A Robust Deep Learning-Based Video Watermarking Using Mosaic Generation

Souha Mansour[1][a], Saoussen Ben Jabra[2] and Ezzedine Zagrouba[1]
[1]*High Institute of Computer Science, University of Tunis El Manar, 2 Rue Abou Rayhane Bayrouni, Tunis, Tunisia*
[2]*National Engineering School of Sousse, University of Sousse, BP 264 Riadh, Sousse, Tunisia*

Keywords: Deep Learning, CNN, Mosaic Generation, Video Watermarking, Embedding Network, Attack Simulation.

Abstract: Recently, digital watermarking has benefited from the rise of deep learning and machine learning approaches. Even while effective deep learning-based watermarking techniques have been proposed for images, video still introduces extra difficulties, such as motion, temporal consistency, and spatial location. In this paper, a robust and imperceptible deep-learning-based video watermarking method based on CNN architecture and mosaic generation is suggested. The proposed approach is decomposed into two main steps: mosaic generation and signature embedding. This last one includes four stages: pre-processing networks for both the obtained mosaic and the watermark, embedding network, attack simulation, and extraction network. In fact, the main purpose of mosaic generation is to create an image from the original video and to provide robustness against malicious attacks, particularly against collusion attacks. CNN architecture is used to embed signature to maximize invisibility and robustness compromise. The proposed solution outperforms both traditional video watermarking and deep learning video watermarking, according to experimental evaluations on a variety of distortions.

1 INTRODUCTION

Recently, neural networks and especially Convolution Neural Network (CNN)-based deep learning models are utilized extensively in image processing and classification for a wide range of applications and have shown high efficiency for different domains. Watermarking is one of these domains which has profit from the advantages of deep learning. In fact, watermarking is the process of adding a signature into original support, such as images, audio, or video, then extracting the integrated information after applying various distortions to the marked data. Several traditional watermarking techniques have been proposed in the two last decades and depending on the embedding domain that has been selected, these techniques can be roughly divided into two types. The first type is spatial domain-based watermarking where the watermark is directly embedded in the cover data by adding a low amplitude spread spectrum signal or by substituting the pixel values' least significant bits (LSB) (Bayoudh et al., 2018; Bahrami and Akhlaghian Tab, 2018). The second class of watermarking is frequency-based techniques, which entail performing a chosen transformation to the cover data before embedding the watermark. These transforms include Discrete Cosine Transform DCT (Yang et al., 2021; Hou, 2021), Discrete Fourier Transform DFT (Jamal et al., 2016a; Raut and Mune, 2017a), and Discrete Wavelet Transform DWT (Jamal et al., 2016b; Raut and Mune, 2017b). A combination of these transforms can also be used to profit from the advantages of all transforms (Sang et al., 2020) , (Kerbiche et al., 2017). Many factors should be considered when evaluating the performance of watermarking techniques. The three most important are capacity, which refers to how much data can be inserted in the cover media, invisibility, which refers to how easy it is to identify the data, and robustness, which refers to how resistant the data is to attacks. There is an implicit trade-off between these factors. For instance, if the data has a large payload capacity, it will be easier to detect, resulting in a lower level of invisibility. Then as well, increasing resistance to attacks has the potential to reduce both payload capacity and invisibility. Overall, watermarking may be compromised by a malicious attack designed to damage or remove the embedded watermark information, or by a non-malicious attack caused by an unavoidable process used to distribute the content. Traditional watermarking techniques have proved a high

[a] https://orcid.org/0000-0001-9199-2698

Mansour, S., Ben Jabra, S. and Zagrouba, E.
A Robust Deep Learning-Based Video Watermarking Using Mosaic Generation.
DOI: 10.5220/0011691700003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 668-675
ISBN: 978-989-758-634-7; ISSN: 2184-4321

level of invisibility with a robustness against several attacks but their resistance against attacks and security need to be proved. Recently, there has been a greater focus on applying deep learning to watermarking (Byrnes et al., 2021). Due to its capacity to adapt and generalize complex properties, deep learning (LeCun et al., 2015) as a representational learning technique has enabled appreciable advancements in computer vision. Deep learning techniques for data watermarking provide the learning dynamic algorithms to extract high-level and low-level data watermarking properties from massive volumes of data. A variety of deep learning-based algorithms have been investigated for image watermarking, but video watermarking is still in its infancy. Based on our knowledge, only two techniques of deep learning based watermarking were proposed for video contents (Luo et al., 2021).

