Blind Deep-Learning-Based Image Watermarking Robust Against Geometric Transformations

Hannes Mareen, Lucas Antchougov, Glenn Van Wallendael, and Peter Lambert
Ghent University – imec, IDLab, Department of Electronics and Information Systems, Ghent, Belgium,
firstname.lastname@ugent.be, https://media.idlab.ugent.be

***Abstract*—Digital watermarking enables protection against copyright infringement of images. Although existing methods embed watermarks imperceptibly and demonstrate robustness against attacks, they typically lack resilience against geometric transformations. Therefore, this paper proposes a new watermarking method that is robust against geometric attacks. The proposed method is based on the existing HiDDeN architecture that uses deep learning for watermark encoding and decoding. We add new noise layers to this architecture, namely for a differentiable JPEG estimation, rotation, rescaling, translation, shearing and mirroring. We demonstrate that our method outperforms the state of the art when it comes to geometric robustness. In conclusion, the proposed method can be used to protect images when viewed on consumers' devices.**

***Index Terms*—Watermarking, Image Forensics.**

I. Introduction

Image watermarking has applications such as copyright protection, authentication, and fingerprinting or traitor tracing [1], [2]. It is typically important for the watermark to be robust. For example, the watermark should survive attacks such as compression and geometric transformations. Additionally, to not hinder the consumers' viewing experience, it is desirable for the watermark to be imperceptible. Moreover, for the detection process to be applicable on a broad scale (such as on consumer devices), it is also important for the watermark to be blind, meaning the watermarking can be detected without requiring the original image nor additional information.

HiDDeN (Hiding Data with Deep Networks) is an advanced, deep-learning-based image watermarking technique [3]. This technique involves the integration of multiple networks into one end-to-end Convolutional Neural Network (CNN). A first network aims to embeds a watermark imperceptibly into the image, and second network aims to accurately decode the embedded watermark. The decoding does not require the original content, making it a blind watermarking technique. Although HiDDeN demonstrated good performance, it is not robust against geometric transformations.

This paper proposes a novel watermarking method robust against geometric attacks, by extending the HiDDeN architecture [3] with noise layers that simulate geometric attacks.

This work was funded in part by the Research Foundation – Flanders (FWO), IDLab (Ghent University – imec), Flanders Innovation & Entrepreneurship (VLAIO), the Flemish Government, and the European Union. The computational resources (STEVIN Supercomputer Infrastructure) and services used in this work were kindly provided by Ghent University, the Flemish Supercomputer Center (VSC), the Hercules Foundation and the Flemish Government department EWI.

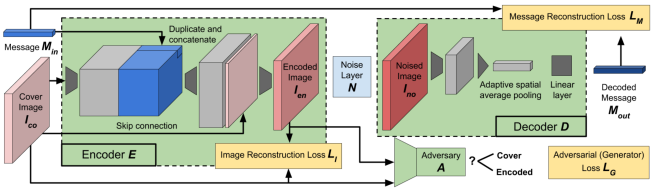

Fig. 1. Architecture of HiDDeN [3], consisting of an encoder and decoder, separated by noise layers. We include noise layers that simulate geometric transformations to provide robustness against geometric attacks.

II. Proposed Method

We propose to extend the HiDDeN architecture [3] in order to make it robust against gemoetric distortions. The HiDDeN architecture is visualized in Fig. 1, and consists of three networks to achieve this task: an encoder, decoder, and adversary. The encoder network takes an image and a bit message as input and produces an output image with the message embedded into it. The decoder network then takes the encoded image as input and returns a decoded message. The decoder is blind, i.e., it does not take any additional information as input. The adversary network plays a crucial role by being trained to differentiate between watermarked and non-watermarked images, essentially acting as a discriminator. This dynamic enables the encoder and decoder networks to learn how to "outsmart" the adversary network, resulting in a substantial improvement in imperceptibility for the embedded watermark. Using the adversary network results in a substantial improvement in imperceptibility [3].

Between the encoder and the decoder networks, multiple so-called noise layers can be inserted. These layers simulate attacks on the watermark. The originally proposed HiDDeN model [3] used the following noise layers: Crop, Gaussian Blur, Cropout, Dropout, and JPEG compression through JPEG-Mask & JPEG-Drop. We opted to not use Cropout and Dropout, as these use information from the unwatermarked image and are therefore not realistic attacks. Additionally, we replaced the JPEG-Mask and JPEG-Drop layers because we could not achieve good robustness against JPEG compression with them. Instead, we propose to use a JPEG noise layer on a differentiable approximation of JPEG compression called $\text{JPEG}_{\text{diff}}$ [4] (instead of the JPEG-Mask and JPEG-Crop noise layers that the original method used). $\text{JPEG}_{\text{diff}}$ works in a similar way as JPEG, with the difference that the rounding operation after quantization uses a rounding approximation instead. That is because the rounding operation $\lfloor I \rceil$ has a derivative of 0 nearly everywhere. The differentiable approximate rounding

TABLE I
IMPERCEPTIBILITY RESULTS

Identity	Rescale	Translate	Rotate	Shear
0.975	0.960	0.977	0.963	0.966
Mirror	**Gaussian Blur**	**$\text{JPEG}_{\text{diff}}$**	**Combined**	**RivaGAN**
0.980	0.963	0.963	0.930	0.977

is done in the following way: $\lfloor I \rceil_{approx} = \lfloor I \rceil + (I - \lfloor I \rceil)^3$. This network is trained separately to simulate JPEG and is then used as a noise layer.

