Watermark Warriors: Comparative Study of Machine Learning Models for Watermark Removal Leveraging GAN, U-Net, and DnCNN Techniques

Ziwei Guo [1] **Jack K. Garritano** [1] **Noah Z. Ong** [1] **Marcus G. Rim** [1] **Nino Zhou** [1]

Abstract

This paper presents a systematic comparison of machine learning approaches for removing watermarks from images. We evaluate Generative Adversarial Networks (GANs), Convolutional Neural Networks (CNNs), and U-Net architectures applied to image processing tasks. A new dataset is generated for training and evaluation to explore models' performances on random generated string watermarks. We present a quantitative and qualitative analysis of model performance based on accuracy, computational cost, and scalability. The results provide insights into the efficacy of different ML algorithms for watermark removal from images.

1. Introduction

Watermarks are widely used for copyright protection, but can be vulnerable to attacks based on modern machine learning techniques. To enforce copyright in the modern world, new protection techniques may become necessary if traditional watermarking is ineffectual. This study explores advanced machine learning techniques for the automated removal of watermarks from images while preserving image quality. We compare GANs, CNNs, and U-Nets repurposed for visual tasks.

If successful, this study will provide insight into the capabilities of automated watermark removal, informing on which techniques stand up to modern methods. By comparing model strengths, it will inform applications such as digital archiving, copyright management, and multimedia editing.

Currently, watermark removal relies heavily on heuristic-based image processing methods or simple neural network architectures like CNNs. These approaches often fail to preserve fine details, leaving visible artifacts or degrading the overall image quality. Manual watermark removal, while precise, is time-consuming, historically making watermarks a strong copyright mechanism. This paper aims to explore the impacts of modern machine learning approaches on the watermarking as a copyright mechanism.

2. Dataset Creation

2.1. Comprehensive Dataset Development

We develop a novel dataset consisting of consisting of 10,000 watermarked and non-watermarked images with diverse styles, complexities, and opacity levels. This ensures that the trained models can generalize across a wide variety of watermark patterns, ranging from simple text overlays to semi-transparent logos embedded in intricate backgrounds.

Key features include:

- **Source Images:** Images were randomly queried from open-license repositories such as Unsplash and Pexels. Query strings were varied randomly to retrieve images with diverse of subject matter, with the goal of helping the model better generalize. The image retrieval process ensures a variety of styles, resolutions, and content types.
- **Watermark Design:** Watermarks were synthetically generated using text, and semi-transparent patterns. Size, opacity, rotation, and complexity were randomly adjusted to produce images representing a wide range of potential watermarking schemes.
- **Training/Testing Split:** The dataset is divided into 80% training, 10% validation, and 10% testing sets.
- **Augmentation:** Techniques like rotation, scaling, and color jittering were applied to enhance robustness. Watermark patterns were randomized and included grid patterns and centralized watermarks.

3. Methodology

We employed several machine learning approaches to train and evaluate watermark removal models. Below is a summary of our methods and approaches:

3.1. Quantitative Hybrid Metric-Based Training

We combine PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) metrics during training to ensure that the output maintains both pixel fidelity and perceptual quality. This dual-objective training strategy allows

for more balanced optimization, addressing both low-level pixel differences and high-level structural similarity.

3.2. Qualitative Comparison of GAN, U-Net, and DnCNN Methods

A qualitative comparison of GAN, U-Net, and DnCNN architectures aims to explore their respective strengths and weaknesses for watermark removal tasks. GANs are evaluated for their ability to generate visually realistic results, especially in complex image scenarios. U-Net is assessed for its spatial precision and structural fidelity, while DnCNN is analyzed for its robustness to noise and generalization capabilities.

3.3. Model Type: Generative Adversarial Networks (GANs)

GANs leverage a generator and discriminator to produce realistic watermark-free images:

- **Architecture:** A custom GAN architecture was implemented, where the generator is based on a convolutional-deconvolutional structure to produce high-resolution images, and the discriminator uses a PatchGAN-like structure to evaluate real versus fake image pairs.
- **Loss Function:** The adversarial loss ensures realistic outputs, while a pixel-wise MSE loss encourages similarity between generated and ground truth images.
- **Training Details:** The GAN was trained for 25 epochs using 5 percent of the dataset with a batch size of 16, a learning rate of 0.0002, and the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$.

3.4. Model Type: Convolutional Neural Networks (CNNs)

CNNs were trained using supervised learning to reconstruct images without watermarks. Specifically, UNet models and DnCNN architectures were tested.

DnCNN is a CNN designed for image denoising. It predicts noise residuals instead of directly reconstructing clean images, making the optimization process more efficient. We test the extent to which a watermark can be detected as noise. We also test various CNN models' performances in image watermark removal tasks.