In this paper, we propose a robust deep learning method for video watermarking that ensures robustness against a variety of attacks, particularly temporal attacks like compression and collusion. In fact, it is based on mosaic generation which transforms the original video to a mosaic image. The signature will then be embedded by applying a CNN based method on the obtained mosaic. The combination of the mosaic image and the CNN architecture allows obtaining high visual quality with a robustness against several usual and malicious attacks.

The rest of this paper is organized as follows: Section 2 provides an overview of deep neural network-based image and video watermarking. Section 3 details the proposed approach and its overall architecture, while Section 4 draws the experimental results and comparison with existing works. Section 5 summarizes the proposed work and shows the future work.

2 RELATED WORKS

Several proposals for watermark embedding based on deep learning into images have been made. These recent deep learning methods for digital watermarking can be classified based on their network architecture design which can be an Auto-encoder CNN, a Convolutions neural network CNN, and Generative adversarial networks GAN. H. Kandi's method (Kandi et al., 2017) is the first deep learning-based method that used two convolutional auto-encoders to generate a watermarked image. Mun et al (Mun et al., 2019) also proposed a watermarking approach with an auto-encoder structured neural network composed of residual blocks.

HiDDeN (Zhu et al., 2018) is a steganography and watermarking scheme proposed by Zhu et al (Zhu et al., 2018) which is the first end-to-end trainable framework model for data hiding that uses an adversarial discriminator. Liu et al (Liu et al., 2019) proposed a novel two-stage separable deep learning (TSDL) framework for blind watermarking that includes noise-free end-to-end adversary training (FEAT) and noise-aware-decoder-only training (ADOT). Ahmadi et al (Ahmadi et al., 2020) proposed a deep diffusion watermarking framework consisting of two fully convolutional neuronal networks with a residual structure that handles embedding and extraction operations.

Zhong et al (Zhong et al., 2020) proposed a CNN-based blind and robust watermarking technique where the main objective is to generalize the watermarking process by training a deep neural network to learn the general rules of watermark embedding and extraction. Lee et al (Lee et al., 2020) proposed a CNN-based digital image watermarking method that does not limit the host image's resolution or watermark information. To the best of our knowledge, the architecture based on convolutional neural networks (CNN) is the most used and applied in various methods. Indeed, the successful results of CNN for image watermarking can be attributed to its high modeling capability and enormous advances in network formation and design. CNN with deep architecture enhances for exploiting image characteristics and improving the training process significantly. Thus, to create an efficient digital watermarking scheme, it should use a framework with CNN architecture including pre-processing networks for watermark data and host data, as proposed in (Lee et al., 2020). Table 1 shows the analytical comparison of the existing deep learning-based image watermarking techniques. Despite the advancement of deep neural network-based image watermarking techniques, video content still has extra challenges, such as temporal coherence. Furthermore, it is difficult to incorporate video into a deep neural network training framework. Besides, visualizing a robust model that uses temporal correlations in a video while preserving temporal coherence and perceptual quality is difficult. Actually, deep learning-based video watermarking is still in its early stages. In fact, Luo et al (Luo et al., 2021) proposed a deep learning-based video watermarking method in 2021. This method includes an encoder, decoder, distortion layer, and video discriminator. The transform layer and the embedding layer are the two main components of the encoder network architecture. As a result, the transform layer's function maps the input video sequence to a feature map with the same di-

Table 1: Characteristics of the recent image watermarking schemes based on deep learning.

