**CSE 572: Data Mining**

**Final Project Literature Review**

**Project Title:** Analysis of Effective Methods for Watermark Detection and Removal

**Team members:**

| Full name | ASU ID |
| --- | --- |
| Meesum Ali Khan | 1225453632 |
| Abichal Ghosh | 1225427294 |
| Zuy Pham | 1216711899 |
| Nikhil Karthik Bindem | 1222520364 |

**Step 1: Summary of relevant work**

To identify relevant work, you should search key words related to your chosen topic in search engines such as Google Scholar. Then you will select the most relevant papers to your proposed project. The number of relevant papers will vary for each project, but most projects will probably find 8-10 key relevant papers.

Once you’ve selected your list of papers, at least one member of your group should read each paper and make notes about the key points. To help you organize your notes about each paper, fill out the following template for each of the key papers (thus you will have around 8-10 of these blocks below, though the exact number of papers will depend on what is relevant for your project).

Citation in [ACM citation style](https://www.acm.org/publications/authors/reference-formatting).

Brief summary:

* 1-3 bullets that concisely summarize the key innovation and results in the paper

Strengths:

* 1-3 bullets that concisely summarize the key strengths of the paper

Limitations:

* 1-3 bullets that concisely summarize the key limitations of the paper

Here is an example:

Rußwurm, M., Courty, N., Emonet, R., Lefèvre, S., Tuia, D., & Tavenard, R. (2023). End-to-end learned early classification of time series for in-season crop type mapping. *ISPRS Journal of Photogrammetry and Remote Sensing, 196*, 445-456.

Brief summary:

* Proposed loss function that optimizes dual objective of classification accuracy and earliness of classification
* Model outputs crop type class prediction in addition to probability that the prediction should be used at that timestep or wait for more data from later timesteps
* Demonstrated using LSTM with Sentinel-2 time series, but can be implemented for any deep learning model

Strengths:

* Simple approach that would be easy to implement for any neural network architecture
* Provides information that can be used to judge reliability of predictions at a given time in the growing season (which can be used to inform end-user decision-making)

Limitations:

* Poor performance for minority classes (subject to class imbalance issues)
* Poor performance for small datasets (as with many deep learning models)
* Outperformed by random forest baseline

1. Xinyun Chen, Wenxiao Wang, Chris Bender, Yiming Ding, Ruoxi Jia, Bo Li, and Dawn Song. 2021. REFIT: A Unified Watermark Removal Framework For Deep Learning Systems With Limited Data. In Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security (Virtual Event, Hong Kong) (ASIA CCS ’21). Association for Computing Machinery, New York, NY, USA, 321–335. <https://doi.org/10.1145/3433210.3453079>

Brief summary:

* This paper proposes a general-purpose watermark removal framework based on fine-tuning called REFIT, which is effective against a wide range of watermarking schemes
* This highlights the importance of investigating more robust watermark embedding schemes against attacks and proposes two techniques for removing watermarks under a weak threat model, an adaptation of elastic weight consolidation (EWC) algorithm, and unlabeled data augmentation (AU).

Strengths:

* A realistic attack scenario is evaluated, where the adversary has limited training data.
* The framework proposed in the study uses two techniques that can remove watermarks without affecting the model's functionality under this weak threat model.

Limitations:

* The study does not analyze how the model's accuracy or performance is affected by the watermark removal attacks.
* The paper considers a weak threat model where the adversary has limited access to data, and it only investigates image classification models, not other types of models or applications.

1. Danni Cheng, Xiang Li, Wei-Hong Li, Chan Lu, Fake Li, Hua Zhao, and Wei-Shi Zheng. 2018. Large-Scale Visible Watermark Detection and Removal with Deep Convolutional Networks. In Pattern Recognition and Computer Vision, Jian-Huang Lai, Cheng-Lin Liu, Xilin Chen, Jie Zhou, Tieniu Tan, Nanning Zheng, and Hongbin Zha (Eds.). Springer International Publishing, Cham, 27–40. <https://doi.org/10.1007/978-3-030-03338-5_3>  
   Brief Summary:

* Watermark region is detected using a deep learning framework which evaluates many regions on the image and outputs the most likely location
* Using an Image Transformation Deep Learning Implementation, the area with watermark is transformed, and the watermark is removed

Strengths:

* By splitting the approach into two steps with two dedicated models, we see more reliability from this method
* This approach can be applied to watermarks of all sizes, while some others may struggle to work with smaller ones

Limitations:

* Dataset needs to have a variety of watermarks, or the model will overfit on one type and not be able to detect others

1. Tali Dekel, Michael Rubinstein, Ce Liu, and William T. Freeman. 2017. On the Effectiveness of Visible Watermarks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 6864–6872. <https://doi.org/10.1109/CVPR.2017.726>

Brief summary:

* This work provides a comprehensive study regarding the effectiveness of visible watermarks by exploring the different types of majorly used watermarks and trying to remove these.
* The study uses an algorithm based on image processing techniques for the detection and removal of watermarks and their alpha matte based on the consistency and uniformity of watermarks used.
* The results of the work show the effectiveness of certain techniques utilized for incorporating watermarks such as opacity, differing size/location, etc.

Strengths:

* This study uses multiple image datasets based on different techniques for creating watermarks to assess these watermarking methods on a large scale. This approach helps to make more accurate comparisons and generalizations.
* Multiple methods were tested which helped in providing practical recommendations for more robust watermarking techniques, such as using opacity variation across the image, adding translation, adding geometric perturbations, using multiple watermarks in different positions and more.

Limitations:

* The method used here exclusively depends on image processing techniques and does not use any Deep Learning methods. If the watermark used is not transparent, then this method will fail.
* The method used in this study is heavily reliant on the consistency and uniformity of watermarks across the image dataset; i.e. given a set of images it assumes the use of a single watermark and alpha matte for all the images.
* For any unseen image/watermark, it requires the user to provide a rough box around the watermark.

1. Yongjian Hu, S. Kwong, and Jiwu Huang. 2005. An algorithm for removable visible watermarking. *IEEE Transactions on Circuits and Systems for Video Technology* 16, 1 (December 2005), 129–133. <http://dx.doi.org/10.1109/tcsvt.2005.858742>

Brief summary:

* Presented a new algorithm for removable visible watermarking that uses a user-key structure for embedding watermarks that adapt to the host image’s features
* With correct user keys, watermarks can be removed, and thus restore the high quality unmarked image
* Unauthorized removal of watermarks will result in a low-quality image due to wrong estimation of embedding parameters

Strengths:

* The algorithm can add a visible watermark to an image, which can be removed without affecting the original image
* The algorithm generates the watermark based on the local features of the host image, making the watermark harder to attack

Limitations:

* The algorithm is not suitable for invisible watermarks
* The algorithm’s performance may vary depending on the characteristics of the host image

1. M.S. Kankanhalli, Rajmohan, and K.R. Ramakrishnan. 1999. Adaptive visible watermarking of images. In Proceedings IEEE International Conference on Multimedia Computing and Systems, Vol. 1. 568–573 vol.1. <https://doi.org/10.1109/MMCS.1999.779263>

Brief summary:

* The paper discusses the need for copyright protection for high-quality images and video, and proposes a visible watermarking technique that varies the location and strength of the watermark according to the underlying content of the image.
* The proposed method analyzes the texture, edge, and luminance information of each block of pixels to classify them into one of eight classes and embeds the watermark in the DCT transform domain.
* The paper highlights the advantages of visible watermarks and discusses the desired characteristics of visible watermarks.

Strengths:

* The proposed method varies the intensity of the watermark in different regions of the image depending on the underlying content of the image, ensuring that it is perceptually uniform over different regions of the image and the process can be automated over a wide range of images.
* The watermarking technique exploits redundancies in the coded image to embed the information bits, and the insertion process can be automated for all kinds of images.
* Provides an immediate claim of ownership and serves as a deterrent against unauthorized use of copyrighted high-quality images serves the dual purpose of providing a recognizable identity to the content.

Limitations:

* This method has a narrowed scope and uses limited watermarking techniques
* The desired characteristics of visible watermarks, such as being visible only on careful examination of the image and not significantly obscuring the image details beneath it, can be conflicting, which makes the creation of a robust and perceptually uniform watermark a hard and interesting problem.