The U-Net architecture is a powerful and versatile deep learning model, specifically designed for pixel-level tasks such as image segmentation and image-to-image translation. As shown in recent work (Fu et al., 2022), watermark removal using U-Net achieves promising results. It consists of an encoder-decoder structure, where the encoder extracts high-level semantic features through down-sampling, and the decoder reconstructs the original resolution using upsampling. To preserve fine-grained spatial details during reconstruction, skip connections are incorporated between corresponding layers of the encoder and decoder. This combination allows the U-Net to efficiently learn both global and local features, making it highly effective for tasks like watermark removal.

So we use three different CNNs here to test its performance in watermark removal:

- **Model1:** A classic UNet.
- **Model2:** A modified UNet which uses fixed-width encoding channels, learned upsampling, and additional skip connections, making it more lightweight and suited for image reconstruction tasks..
- **Model3:** A 3 channel, 17 layer DnCNN.

3.4.1. Loss Function

To optimize the performance of the U-Net for watermark removal, a combined loss function was employed, integrating Mean Squared Error (MSE) and Structural Similarity Index Measure (SSIM). The MSE component minimizes pixel-wise differences between the predicted and ground truth images, ensuring numerical accuracy. The SSIM component, on the other hand, emphasizes the perceptual quality of the reconstructed images by focusing on structural and visual similarities. The combined loss function is formulated as:

$$\text{Loss} = \alpha \cdot \text{MSE} + \beta \cdot (1 - \text{SSIM}),$$

where $\alpha = 0.5$ and $\beta = 1.5$. This weighting ensures that the model balances pixel-level precision with perceptual quality, leading to outputs that are both accurate and visually coherent.

3.4.2. Key Strengths of the Approach

The combination of the U-Net architecture and our new loss function provides a robust approach to watermark removal. The encoder-decoder structure, augmented by skip connections, enables the model to retain spatial details while reconstructing watermark-free images. By incorporating SSIM into the loss function, the approach prioritizes perceptual similarity, ensuring that the reconstructed images are visually appealing and free from artifacts. Additionally, techniques such as learning rate scheduling (using ReduceLROnPlateau) and output clamping ensure stable training and prevent invalid pixel values. These design choices collectively contribute to a model that is effective, reliable, and capable of generalizing well to unseen data.

3.5. Deep Image Prior

The model for this section is based on the paper: (Ulyanov et al., 2020)

Deep convolutional neural networks (CNNs) have revolutionized image generation and restoration by producing impressive results in tasks like denoising, super-resolution, and inpainting. Traditionally, these successes are attributed to the networks' ability to learn complex image features from large datasets. However, recent findings reveal that even without any training, the inherent structure of a CNN can capture significant low-level image statistics.

The Deep Image Prior (DIP) model leverages this insight by using an untrained, randomly initialized neural network as a handcrafted image prior. Instead of learning from external data, the network relies solely on the corrupted image itself to guide the restoration process.

Key Concepts:

- Optimization Over Network Parameters: Rather than searching for the restored image directly in the image space, DIP searches within the parameter space of a neural network. The network generates an image from a fixed random input (often just noise), and the goal is to adjust the network's parameters so that its output matches the corrupted image where it is known to be correct.
- No External Training Data: The DIP model does not require any pre-training on large image datasets. The network starts with random weights and uses the corrupted image as the only source of information during the optimization.
- Implicit Image Prior: The architecture of the CNN itself acts as a prior. Its built-in inductive biases favor natural image structures, such as smooth regions and repetitive patterns, which are common in real-world images. This means the network is more likely to produce plausible, natural-looking images even without explicit training.

How It Works:

- Initialize the Network: Start with a deep CNN with random weights. The network takes a fixed random noise image as input.
- Define the Loss Function: The loss measures the difference between the network's output and the corrupted image, but only in the regions where the image is not corrupted. In other words, it compares the known good pixels in the corrupted image to the corresponding pixels in the network's output.
- Optimize the Network Parameters: Adjust the network's weights to minimize the loss. Since the network is untrained, it gradually learns to reproduce the known parts of the image.
- Infer Missing or Corrupted Regions: As the network fits the known data, its inherent structure helps it to fill in the missing or corrupted regions with plausible content, effectively restoring the image.

4. Results and Discussion

The overall performances of all methods is as follows:

Table 1. Comparison of Models on Average PSNR and SSIM

Model	Average PSNR (dB)	Average SSIM
U-Net	12.8880	0.5123
GAN	17.28	0.2733
Modified U-Net	20.7129	0.6062
DnCNN	30.5602	0.9623
Deep Prior	36.8028	0.9480

4.1. GAN Performance

The PSNR and SSIM scores are relatively low compared to the other methods that we tried. The model achieved an average PSNR of 17.28 dB and SSIM of 0.2733. The model generally seems to make all the images a little bit fuzzy, but it appears that the watermarks are also being blurred out. The model seemed to struggle when watermarks were large in size as a whole, or had a large font size, while performing better on smaller watermarks. Similarly to the U-Net model, the results seem to show whitening and over-smoothing.