Methods	*Techniques*	*Robustness*
Kandi et al (Kandi et al., 2017)	Frequency/ Auto-encoder CNN	common image processing attack
Mun et al (Mun et al., 2019)	Spatial/ Auto-encoder residual block	JPEG attack, Geometric attack, Signal processing attack
Zhu et al (Zhu et al., 2018)	Spatial/ Adversarial network	Dropout,Cropout, Crop,Gaussian, JPEG mask, JPEG drop, Combined
Liu et al (Liu et al., 2019)	Spatial/ Adversarial loss	traditional noise attack and some black box noise attack
Ahmadi et al (Ahmadi et al., 2020)	Frequency DCT/ Circular convolutional layer	common image processing attack
Zhong et al (Zhong et al., 2020)	Spatial/ CNN	common image processing attack
Lee et al (Lee et al., 2020)	Spatial/CNN average pooling	common image processing attack

mensions as the input. It is composed of four layers of three-dimensional convolutions that transform the input video block into a three-dimensional feature block with the same spatial-temporal dimensions as the input video block. The message is first repeated along the spatial-temporal dimensions for the embedding layer and then combined with the transformed video at two scale levels. During the training process, common distortions such as temporal distortions, spatial distortions, and video compression are combined in the distortion layer to teach both the encoder and the decoder to be robust under diverse distortions.

Zhang et al(Zhang et al., 2019) introduce RIVAGAN, a new architecture for video watermarking comprised of two adversaries: a critic and an adversary network The first assesses the quality of the marked video, while the second attempts to remove the watermark. These two components interact with the encoder and decoder networks, which embed and extract the video watermark, respectively. The proposed architecture is based on an attention-based mechanism that identifies regions with high visual quality that are robust for embedding. The attention module is made up of two convolutional layers that are shared by the encoder and decoder. Using the two convolutional blocks creates an attention mask from the original frames. This mask contains the dimensions of data, time, and size. We note that there exist two other works for video watermarking based on deep learning but they are not related to our problematic.

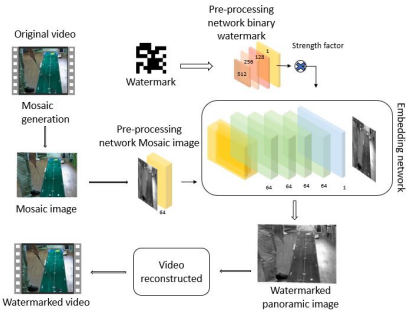

Figure 1: Flowchart of the proposed digital watermarking system.

3 PROPOSED METHOD

In this work, we propose a blind, invisible, and robust video watermarking based on deep learning and mosaic generation. Figure 1 depicts a flowchart of the proposed approach which is composed of four steps: pre-processing network for host data and watermark information, embedding network, attack simulation, and extraction network. The proposed method adapts the resolution of the mosaic image and the watermark information and includes learning methods for training CNN such as average pooling, batch normalization, and Rectifier Linear Unit (ReLU). During the attack simulation, all attacks except collusion and MPEG compression are included in each mini-batch. Furthermore, the content data and watermark information used during training are generated at random.

3.1 Pre-Processing Network

The input of the proposed approach is the mosaic image generated from the original video. The first reason for choosing mosaic image as input is certainly to gain resistance to malicious attacks, particularly collusion and temporal manipulations, which were not considered in previous works. Besides, the second reason is that mosaic generation permits obtaining a 2D image from a given video. Hence, we can apply an efficient image deep learning based technique on the original video. In fact, the mosaic provides a panoramic view of all the information dispersed throughout the sequence and the major benefit of mosaic images is that each repeated point along the video is represented by a single physical point (Ben Jabra and Zagrouba, 2021). Besides, the construction is

performed in three steps: aligning the sequence's images, integrating the images into the panoramic view, and calculating the residual between the mosaic and each image to estimate the mask. The resulting mosaic will represent the background of the scene by removing moving objects or leaving only traces of opacity on these objects. As a result, by combining the final mosaic and the set of transformation parameters associated with each frame, a complete representation of the sequence's evolution can be obtained, but only after including the residual information for each frame. Several video tradition techniques based on mosaic generation have been proposed in the literature (Kerbiche et al., 2018; Bayoudh et al., 2015; Koubaa et al., 2012) and they proved a high robustness against malicious attacks and especially against collusion which resents a dangerous attack, which must be considered for video watermarking. Figure 2 shows some mosaic generation examples, where the first line shows the original videos, and the second one shows the obtained mosaic. Indeed, The host image

