension, and then completely coupled in the deep dimension by STB. Finally, shallow features and deep features are fully integrated to generate the watermarked image. For the watermark decoder, it is mainly composed of DCN [14] and STB, etc. Due to transmission distortions or malicious attacks towards the watermarked image, the input to the watermark decoder is usually a noised image. It is firstly pre-processed by DCN-related layers, then features are further extracted in the deep dimension by the STB. Finally, a linear layer outputs the extracted watermark. Details about watermark model structures and processing workflow can be seen in the Appendix.

3.3 Step-1: Initialization of Style Transfer Network

As aforementioned, in step-1, we need to obtain a style transfer network that can decouple the style and the image content, and then effectively transfer the original style to the distortion's style. Inspired by the idea of stylebank [9], our style transfer network $\mathbf{N_S}$ consists of an image-encoder $\mathbf{E_I}$, an image-decoder $\mathbf{D_I}$ and distortion-style layers $\mathcal{T}^i \in \mathbf{T}$. The training of $\mathbf{N_S}$ includes two branches, the image-content branch (i.e., $\mathbf{E_I} \rightarrow \mathbf{D_I}$) and distortion style branch (i.e., $\mathbf{E_I} \rightarrow \mathcal{T}^i \rightarrow \mathbf{D_I}$). These two branches share the same parameters of $\mathbf{E_I}$ and $\mathbf{D_I}$.

Training Framework Firstly, we select n distortions from 3 representative non-geometric distortion classes (Noise, Blur, and Compression). Then image pairs are generated, denoted as $\left\{I_o^i, I_d^i\right\} \in \left\{\mathbb{I}_o^i, \mathbb{I}_d^i\right\}, i \in [1, n]$, where I_o^i and I_d^i represents the original image and the distorted image. I_o^i is sent to $\mathbf{E_I}$ to form image feature map F_I. For the image-content branch, F_I will be directly sent to $\mathbf{D_I}$ and get the recovered image I_{Re}^i. For the distortion style branch, the distortion-style layer $\mathcal{T}^i$ for i^{th} distortion's style is inserted between $\mathbf{E_I}$ and $\mathbf{D_I}$. After obtaining F_I, we process it with $\mathcal{T}^i$ and get $\hat{F}_I^i = \mathcal{T}^i(F_I)$, which is the image feature map after style transfer. Then I_{st}^i, the image with transferred style can be obtained after the processing of $\mathbf{D_I}$.

After training two branches, the image content is coupled into $\mathbf{E_I}$ and $\mathbf{D_I}$, the distortion styles are coupled into $\mathbf{T}$. The styles simulated by $\mathcal{T}^i$ that belong to the same class are fused with the linear manner of [9], $\hat{F}_I^i$ will be linearly added with the weight. In our method, the weights of different $\hat{F}_I^i$ are equal.

Finally, with style fusion, we can simulate styles of the above-mentioned 3 distortion classes with correspondingly fused distortion-style layers, denoted as $\mathcal{T}^N$, $\mathcal{T}^B$, and $\mathcal{T}^C$. Along with $\mathbf{E_I}$ and $\mathbf{D_I}$, they will be utilized to form a distortion-style-ensembled noise layer and train our watermarking model in step-2. The training algorithm and model structures can be seen in Appendix.

Loss Function During the training of our style network $\mathbf{N_S}$, the image-content branch and distortion style branch are iteratively trained, and there exists some difference between their loss function. The loss function of image-content branch is $\mathcal{L}_c$, the MSE (mean square error) loss between I_o^i and I_{Re}^i. For distortion style branch, the loss function $\mathcal{L}_\mathcal{T}\left(I_o^i, I_{st}^i, I_d^i\right)$ can be calculated as:

$$\mathcal{L}_\mathcal{T}\left(I_o^i, I_{st}^i, I_d^i\right) = \alpha\mathcal{L}_d\left(I_d^i, I_{st}^i\right) + \beta\mathcal{L}_v\left(I_o^i, I_{st}^i\right) + \gamma\mathcal{L}_s\left(I_d^i, I_{st}^i\right), \tag{1}$$