1. Xiang Li, Chan Lu, Danni Cheng, Wei-Hong Li, Mei Cao, Bo Liu, Jiechao Ma, and Wei-Shi Zheng. 2019. Towards Photo-Realistic Visible Watermark Removal with Conditional Generative Adversarial Networks. In Image and Graphics, Yao Zhao, Nick Barnes, Baoquan Chen, Rüdiger Westermann, Xiangwei Kong, and Chunyu Lin (Eds.). Springer International Publishing, Cham, 345–356. <https://doi.org/10.48550/arXiv.1905.12845>

Brief Summary:

* Proposed Conditional GAN model to remove watermarks from images and give a photorealistic reconstruction.
* Incorporated a patch-based discriminator trained on reconstructed images and original images to improve reconstruction by reducing traces of watermarks left by the original model.

Strengths:

* This is the first novel CGAN approach to removing watermarks from images, and shows very promising results.
* A new loss function is introduced, which represents the earlier described discriminator through its adversarial loss, and also considers pixel-wise content loss. This leads to more photorealistic results.

Limitations:

* Difficult to train the discriminator to identify the differences between reconstructed images and original images, as it may overfit the model that is giving the reconstructed images

1. Jing Liang, Li Niu, Fengjun Guo, Teng Long, and Liqing Zhang. 2021. Visible Watermark Removal via Self-calibrated Localization and Background Refinement. CoRR abs/2108.03581 (2021). arXiv:2108.03581 <https://doi.org/10.48550/arXiv.2108.03581>

Brief summary:

* The work done in this paper includes a deep learning framework based on an encoder-decoder (generator-discriminator) structure for the task of visible watermark removal
* The authors also conduct an ablation study on the model which emphasizes the significance of several elements of the model.
* The proposed method is able to perform the task of watermark removal with performance and is a significant improvement over pre-existing methods.

Strengths:

* The approach outperforms existing state-of-the-art methods, including W2Net, BVMR, Split-Net, U-Net, and more.
* This model performs well even on the cases in which the image has been compressed or resized.

Limitations:

* The dataset used uses watermarks which are very large in size. This is not the case in images found online, as watermarks are usually much much smaller and occupy much smaller spaces of the image.
* The watermarks themselves are simply added with a reduced opacity. No other forms of watermarking such as geometric perturbation, varying opacity, multiple opacity over the image has been used.

1. Hector Santoyo-Garcia, Eduardo Fragoso-Navarro, Rogelio Reyes-Reyes, Gabriel Sanchez-Perez, Mariko Nakano-Miyatake, and Hector Perez-Meana. 2017. An automatic visible watermark detection method using total variation. *2017 5th International Workshop on Biometrics and Forensics (IWBF)* (April 2017), 1–5. <http://dx.doi.org/10.1109/iwbf.2017.7935109>

Brief summary:

* Proposed an automatic visible watermark detection method using total variation
* The visible watermarked image is decomposed into a structure image and texture image by applying the total variation method based on L1 norm
* The structure image contains the watermark pattern, and the texture image contains the texture data of the host image

Strengths:

* Total variation filtering improves the visibility of watermarks, making it easier to detect
* Experimental results show that the proposed method performs well with detecting watermarks

Limitations:

* The proposed method is not applicable for invisible watermarks
* The proposed method may not work well on images that are low quality or highly-compressed

**Step 2: Organization of relevant work**

In this section, you will organize the papers from above into groups of papers that have similar techniques, strengths, and/or limitations. For example, you might group papers by the type of methods used (e.g., deep learning vs. other techniques for classification) or by their limitations (e.g., studies that showed poor vs. strong performance on imbalance datasets).

There is not a specific format for this section, as long as you clearly show how you have organized your papers from Step 1. This is meant to help you prepare to write your Related Work section in your written report. You can refer to each paper by its in-text citation (e.g., Rußwurm et al., 2023 in the earlier example).