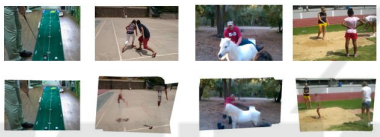

Figure 2: Examples of mosaic generation.

pre-processing network, i.e. the generated mosaic image pre-processing network, keeps the original image resolution and is composed of only one convolutional layer with (3x3) filters to obtain as many features as possible. As well as, the watermark pre-processing network is set to raise the resolution to fit the resolution of the host image pre-processing network. This network includes a convolutional layer (CL), batch normalized (BN), activation function (AF), and average pooling (AP) in order to obtain the desired resolution. After mosaic generation, the pre-processed host image and random watermark outputs are then merged and used as inputs in the watermark embedding network.

3.2 Embedding Network

The embedding network is made up of five blocks: the convolutional layer, batch normalization, activation function ReLU, and the last one consists of the convolutional layer, activation function Tanh to produce the watermarked host image. The watermarked content (WmImg) procures several attacks during training to improve robustness. Algorithm 1 describes the different blocks of the proposed network. With regard to CNN, it is a variety of deep learning (DL) architecture which is a part of machine learning inspired by brain function and structure. These CNN models are now dominant in several computer vision tasks and have achieved incredible results in a wide range of domains, especially in the watermarking domain due to their ability to extract powerful features and represent data with a limited number of parameters (Li et al., 2021).

Data: Vid,Wm,$Mimg$
; /* Vid is the original video */
; /* Wm is the binary watermark */
; /* Mimg is the generated mosaic image */
; /* α is invisibility factor */
Result: $WmImg$: watermarked mosaic image
; /* Processing off network */
$Mimg \leftarrow getmosaic(Vid)$;
$MimG \leftarrow rgb2gray(Mimg)$;
$Nimg \leftarrow normalize(MimG)$; /* Transform from [min, max] into [-1,1] */
; /* Processing in network */
$Pre_M \leftarrow preprocess(Nimg)$; /* Same resolution */
$Pre_W \leftarrow preprocess(Wm)$; /* Increase resolution */
; /* Watermark embedding network */
$Concat \leftarrow Pre_M + \alpha * Pre_W$;
$FeedForward : CNN$; /* input[128x128], filter[3x3], stride [1], No padding */
$Backpropagation$; /* Gradients for all weights, kernels and biais */
$Testing$; /* Calculating error between the Mimg and WmImg */
if $ValAbs(error) \geq tolerance$ **then**
  $Go - Training$;
else
  $End - Training$
end

Algorithm 1: Watermarking embedding.

3.3 Attack Simulation

Simulating such distortions during training is a simple and widely used strategy for resisting noise attacks in robust watermarking. In our approach, for high robustness, we use a dividing strategy that divides the mini-batch evenly into multiple groups, with each group applying a different type of image distortion during the distribution process. This dividing strategy, evidently, applies all of the investigated image distortions in every iteration at the same time, resulting in a significant performance increase. Table 2 shows the types, strengths, and ratios of each attack

chosen and proven in most of the works for testing used in one mini-batch of training. Concerning the collusion attack, which is the main attack we work on, we will discuss it in another section because it is examined outside of the attack simulation step.

Table 2: Attacks used during the training.

Attacks	*Strength*	*Ratio*
Gaussian filtering	3x3,5x5,7x7,9x9	2/12
Average filtering	3x3,5x5	2/12
Median filtering	3x3,5x5	2/12
Salt and pepper	p=0.1	0.5/12
Gaussian noise	sigma=0.1	1/12
JPEG	QF=50	0.5/12
Dropout	0.7	1/12
Rotation	0-90°	1/12
Crop	0.7	1/12

3.4 Extraction Network

The extraction network employs the same technique as the watermark embedding network, but in reverse. The extraction process takes the watermarked video as input and generates a watermarked panoramic image using a mosaic-based approach. It has three convolutional layers, batch normalization, activation function ReLU blocks, and one convolutional layer with Tanh activation function. Finally, it extracts the watermark information as an output.