in which $\mathcal{L}_d$ is the MSE loss between I_d^i and I_{st}^i. And $\mathcal{L}_v$ and $\mathcal{L}_s$ are evaluated by a pre-trained VGG network [45]. α, β, and γ are weights of $\mathcal{L}_d$, $\mathcal{L}_v$, and $\mathcal{L}_s$, respectively. More details including the calculation of $\mathcal{L}_v$ and $\mathcal{L}_s$ are in Appendix.

3.4 Step-2: Watermarking Model Training

Training Framework The classical end-to-end training manner [57] can be kept for our watermark model. Our watermark model includes a watermark

encoder E and a watermark decoder D. The distortion-style-ensembled noise layer includes $\mathbf{E_I}$, $\mathbf{D_I}$, $\mathcal{T}^N$, $\mathcal{T}^B$, and $\mathcal{T}^C$, which are all obtained in step-1 and fixed. E receives the input cover image I_c and the original watermark M as its input, and outputs watermarked image I_w that containing M. I_w is processed by $\mathbf{E_I}$ to extract its image feature F_I. Then one distortion-style layer will be randomly chosen from $\mathcal{T}^N$, $\mathcal{T}^B$, and $\mathcal{T}^C$ to exert distortion class style to F_I and get $\hat{F}_I^i$. $\mathbf{D_I}$ receives $\hat{F}_I^i$ then outputs the distorted watermarked image I_{no}. Finally, D takes I_{no} as input and outputs the extracted watermark m_{ex}.

Loss Function The overall loss function $\mathcal{L}$ for our watermark model's training is composed of 3 losses, which can be formulated as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_M \left(m, m_{ex}\right) + \lambda_2 \mathcal{L}_I \left(I_c, I_w\right) + \lambda_3 \mathcal{L}_S \left(I_c, I_w\right), \tag{2}$$

where $\mathcal{L}_M$ is the MSE loss between m and m_{ex}, $\mathcal{L}_I$ is the MSE loss between I_c and I_w, $\mathcal{L}_S$ is the SSIM loss [54] between I_c and I_w. $\mathcal{L}_M$ is used for optimizing the watermark extraction accuracy, $\mathcal{L}_I$ and $\mathcal{L}_S$ are used for improving the visual quality of watermarked images. λ_1, λ_2, and λ_3 are weights for $\mathcal{L}_M$, $\mathcal{L}_I$, and $\mathcal{L}_S$, respectively. The whole watermarking model is trained in the end-to-end mode according to $\mathcal{L}$. Besides, it is worth noting that all parameters belonging to our well-trained style transfer network are fixed during step-2. More details about Equation 2 can be seen in Appendix.

4 Experiment Evaluation

4.1 Implementation Details and Metrics

To train the style transfer network, we choose DIV2K [1] as our training dataset. Gaussian Noise ($\sigma = 0.05$), Poisson Noise ($\lambda = 0.3$), and Salt-and-Pepper (SP) Noise ($p = 0.15$) are chosen to be simulated and fused to represent the Noise class. Gaussian Blur ($\sigma = 6$), Median Blur (kernel size $= 7 \times 7$), and Box Blur (kernel size $= 5 \times 5$) are chosen to be simulated and fused to represent the Blur class. As the most representative compression algorithm, JPEG compression ($Q = 50$) is chosen to represent the Compression class. α, β, and γ in Equation 1 are set as 1, 0.01, and 0.0001. For the watermark model training in step-2, we randomly choose 5000 images from COCO dataset [36] as our training dataset. Data augmentation will be randomly applied to each mini-batch during training for better generalizability. λ_1, λ_2, and λ_3 in Equation 2 are set as 3, 1, and 1. Other randomly chosen 500 images from COCO dataset compose our testing set for both two steps. The images' size is standardized to $3 \times 224 \times 224$ during the experiment and the length of the randomly generated watermark is 196 bits. The learning rate is respectively set to 0.001 and 0.0001 in step-1 and step-2.