Here are some suggested resources to review to help you prepare to write a good Related Work section based on your literature review:

* Carnegie Mellon University pdf and video on preparing a literature review: <https://www.cmu.edu/student-success/other-resources/resource-descriptions/related-work.html>
* Related Work slides from Penn State University: <https://www.cse.psu.edu/~pdm12/cse544/slides/cse544-relwork.pdf>

The papers are organized by the following types of methods used to embed, detect, or remove watermarks: model free methods, adversarial learning, and convolutional neural networks.

Model Free Methods:

[3] Tali Dekel et al., 2017

[4] Yongjian Hu et al., 2005

[5] Kankanhalli, M. S. et al., 1999

[8] Hector Santoyo-Garcia et al., 2017

The most primitive methods of embedding, detecting, and removing watermarks from images use various algorithms without any deep learning techniques. There has been lots of research done over the last few decades to make watermarks harder to detect and remove in order to protect the images’ copyrights. One method to make watermarks harder to detect is by embedding watermarks in a way that adapts to the texture, edge, and luminance of the pixels in the underlying image [5]. This causes variations between different watermarks, making it difficult for detection algorithms to learn the patterns. Another method for embedding watermarks is an algorithm based on a user-key structure that adapts to the host image’s features [4]. With the correct key, the watermark can be removed without affecting the original image. If the watermark is removed illegally, then the image will have a lower quality.

For automatic detection of watermarks, one algorithm uses total variation based on L1 norm [8]. The watermarked image is decomposed based on its structure and texture, and the watermark’s edges are distinguished from the host image’s edges. Another model free method [3] uses a simple approach for the detection and removal of watermarks simply using mathematical equations and image processing techniques. This method is based on the assumption of consistency and uniformity of the locations, size, and types of watermarks.

We believe that these methods can be effectively used once we have a localized region for the watermark, and in the next section we discuss the role of CNNs in achieving this.

Convolutional Neural Networks:

[2] Cheng, D. et al., 2018

Convolutional Neural Networks is the traditional approach for any kind of object detection. This is due to the fact that they are much better suited to detect features, and simplify the processing of features in high dimensional data such as images. There is only one paper in this given grouping as this is the only approach we found that was using object detection using CNNs as an initial step. The method proposed in [2] is a two-step method, consisting of two specialized CNN models for two different tasks. The first task is to locate the watermark in the given image, and thus the first CNN model is an object detection model. The second model is trained as an Image Translation model, which takes in the region of pixels affected by the watermark and attempts to convert them back to their original form. We believe that by taking the learnings of this paper from task one, we can build a strong watermark detection model, and then try different approaches to convert the pixels in the watermarked region, using both Model Free and Model Based approaches. This is how we arrived at our initial technical approach in the proposal.

The CNN-based model’s method is adept at detecting/identifying the watermarks in the images, but the step for the reconstruction of the images leads to the creation of some artifacts. Adversarial methods are able to deal with this problem by using a generator-discriminator model.

Adversarial Learning:

[1] Chen, X. et al., 2021

[6] Zhao et al., 2019

[7] Jing Liang et al., 2021

Adversarial Learning-based approaches have become more prevalent for watermark removal techniques. An approach using a Conditional GAN model that reconstructs images with a patch-based discriminator trained on reconstructed and original images [6] is able to deal with the problem of reconstruction of artifacts produced. This model incorporates a patch-based discriminator which uses the reconstructed image and original image to improve the reconstruction by reducing traces of watermarks (artifacts) that were left by the original model. This patch-based discriminator can thus be used on top of the CNN-based model previously mentioned to improve the results by removing the artifacts. However, it is difficult to train the discriminator to differentiate between the reconstructed and original images, as it could lead to overfitting.

Other recent approaches using Adversarial Learning are able to perform significantly well on the watermark removal task, but these methods utilize substantially greater computational power. One such approach is a general-purpose watermark removal framework called REFIT [1]. This technique uses fine-tuning and two different methods, EWC algorithm and unlabeled data augmentation, to remove watermarks without affecting the model's functionality under a weak threat model. It can remove watermarks against a wide range of watermarking schemes. Although it only evaluated image classification models under a weak threat model, this technique shows promising results in removing watermarks without affecting the model's performance or accuracy. An alternative approach used in [7] involves a deep learning framework based on an encoder-decoder structure for the task and is able to perform well even in the case when the images used have lower quality due to image compression or image resizing.