To provide robustness against geometric transformations, we add the following noise layers: Rescale (factor between 50% and 200%), Translate (5% to 50% of the image width and height), Rotate (angles between +/- 10 and 60 degrees), Shear (angles between +-/ 10 and 45 degrees), and Mirror. These operations are all differentiable. Additionally, we deleted the Crop noise layer, as we found that the base model (without noise layers) already provided sufficient robustness against cropping (see Section III and Fig. 2).

III. EVALUATION

A. Experimental Setup

To create the training, validation, and testing, we selected 10000, 1000, and 1000 images from the COCO dataset [5], respectively. During training, random crops of 128×128 pixels are taken, whereas testing was done on rescaled images of 512×512 pixels (to evaluate imperceptibility) and 256×256 pixels (to evaluate robustness). Additionally, we embedded randomly-generated 30-bit messages. We train and evaluate an Identity model (no noise layers), 7 specialized models (single noise layer), and a Combined model (6 noise layers, selected at random during training). The Shear noise layer was not added in the Combined model, because the Rotate layer already sufficient provided robustness against shearing. We additionally evaluate and compare against the state-of-the-art RivaGAN method [6].

B. Imperceptibility

To measure the imperceptibility, we calculate the structured similarity index measure (SSIM) between the original image and its watermarked version. A higher value signifies a less perceptible watermark. Table I gives average SSIM values for all methods. Although the combined models scores the lowest SSIM score (0.930), it remains an imperceptible result. Furthermore, RivaGAN exhibits an SSIM score of 0.977, performing similarly to most specialized models. We manually inspected the watermarked images and confirm that it is very hard (if not impossible) to spot artefacts at native resolution. Example images are available on our website: https://media.idlab.ugent.be/watermarking-blind-icce.

C. Robustness

To measure the robustness, we evaluate the bit accuracy of the decoded messages for a range of attacks. It can be observed that the identity model (blue line) performs relatively poorly against most attacks. The surprising exception is cropping, for

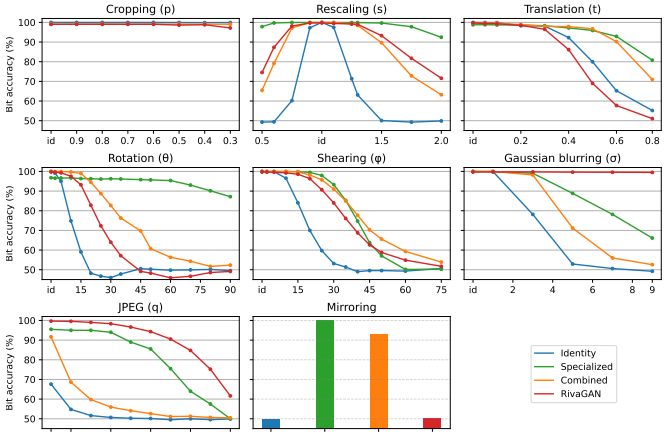

Fig. 2. Robustness results.

which it maintains a bit accuracy close to 100% for crops with ratio $p \geq 0.2$. The specialized models (green lines) never fall below 65% bit accuracy for the considered attacks. The combined model shows significantly improved robustness compared to the identity model. Unsurprisingly, the combined model has a lower resilience than the specialized models for their targeted attacks, though. This is especially the case for JPEG compression and Gaussian Blurring. Future work should investigate how to provide additional robustness in these cases.

In contrast, RivaGAN shows very good performance against JPEG compression and Gaussian blurring. This demonstrates RivaGAN's ability to extract low-level features of images that can survive certain attacks without an explicit need for noise layers [6]. However, most notably, our specialized and combined methods outperform RivaGAN when exposed to geometric attacks.

IV. CONCLUSION

We presented a blind deep-learning-based watermarking method based on the HiDDeN architecture [3], which uses noise layers to provide robustness against specific attacks. We demonstrate that the embedded watermarks are imperceptible and, most notably, we outperform state-of-the-art methods in robustness against geometric attacks. As such, the watermarking method can be used to protect copyright of images viewed on consumer electronic devices.

REFERENCES

[1] M. Asikuzzaman and M. R. Pickering, "An overview of digital video watermarking," *Proc. IEEE Int. Symposium Circuits Syst. (ISCAS)*, vol. 28, no. 9, pp. 2131–2153, 2018.

[2] H. Mareen, J. De Praeter, G. Van Wallendael, and P. Lambert, "A novel video watermarking approach based on implicit distortions," *IEEE Trans. Consum. Electron.*, vol. 64, no. 3, pp. 250–258, 2018.

[3] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "HiDDeN: Hiding data with deep networks," *CoRR*, vol. abs/1807.09937, 2018.

[4] R. Shin and D. Song, "JPEG-resistant adversarial images," in *NIPS Workshop Mach. Learning Computer Security,*, vol. 1, 2017, p. 8.

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Computer Vis.* Springer, 2014, pp. 740–755.

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