For non-geometric distortions, our proposed distortion-style-ensembled noise layer is used for the multi-robustness, as mentioned in Section 3.3. For geometric distortions, on the one hand, Cropout, Dropout, Crop, and Resize are deemed

Table 1: The detailed parameters of different models in the comparative experiment.

Model	SSL [21]	MBRS [31]	CIN [40]	Ours
Image Size	$3 \times 128 \times 128$	$3 \times 128 \times 128$	$3 \times 128 \times 128$	$3 \times 224 \times 224$
Message Length	64 bits	64 bits	64 bits	196 bits
Information Density	0.003906 bits/pixel	0.003906 bits/pixel	0.003906 bits/pixel	**0.003906 bits/pixel**
PSNR(wo G)	25.562 dB	35.382 dB	33.506 dB	**35.973 dB**
SSIM(wo G)	0.765	0.924	0.949	**0.992**
PSNR(w G)	25.562 dB	31.830 dB	24.180 dB	**33.720 dB**
SSIM(w G)	0.765	0.921	0.738	**0.986**

incapable of bringing the serious desynchronization problem. And most representative deep watermarking schemes [17, 21, 31, 37, 40, 57] test them together with non-geometric distortions in experiments rather than using them to evaluate the robustness against geometric distortions, so we follow the same experimental settings as previous works. On the other hand, considering that Rotation, Shear, and Affine Transformation (including both Rotation and Shear), which are all usually applied during image processing, will representatively cause a significant desynchronization problem, we choose Affine Transformation as the noise layer to train for corresponding robustness, with rotation angle and shear value randomly chosen from $(-180, +180)$ and $(-30, +30)$ for each mini-batch. Besides, these three are chosen to evaluate the robustness under geometric distortions. The detailed explanation of included distortions and their parameters in experiments can be seen in the Appendix.

To evaluate robustness, the bit correct rate (BCR) of the extracted watermark is used. Meanwhile, Peak Signal-to-Noise Ratio (PSNR) [3] and Structural Similarity (SSIM) [54] are used to evaluate the visual quality impact of the encoder to cover images.

4.2 Baseline

Baselines used in comparative experiments are SSL [21], MBRS [31], and CIN [40]. They are all SOTA deep watermarking methods that are claimed or practiced to be robust against both geometric and non-geometric distortions. For MBRS and CIN, we use the same dataset as ours to re-train their two models. For SSL, we use its released pre-trained model recommended by its authors. For geometric distortions, the same noise layer mentioned above is used. For non-geometric distortions, the traditional combined noise layer including the same 7 directly implemented distortions and parameters mentioned in Section 4.1 is used. To achieve the differentiable training, JPEGSS [48] is used to replace the original JPEG compression. The images' size is standardized to $3 \times 128 \times 128$ for baselines and the length of the randomly generated watermark is 64 bits, maintaining the same information density as ours. Table 1 shows detailed parameters of methods in the comparative experiment, in which PSNR (wo G), SSIM (wo G), and PSNR (w G), SSIM (w G) evaluate the visual quality under non-geometric (wo G) or geometric distortions (w G), respectively. It can be seen from Table 1

Table 2: BCR of different models under geometric distortions. Our method achieves 100% accuracy except for scarce extreme cases with weak image availability.