4.2. U-Net Performance

We evaluated the U-Net model on the test set using two primary metrics: **Peak Signal-to-Noise Ratio (PSNR)** and **Structural Similarity Index (SSIM)**, which quantify reconstruction quality and perceptual similarity, respectively. The model achieved an average PSNR of 12.89 dB and SSIM of 0.5123, performing well on simpler watermarks or uniform backgrounds but struggling with heavily textured regions where the metrics were significantly lower. Visual inspection revealed whitening and over-smoothing effects in complex areas, likely due to pixel-based loss functions (e.g., MSE) that prioritize global consistency over fine detail restoration, as well as the U-Net architecture's limited ability to capture high-frequency features. Additionally, the diversity of the training dataset may have constrained the model's ability to generalize to more intricate scenarios. These observations suggest that while the U-Net provides a solid foundation for watermark removal, improvements such as more diverse training data, perceptual loss functions, and

enhanced architectures are necessary for better robustness and fidelity.

4.3. Modified U-Net Performance

4.3.1. Performance Analysis

The score this model can get is not so good, but it seems good in specific images, it can get PSNR values over 25dB in some pictures, and the model reduces its watermark but reduce its resolution at the same time, which causes not showing good in both PSRN and SSIM values. But if we can fix the problem that makes the picture blurry, it's still a potential method, and that's why we add SSIM as a part of our loss function.

4.4. DnCNN Performance

4.4.1. Performance Analysis

We can see The combination of the DnCNN architecture with our new loss function can get a relatively high score: average PSNR: 30.5602, average SSIM: 0.9623, but seems not good in the output images, the output image is actually the same as the watermarked images.

The reason we think is that the watermark we used in this project is random and not so influential to the origin image, so it's not so harmful to PSNR score and SSIM score, but the model can't figure this noise out, So the model's output is not what we want.

In conclusion, DnCNN doesn't show good performance on this kind of watermarks.

4.5. Deep Prior Performance

The deep prior image processing system shows excellent performance. With an average PSNR of 36.80 dB, the images are of high quality with minimal distortion. The average SSIM is 0.948, indicating that the images retain their structural similarity to the originals very well. However, it is important to note that the use of a mask in the deep prior model means that its task removing the watermark is much easier, as the model is initially given the general location of the watermark. But overall, the system effectively preserves image quality and detail.

5. Comparison of Model Performance

The performance comparison across models highlights significant variations in their capabilities for watermark removal. The U-Net achieved a moderate average PSNR of 12.8880 dB and SSIM of 0.5123, demonstrating its effectiveness in preserving structural fidelity but limited capacity for high-fidelity reconstruction in complex scenarios. The GAN model, with an average PSNR of 17.28 dB and a lower SSIM of 0.2733, showed better numerical reconstruction quality but struggled with perceptual consistency, likely due to the adversarial training's emphasis on realism over pixel accuracy. The modified U-Net significantly outperformed the original U-Net, achieving an average PSNR of 20.7129 dB and SSIM of 0.6062, suggesting that architectural enhancements improved both fidelity and perceptual quality. Finally, the DnCNN model demonstrated the best performance with an average PSNR of 30.5602 dB and SSIM of 0.9623, reflecting its superior ability to handle noise and generalize across diverse images. These differences may be attributed to the distinct training strategies, architectural designs, and loss functions employed by each model, with the DnCNN benefiting from its specialized focus on noise reduction and fine-detail reconstruction.

6. Conclusion and Future Work

This study conducted a comparative evaluation of machine learning models for watermark removal, analyzing their performance across key metrics. While U-Net and DnCNN models exhibited strong results in terms of reconstruction quality and perceptual similarity, GANs proved effective in handling complex watermark patterns despite limitations in numerical accuracy. The modified U-Net further demonstrated the potential of architectural enhancements to balance fidelity and efficiency. Future research will aim to enhance model robustness, optimize computational performance, and investigate advanced approaches such as few-shot learning to address novel and diverse watermark designs in real-world scenarios.

References

Fu, L., Shi, B., Sun, L., Zeng, J., Chen, D., Zhao, H., and Tian, C. An improved u-net for watermark removal. *Electronics*, 11(22):3760, 2022. doi: 10.3390/electronics11223760.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. Deep image prior. *International Journal of Computer Vision*, 128 (7):1867–1888, March 2020. ISSN 1573-1405. doi: 10.1007/s11263-020-01303-4. URL http://dx.doi.org/10.1007/s11263-020-01303-4.