4 EXPERIMENTAL RESULTS

To assess the proposed approach, we focused on two main criteria : the invisibility and the robustness against usual and malicious attacks. In fact, the proposed method was applied on a dataset and then several various tests are performed and evaluated by qualitative and quantitative measurements. As database, the Kinetics 400 dataset is chosen for our model (Arnab et al., 2021), which is a YouTube action recognition dataset of realistic action videos. Even though our method is based on mosaic images, the first step was the generation of the mosaic datasets by applying the method proposed in (Lee et al., 2020). As a result, we obtain a dataset made up of 1000 mosaics generated from 1000 original videos. In fact, we limited for this work the number of videos at 1000 videos due to the long time taken by the mosaic generation step. The chosen signature is composed of binary images with an 8x8 pixel resolution. Note that a random signature is generated for each training iteration. The proposed video watermarking network is trained in Tensor-Flow on a PC with an Intel(R) Core(TM) i7-7500U CPU @2.70GHz, 64 GB RAM, and an nVidia GeForce RTX 2080 Ti GPU. For gradient descent, we use Adam which is the most widely used optimizer in deep learning because of its efficiency and stability, with a learning rate of 10^{-3} and defaults hyper-parameters. A mini-batch consists of 30 host images, and each mini-batch uses newly generated random-pattern watermark data. The training will be repeated until the loss value becomes stable, which will take 1000 epochs. The strength factor is set to 1 during training, and the weight decay rate is set to 0.01. The difference between the original host mosaic image (Mimg) and the watermarked mosaic image (WmImg) is used for the first loss function which is defined as the mean square error (MSE) descripted in equation 1.

$$L1 = \frac{1}{MN}\sum_{n=i,j}^{MN}[IHMI(i,j) - IWMI(i,j)]^2 \quad (1)$$

Where M x N is the resolution of the host mosaic image.

The difference between the extracted watermark and the original watermark (Wm) is used as the second loss function which is calculating using the mean absolute error (MAE) defined in equation 2

$$L2 = \frac{1}{XY}\sum_{n=i,j}^{MN}|(WM(i,j) - WEx(i,j))| \quad (2)$$

Where X x Y is the resolution of the watermark information.

The loss function of the entire network for training is constructed using the two-loss terms of equations 1 and 2 for host image (Mimg) and watermark, respectively, as equations 3 and 4 for watermark embedding and watermark extraction, respectively.

$$Lemb = \lambda_1 L1 + \lambda_2 L2 \quad (3)$$

$$Lext = \lambda_3 L2 \quad (4)$$

In these two equations, λ_1, λ_2, and λ_3 are the hyper-parameters that control invisibility and robustness, respectively. We assess the invisibility of the proposed approach using qualitative and quantitative evaluations. The qualitative type is verified by the visual appearance of the marked video, while the quantitative type employs both PSNR and SSIM as a quantitatives metrics to demonstrate the invisibility of the embedded signature. The PSNR is defined as

$$PSNR = 10\log 10(255^2/MSE) \quad (5)$$

The SSIM is defined as

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + k_1)(2\sigma_{xy} + k_2)}{(\mu_x^2 + \mu_y^2 + k_1)(\sigma_x^2\sigma_y^2 + k_2)} \quad (6)$$

Where μ_x and μ_y are the means, σ_x and σ_y are the standard deviations,σ_{xy} is the cross-covariance of x and y,

and k_1 and k_2 are two constants used to avoid a null denominator.