Method	Shear			Rotation				
	$s = 30$	$s = 40$	$s = 50$	$r = 60$	$r = 90$	$r = 120$	$r = 150$	$r = 180$
SSL [21]	90.28%	85.06%	80.31%	85.85%	84.49%	84.39%	81.64%	81.36%
MBRS [31]	87.48%	87.22%	85.34%	88.26%	88.11%	88.04%	87.80%	87.69%
CIN [40]	84.02%	81.16%	77.91%	83.22%	83.07%	82.98%	82.77%	82.71%
Ours	**100.00%**	**99.35%**	**93.60%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
Method	Affine Transformation							
	$s = 10, r = 30$	$s = 10, r = 60$	$s = 20, r = 90$	$s = 20, r = 120$	$s = 30, r = 150$	$s = 30, r = 180$	$s = 40, r = 180$	$s = 50, r = 180$
SSL [21]	90.05%	85.03%	82.73%	81.53%	76.62%	75.91%	72.43%	69.59%
MBRS [31]	88.03%	87.98%	87.76%	87.64%	87.40%	87.28%	87.05%	85.19%
CIN [40]	82.96%	82.32%	82.29%	82.08%	81.04%	80.94%	80.66%	75.06%
Ours	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**99.21%**	**93.37%**

(a) SSL [21] **(b)** MBRS [31] **(c)** CIN [40] **(d)** Ours

Fig. 5: Outputs of different models under geometric distortions, including an example of the original image (1st row), encoded images (2nd row), and residual images (3rd row). For easy visibility, all residual images are multiplied by 2 and shifted by 128.

that under the same information density, our method can achieve the highest PSNR and SSIM under different distortions, especially the values of SSIM. In the Appendix, we also consider the situation when taking AI-generated images (AIGI) as the input of the watermark model, and compare the performance of our method with Stable Signature [20], the SOTA AIGI watermarking method. Experimental results in the Appendix show our method's better robustness and visual quality of watermarked images compared to Stable Signature.

4.3 Evaluation under Geometric Distortions

Experimental results about geometric distortions are shown in Table 2, in which the value of Shear is randomly chosen from $(-s, +s)$ and the angle of Rotation is randomly chosen from $(-r, +r)$ for each image. Affine Transformation includes both Rotation and Shear. Our method achieves 100% accuracy in nearly

Table 3: BCR of different models under non-geometric distortions.

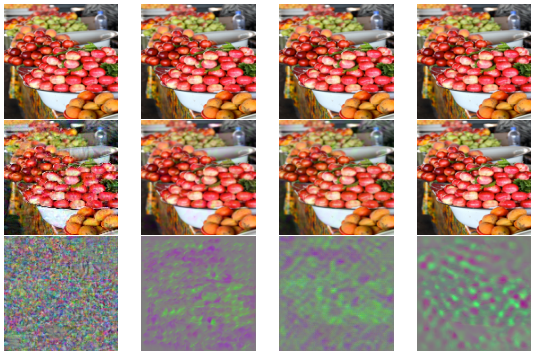
Method	GB $\sigma = 6$	MB 7×7	BB 7×7	GN $\sigma = 0.04$	SPN $p = 0.25$	JPEG $Q = 40$	DP $p = 0.7$
SSL [21]	88.21%	69.71%	76.26%	97.03%	61.79%	74.83%	80.99%
MBRS [31]	97.05%	95.45%	95.48%	96.93%	99.79%	96.48%	79.57%
CIN [40]	96.77%	96.33%	96.61%	96.60%	98.81%	71.54%	84.58%
Ours	**100.00%**	**100.00%**	**100.00%**	**98.09%**	**100.00%**	**99.93%**	**95.44%**
Method	CP $p = 0.5$	Crop $p = 0.2$	Resize $p = 0.4$	Contrast $[0.1, 1.9]$	Hue $[-0.2, 0.2]$	Saturation $[0.1, 1.9]$	Bright $[0.2, 1.8]$
SSL [21]	90.83%	91.34%	68.29%	96.63%	96.69%	97.72%	95.73%
MBRS [31]	94.65%	95.43%	88.41%	97.38%	95.60%	98.30%	98.43%
CIN [40]	89.96%	94.38%	95.53%	98.76%	91.17%	98.82%	98.66%
Ours	**96.87%**	**96.75%**	**99.38%**	**100.00%**	**99.60%**	**100.00%**	**99.97%**

all tested cases, such as Rotation with arbitrary angles and Affine Transformation with most of the parameters, demonstrating excellent immunization ability against geometric distortions. Besides, the superiority compared to other methods is also significant, improving the average BCR of around 80% achieved by previous methods to almost 100%. Although in scarce extreme cases, the accuracy is a little lower than exact correctness, the availability of images is drastically destroyed under these situations, which can be seen in the Appendix.

Fig. 5 shows the comparison between the output samples of different methods. The overall modification generated by our method has an explicitly slight magnitude, maintaining the superior visual quality of our **watermarked** images. However, other comparative methods all exhibit a more apparent trend to strengthen the degree of modification, which may help maintain limited accuracy under geometric distortions. The experimental data demonstrates the unsatisfactory convergence of CNN-based models under geometric distortions, further highlighting the superior immunization of our method against geometric distortions compared to previous model structures.

4.4 Evaluation under Non-Geometric Distortions

In this subsection, we test methods in comparative experiments under non-geometric distortions, including Gaussian Blur (GB), Median Blur (MB), Box Blur (BB), Gaussian Noise (GN), Salt-and-Pepper Noise (SPN), and JPEG Compression (JPEG), and they mostly do not follow the same parameters as in training. Distortions that do not belong to 3 distortion classes in our distortion-style-ensembled noise layer are also used to evaluate the generalization ability, consisting of Cropout (CP), Dropout (DP), Crop, Resize, and image tone adjustment, including contrast, hue, saturation, and brightness. As shown in Table 3, under various non-geometric distortions, our method consistently shows the prominently highest BCR. For example, our method achieves excellent extraction accuracy under JPEG (Q=40) (99.93%) and Dropout ($p = 0.7$) (95.44%), at least 3.45% and 10.86% higher than other methods. This indicates the prominent generalization ability of our method.



(a) SSL [21] **(b)** MBRS [31] **(c)** CIN [40] **(d)** Ours

Fig. 6: Outputs of different models under non-geometric distortions, arranged as Fig. 5.

Fig. 6 shows outputs of different models under non-geometric distortions. It can be seen that our model embeds the watermark in a mode more suitable for various distortions. Specifically, compared to SSL which mainly embeds watermarks in the high-frequency domain, thus lacking the robustness against JPEG compression [57] and blurs, our method does not mainly rely on the high-frequency domain, leading to better generalizability for more distortions. This can also be demonstrated by relevant data in Table 3. All of the above experimental results demonstrate the superior ability of our model to handle non-geometric distortions, and the effectiveness of our distortion-style-ensembled noise layer in better improving the watermark model's generalization ability.

4.5 Ablation Study

Replacing CNN with DCN To evaluate the impact of replacing some CNN in the watermark decoder with DCN, we change the DCN in our watermark decoder with traditional CNN and re-train this model, called Ours$_{woDCN}$, with the same setting of our model's training. Ours$_{woDCN}$ achieves the PSNR value of 32.415dB and SSIM value of 0.978, which are lower than our model; see Table 1. More details about the output image of Ours$_{woDCN}$ can be seen in Appendix. The experimental results are shown in Table 4. We find that Ours$_{woDCN}$ exhibits high extraction accuracy under various geometric distortions, yet still at least 2.11% lower than our proposed model, indicating the effectiveness of DCN in terms of boosting geometric distortion robustness.

The Distortion-Style-Ensembled Noise Layer To evaluate the effect of replacing the previous traditional combined noise layer with our distortion-style-ensembled noise layer, we re-train our model with the same combined noise layer mentioned in Section 4.2, called Ours$_c$, and MBRS with our distortion-style-ensembled noise layer mentioned in Section 4.1, called MBRS$_s$. Ours$_c$ and

Table 4: BCR of our model with or without DCN under geometric distortions.