Concerning quantitative evaluation, marked videos obtained after embedding the signature in the videos composed the chosen dataset are provided to the laboratory members, and they confirm that the original videos and the marked ones are identical, and it is hard to distinguish between the original video and the marked ones. Figure 4 shows some examples of original and marked frames for two test videos, and it proves that no significant degradation exists between these frames. To quantitatively prove this invisibility, PSNR is measured between original and watermarked video frames. The obtained mean-PSNR values for video tests shown in figure 3 which are detecting during epochs, confirm that the suggested technique guarantees a high visual quality. Certainly, the more we advance epochs, the better we get at invisibility. The robustness of signature is

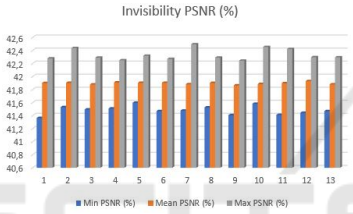

Figure 3: Invisibility percentage.

assessed by performing malicious and non-malicious attacks on watermarked videos and then attempting to extract the embedded mark. In general, watermarking gives priority to robustness over capacity and invisibility. It is a matter of inserting a signature into a video that must be extracted after being exposed to multiple attacks on the watermarked video. We show that by varying the type of image distortion used during training, our model can learn robustness to a wide range of image distortions. Our approach also yields promising results for the MPEG compression and collusion attacks, which are regarded as the most dangerous video attacks. To test the robustness of the proposed approach, various attacks are applied to the watermarked video, and the correlation between the extracted and original signatures is measured. There are two types of attacks tested: classical attacks such

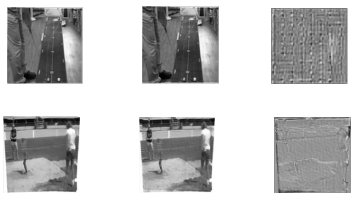

Figure 4: Original, watermarked mosaic image and their differences.

as geometric attacks, filtering, noises, and compression, and temporal attacks such as frame suppression and collusion attacks, which intent to estimate the signature and delete it from the watermarked sequence. Table 3 shows the experimental robustness

Table 3: Robustness results after attacks.

Attacks	***Strength***	***BER***
No attack		0.441576
Gaussian filtering	3x3	4.470109
Average filtering	3x3	11.440217
Median filtering	3x3	4.578804
Salt and pepper noise	0.01	2.792120
Gaussian noise	0.01	0.237772
JPEG	50	3.220109
Rotation	45	50.360054
Cropout	0.5	29.436141
Dropout	0.3	34.932065

results. As demonstrated, the proposed approach is robust against the majority of attacks where the Bit Error Rate (BER) values are less than 10%. In fact, the proposed scheme's robustness against various distortions on the marked image is evaluated by analyzing the distortion tolerance range. As a result, the proposed scheme responses to some difficult image processing attacks are discussed. For these challenges, our approach has a wide tolerance range, particularly for salt and pepper noise, gaussian noise, gaussian and median filtering, and JPEG compression. For example, the extracted watermarks have low average BERs of 2.79%, 0.23%, 8.09%,4.57%, and 3.22% under severe distortions including a cropping 34,93% rotation 50,63% of the marked image. It is worth noting that at 45 degrees, the rotation attack has the biggest effect on image information. As a result, the BER increases as the rotation angle increases but decreases after 45 degrees. It shows that the proposed network was well-trained without being over-fitted to a specific type of strength. The values obtained by the BER can be improved by exploiting more data during the training but since the generation stage of the mosaic images is very time-consuming we are restricted to these values. The simulation attack can be updated to a more general form in future efforts. It is also possible to investigate famous neural network architectures In order to achieve efficiency. To further validate the proposed scheme performance, we tested its robustness against attacks that were not used in training which the correlation value between the extracted and original watermark for collusion measured 0.987 and for MPEG-4 200kbs measured 0.986. For many video watermarking researchers, resistance to collusion attacks has become a critical challenge. Collusion, in fact, is a dangerous and

critical attack, it involves averaging the successive images in order to remove the mark without reducing the quality of the sequence. The proposed approach shows a high robustness against this attack thanks to the use of mosaic image which allow marking every physical point of the video with a same way.