Method	Shear		Rotation		Affine Transformation				
	$s = 30$	$s = 40$	$r = 90$	$r = 150$	$s = 20, r = 120$	$s = 30, r = 150$	$s = 30, r = 180$	$s = 40, r = 180$	$s = 50, r = 180$
Ours_{woDCN}	97.88%	95.72%	97.89%	97.82%	97.77%	97.58%	97.50%	95.36%	90.44%
Ours	**100.00%**	**99.35%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**99.21%**	**93.37%**

Table 5: BCR of different models under non-geometric distortions.

Method	JPEG $Q = 40$	MB 7×7	BB 8×8	GB $\sigma = 8$	CP $p = 0.7$	DP $p = 0.7$	Crop $p = 0.2$	Resize $p = 0.6$
MBRS [31]	96.48%	95.45%	95.46%	97.08%	83.80%	79.57%	95.43%	97.45%
MBRS_s	98.86%	99.97%	99.93%	99.99%	84.78%	91.57%	96.03%	99.19%
Ours_c	98.24%	98.09%	98.76%	98.78%	84.22%	85.75%	95.60%	98.48%
Ours	**99.93%**	**100.00%**	**100.00%**	**100.00%**	**85.68%**	**95.44%**	**96.75%**	**99.96%**

MBRS_s achieve PSNR values of 35.550dB and 35.748dB, SSIM values of 0.989 and 0.986, respectively, which are all lower than our model; see Table 1. More details about the output image can be seen in Appendix. Then we evaluate the robustness of MBRS, MBRS_s, Ours_c, and our model under various distortions. As shown in Table 5, our model achieves the best accuracy under various distortions, which is 2.48% higher than Ours_c on average. Besides, compared to MBRS, the accuracy of MBRS_s is 3.70% higher on average. These explicit improvements illustrate the better effect of our distortion-style-ensembled noise layer in terms of improving the model's generalization ability. Additionally, compared with MBRS, the accuracy of Ours_c is 2.15% higher on average, showing better robustness of our watermark model backbone with the same noise layer.

5 Conclusion

In this paper, we propose a geometric distortion immunized deep watermarking framework with robustness generalizability, which utilizes Swin Transformer and DCN to build a novel watermark model backbone, effectively solving the stubborn desynchronization problem caused by geometric distortions, showing 100% accuracy under almost all tested cases. Besides, we propose a distortion-style-ensembled noise layer to enrich the expressed distortion features during training, improving our method's generalization ability. Our framework includes two steps. First, the style transfer network is trained to effectively exert distortion style transfer and style fusion. Then, it is converted to the noise layer for training our watermark model, in the end-to-end manner. Extensive experiments demonstrate the effectiveness of our proposed framework, resulting in better visual quality of watermarked images, robustness, and generalizability.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China under Grant 62121002, 62072421, U2336206, 62102386, 62372423 and U20B2047, and by Fundamental Research Funds for the Central Universities under Grant WK2100000041. We would like to sincerely thank all anonymous reviewers for their constructive feedback on our work.

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 126–135 (2017)
2. Ahmadi, M., Norouzi, A., Karimi, N., Samavi, S., Emami, A.: Redmark: Framework for residual diffusion watermarking based on deep networks. Expert Systems with Applications **146**, 113157 (2020)
3. Almohammad, A., Ghinea, G.: Stego image quality and the reliability of psnr. In: 2010 2nd International Conference on Image Processing Theory, Tools and Applications. pp. 215–220. IEEE (2010)
4. Andalibi, M., Chandler, D.M.: Digital image watermarking via adaptive logo texturization. IEEE Transactions on Image Processing **24**(12), 5060–5073 (2015)
5. Bao, P., Ma, X.: Image adaptive watermarking using wavelet domain singular value decomposition. IEEE transactions on circuits and systems for video technology **15**(1), 96–102 (2005)
6. Bi, N., Sun, Q., Huang, D., Yang, Z., Huang, J.: Robust image watermarking based on multiband wavelets and empirical mode decomposition. IEEE Transactions on Image Processing **16**(8), 1956–1966 (2007)
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
8. Chang, C.S., Shen, J.J.: Features classification forest: a novel development that is adaptable to robust blind watermarking techniques. IEEE Transactions on Image Processing **26**(8), 3921–3935 (2017)
9. Chen, D., Yuan, L., Liao, J., Yu, N., Hua, G.: Stylebank: An explicit representation for neural image style transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1897–1906 (2017)
10. Chen, Z., Li, L., Peng, H., Liu, Y., Yang, Y.: A novel digital watermarking based on general non-negative matrix factorization. IEEE Transactions on Multimedia **20**(8), 1973–1986 (2018)
11. Cho, S., Hong, S., Jeon, S., Lee, Y., Sohn, K., Kim, S.: Cats: Cost aggregation transformers for visual correspondence. Advances in Neural Information Processing Systems **34**, 9011–9023 (2021)
12. Cox, I., Miller, M., Bloom, J., Fridrich, J., Kalker, T.: Digital watermarking and steganography. Morgan kaufmann (2007)
13. Craver, S., Memon, N., Yeo, B.L., Yeung, M.M.: On the invertibility of invisible watermarking techniques. In: Proceedings of International Conference on Image Processing. vol. 1, pp. 540–543. IEEE (1997)

14. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
15. Dinh, L., Krueger, D., Bengio, Y.: NICE: non-linear independent components estimation. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings (2015)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021
17. Fang, H., Chen, D., Huang, Q., Zhang, J., Ma, Z., Zhang, W., Yu, N.: Deep template-based watermarking. IEEE Transactions on Circuits and Systems for Video Technology **31**(4), 1436–1451 (2020)
18. Fang, H., Chen, K., Qiu, Y., Liu, J., Xu, K., Fang, C., Zhang, W., Chang, E.C.: Denol: A few-shot-sample-based decoupling noise layer for cross-channel watermarking robustness. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 7345–7353 (2023)
19. Fang, H., Zhang, W., Zhou, H., Cui, H., Yu, N.: Screen-shooting resilient watermarking. IEEE Transactions on Information Forensics and Security **14**(6), 1403–1418 (2018)
20. Fernandez, P., Couairon, G., Jégou, H., Douze, M., Furon, T.: The stable signature: Rooting watermarks in latent diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22466–22477 (2023)
21. Fernandez, P., Sablayrolles, A., Furon, T., Jégou, H., Douze, M.: Watermarking images in self-supervised latent spaces. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3054–3058. IEEE (2022)
22. Gatys, L., Ecker, A., Bethge, M.: A neural algorithm of artistic style (2016). `https://doi.org/10.1167/16.12.326`
23. Hamidi, M., Haziti, M.E., Cherifi, H., Hassouni, M.E.: Hybrid blind robust image watermarking technique based on dft-dct and arnold transform. Multimedia Tools and Applications **77**, 27181–27214 (2018)
24. Han, B., Wang, H., Qiao, D., Xu, J., Yan, T.: Application of zero-watermarking scheme based on swin transformer for securing the metaverse healthcare data. IEEE Journal of Biomedical and Health Informatics (2023)
25. Hong, S., Cho, S., Nam, J., Lin, S., Kim, S.: Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In: European Conference on Computer Vision. pp. 108–126. Springer (2022)
26. Hosam, O.: Attacking image watermarking and steganography-a survey. International Journal of Information Technology and Computer Science **11**(3), 23–37 (2019)
27. Hsu, C.T., Wu, J.L.: Hidden digital watermarks in images. IEEE Transactions on image processing **8**(1), 58–68 (1999)
28. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
29. Huang, Y., Niu, B., Guan, H., Zhang, S.: Enhancing image watermarking with adaptive embedding parameter and psnr guarantee. IEEE Transactions on Multimedia **21**(10), 2447–2460 (2019)