5 COMPARATIVE STUDY

The performance of the proposed scheme is compared to that of state-of-the-art video methods. We chose to compare our approach to a traditional video watermarking based on mosaic generation (Kerbiche et al., 2018) and a deep learning-based video watermarking (Luo et al., 2021). In order to make a fair comparison, we match the PNSR to the comparable methods. Obviously, our method outperforms the DVMark method (Luo et al., 2021) in terms of filtering, salt pepper, geometric attack, Gaussian noise, and collusion attack. In comparison to (Kerbiche et al., 2018) method's which uses mosaic, our method has a significant result against collusion and compression attacks, as well as comparable results against other attacks. Despite the fact that (Luo et al., 2021) has a good compression result due to the training of a small 3D-CNN named CompressionNet, which was used to mimic the output of an H.264 codec at a fixed Constant Rate Factor (CRF), our method represents a good MPEG compression result due to mosaic generation, even when tested without training. We compares our PSNR

Table 4: Comparison of bit accuracy for watermarking methods.

Methods	***DVMark Luo et al***	***Kerbiche et al***	***Our approach***
Gaussian noise	98.56	99.7	98.64
Salt &pepper		99.5	99.77
Crop	87.72	98.4	98.35
Rotation		99.8	96.26
Frame suppression	96.82	99.8	99.82
Compression	83.09	98.6	98.60

to that of (Luo et al., 2021) and (Kerbiche et al., 2018) (ours =42.03, (Luo et al., 2021)=36.5, (Kerbiche et al., 2018)=58.95). The method in (Luo et al., 2021) has a PSNR of 36.5, but our PSNR is 42.03, indicating that our method has a higher PSNR. Indeed, this invisibility is demonstrated using epochs, where an epoch corresponds to learning about the data set, and the higher the number, the better the accuracy. Then, in some attacks, such as a gaussian attack, frame suppression, salt and pepper attack, and collusion attack, our method outperforms others. This demonstrates that the watermark is distributed throughout the image, and by removing some parts, other parts of the image retain the information. Our network has been trained on Gaussian filtering, Gaussian noise addition, cropout, and salt-pepper attacks, but as shown in the table, our method also performs well against collusion and compression attacks. Due to the mosaic generation, our method is resistant to frame suppression; however, the approach (Luo et al., 2021) is weak against this one. In comparison to (Kerbiche et al., 2018), our method was not tested against zoom. Furthermore, other attacks are not similar to or near this attack. As a result, our network is lacking in some areas. To resume, our approach outperforms state-of-the-art methods dependent on the properties of deep learning and mosaic generation ((Luo et al., 2021) and (Kerbiche et al., 2018)). Due to the step of simulation attacks, deep learning outperforms traditional methods in several attacks, and mosaic generation ensures that the method is resistant to collusion attacks. As a result, when compared to state-of-the-art works, the results demonstrated excellent performance and increased efficiency for several attacks. As a matter of fact, our proposal has proven to be both functional and universal.

6 CONCLUSION

This paper proposes a robust video watermarking method based on deep learning and mosaic generation. This method adjusts the resolution of the mosaic image generated from the original video and watermark data. The proposed model is composed of CNN layers. Distortions are simulated in each mini-batch. Furthermore, collusion and video compression attacks are evaluated outside of training and show promising results. This approach outperforms the traditional techniques of video watermarking based on the mosaic generation and deep learning based methods in terms of robustness and invisibility.

We tested the robustness against several attacks and obtained excellent results. The good perceptual quality is then tested qualitatively and quantitatively. The effectiveness of our method stems from the fact that it can easily adapt the video so that it can be watermarked using any deep learning image watermarking scheme. We intend to continue investigating this topic and discover new techniques for extending the deep neural network for video watermarking, such as the recurrent neural network (RNN) and long-term memory (LSTM), as well as work on exploiting the complexity of temporal attacks for video watermark-

ing with deep learning.

REFERENCES

Ahmadi, M., Norouzi, A., Karimi, N., Samavi, S., and Emami, A. (2020). Redmark: Framework for residual diffusion watermarking based on deep networks. *Expert Systems with Applications*, 146:113157.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846.

Bahrami, Z. and Akhlaghian Tab, F. (2018). A new robust video watermarking algorithm based on surf features and block classification. *Multimedia Tools and Applications*, 77(1):327–345.

Bayoudh, I., Ben Jabra, S., and Zagrouba, E. (2018). Online multi-sprites based video watermarking robust to collusion and transcoding attacks for emerging applications. *Multimedia Tools and Applications*, 77(11):14361–14379.