30. Jia, J., Gao, Z., Zhu, D., Min, X., Zhai, G., Yang, X.: Learning invisible markers for hidden codes in offline-to-online photography. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2273–2282 (2022)
31. Jia, Z., Fang, H., 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. pp. 41–49 (2021)
32. Kandi, H., Mishra, D., Gorthi, S.R.S.: Exploring the learning capabilities of convolutional neural networks for robust image watermarking. Computers & Security **65**, 247–268 (2017)
33. Kang, X., Yang, R., Huang, J.: Geometric invariant audio watermarking based on an lcm feature. IEEE Transactions on Multimedia **13**(2), 181–190 (2010)
34. Karybali, I.G., Berberidis, K.: Efficient spatial image watermarking via new perceptual masking and blind detection schemes. IEEE Transactions on Information Forensics and security **1**(2), 256–274 (2006)
35. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 541–551 (1989)
36. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
37. Liu, Y., Guo, M., Zhang, J., Zhu, Y., Xie, X.: A novel two-stage separable deep learning framework for practical blind watermarking. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 1509–1517 (2019)
38. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
39. Lu, Z., He, S., Zhu, X., Zhang, L., Song, Y.Z., Xiang, T.: Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8741–8750 (2021)
40. Ma, R., Guo, M., Hou, Y., Yang, F., Li, Y., Jia, H., Xie, X.: Towards blind watermarking: Combining invertible and non-invertible mechanisms. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 1532–1542 (2022)
41. Mathon, B., Cayre, F., Bas, P., Macq, B.: Optimal transport for secure spread-spectrum watermarking of still images. IEEE Transactions on Image Processing **23**(4), 1694–1705 (2014)
42. Mordvintsev, A., Olah, C., Tyka, M.: Inceptionism: Going deeper into neural networks (2015)
43. Nasir, I., Weng, Y., Jiang, J.: A new robust watermarking scheme for color image in spatial domain. In: 2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System. pp. 942–947. IEEE (2007)
44. Pramila, A., Keskinarkaus, A., Takala, V., Seppänen, T.: Extracting watermarks from printouts captured with wide angles using computational photography. Multimedia Tools and Applications **76**, 16063–16084 (2017)
45. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**, 211–252 (2015)
46. Selim, A., Elgharib, M., Doyle, L.: Painting style transfer for head portraits using convolutional neural networks. ACM Transactions on Graphics (ToG) **35**(4), 1–18 (2016)

47. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
48. Shin, R., Song, D.: Jpeg-resistant adversarial images. In: NIPS 2017 Workshop on Machine Learning and Computer Security. vol. 1, p. 8 (2017)
49. Tancik, M., Mildenhall, B., Ng, R.: Stegastamp: Invisible hyperlinks in physical photographs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2117–2126 (2020)
50. Van Schyndel, R.G., Tirkel, A.Z., Osborne, C.F.: A digital watermark. In: Proceedings of 1st international conference on image processing. vol. 2, pp. 86–90. IEEE (1994)
51. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
52. Wang, H.J.M., Su, P.C., Kuo, C.C.J.: Wavelet-based digital image watermarking. Optics Express **3**(12), 491–496 (1998)
53. Wang, Z., Zhou, M., Liu, B., Li, T.: Deep image steganography using transformer and recursive permutation. Entropy **24**(7), 878 (2022)
54. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
55. Zhang, C., Benz, P., Karjauv, A., Sun, G., Kweon, I.S.: Udh: Universal deep hiding for steganography, watermarking, and light field messaging. Advances in Neural Information Processing Systems **33**, 10223–10234 (2020)
56. Zhang, X., Peng, F., Long, M.: Robust coverless image steganography based on dct and lda topic classification. IEEE Transactions on Multimedia **20**(12), 3223–3238 (2018)
57. Zhu, J., Kaplan, R., Johnson, J., Fei-Fei, L.: Hidden: Hiding data with deep networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 657–672 (2018)