Bayoudh, I., Jabra, S. B., and Zagrouba, E. (2015). On line video watermarking-a new robust approach of video watermarking based on dynamic multi-sprites generation. In *VISAPP (3)*, pages 158–165.

Ben Jabra, S. and Zagrouba, E. (2021). Robust anaglyph 3d video watermarking based on cyan mosaic generation and dct insertion in krawtchouk moments. *The Visual Computer*, pages 1–15.

Byrnes, O., La, W., Wang, H., Ma, C., Xue, M., and Wu, Q. (2021). Data hiding with deep learning: A survey unifying digital watermarking and steganography. *arXiv preprint arXiv:2107.09287*.

Hou, J.-U. (2021). Mpeg and da-ad resilient dct-based video watermarking using adaptive frame selection. *Electronics*, 10(20):2467.

Jamal, S. S., Khan, M. U., and Shah, T. (2016a). A watermarking technique with chaotic fractional s-box transformation. *Wireless Personal Communications*, 90(4):2033–2049.

Jamal, S. S., Khan, M. U., and Shah, T. (2016b). A watermarking technique with chaotic fractional s-box transformation. *Wireless Personal Communications*, 90(4):2033–2049.

Kandi, H., Mishra, D., and Gorthi, S. R. S. (2017). Exploring the learning capabilities of convolutional neural networks for robust image watermarking. *Computers & Security*, 65:247–268.

Kerbiche, A., Jabra, S. B., Zagrouba, E., and Charvillat, V. (2017). Robust video watermarking approach based on crowdsourcing and hybrid insertion. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE.

Kerbiche, A., Jabra, S. B., Zagrouba, E., and Charvillat, V. (2018). A robust video watermarking based on feature regions and crowdsourcing. *Multimedia Tools and Applications*, 77(20):26769–26791.

Koubaa, M., Elarbi, M., Ben Amar, C., and Nicolas, H. (2012). Collusion, mpeg4 compression and frame dropping resistant video watermarking. *Multimedia tools and applications*, 56(2):281–301.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lee, J.-E., Seo, Y.-H., and Kim, D.-W. (2020). Convolutional neural network-based digital image watermarking adaptive to the resolution of image and watermark. *Applied Sciences*, 10(19):6854.

Li, Y., Wang, H., and Barni, M. (2021). A survey of deep neural network watermarking techniques. *Neurocomputing*, 461:171–193.

Liu, Y., Guo, M., Zhang, J., Zhu, Y., and Xie, X. (2019). A novel two-stage separable deep learning framework for practical blind watermarking. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1509–1517.

Luo, X., Li, Y., Chang, H., Liu, C., Milanfar, P., and Yang, F. (2021). Dvmark: A deep multiscale framework for video watermarking. *arXiv preprint arXiv:2104.12734*.

Mun, S.-M., Nam, S.-H., Jang, H., Kim, D., and Lee, H.-K. (2019). Finding robust domain from attacks: A learning framework for blind watermarking. *Neurocomputing*, 337:191–202.

Raut, S. S. and Mune, A. (2017a). A review paper on digital watermarking techniques. *International Journal of Engineering Science*, 10460.

Raut, S. S. and Mune, A. (2017b). A review paper on digital watermarking techniques. *International Journal of Engineering Science*, 10460.

Sang, J., Liu, Q., and Song, C.-L. (2020). Robust video watermarking using a hybrid dct-dwt approach. *Journal of Electronic Science and Technology*, 18(2):100052.

Yang, L., Wang, H., Zhang, Y., Li, J., He, P., and Meng, S. (2021). A robust dct-based video watermarking scheme against recompression and synchronization attacks. In *International Workshop on Digital Watermarking*, pages 149–162. Springer.

Zhang, K. A., Xu, L., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Robust invisible video watermarking with attention. *arXiv preprint arXiv:1909.01285*.

Zhong, X., Huang, P.-C., Mastorakis, S., and Shih, F. Y. (2020). An automated and robust image watermarking scheme based on deep neural networks. *IEEE Transactions on Multimedia*, 23:1951–1961.

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