Chapter 8

Conclusion

This final chapter of the thesis is structured as follows. It begins by describing to what extent the proposed research questions (in Sec. 1.1) are addressed. Thereafter, a brief summary of the thesis and the future work is elucidated in Sec. 8.2 and Sec. 8.3 respectively.

8.1 Evaluation of research questions

There are two broad research questions concerning each part of the thesis and sub-research questions within each part. A brief evaluation of these questions is listed below:

1. *How can image noise be leveraged for source camera identification?*

Deep convolutional neural networks are a category of neural networks that have significantly furthered the state-of-the-art in several machine-learning-based applications where images are used as inputs to the system. This thesis examines how ConvNets can be used for digital image forensics which leverage on the image noise. Although it is not the first such attempt, the methods proposed in this work advance the state-of-the-art on publicly available datasets. We further point out that, for forensic tools to be admitted in courts of law explainability of the model is necessary. In contrast to the traditional explainable models for source camera identification, ConvNets offer higher performance. ConvNets being a class of black-box machine learning models, it is necessary to conduct large-scale validation of these models. We are of the view that large-scale validation would help increase the trust in ConvNet-based methods. Since access to such large-scale data sets was not available during this project, we show improvements only on the publicly available data sets. The specific sub-research questions are addressed as follows.

1.1 Given a pair of images, how likely are they to be captured using the same camera device?

Quantifying the likelihood of a pair of images coming from the same camera device is very useful for LEAs, without needing to know the specific camera device. To this end, in Ch. 2, a two-part network is proposed where the first part extracts a fingerprint from each image and the second part compares the pair of fingerprints and gives a similarity score. The results achieved in these experiments are promising and show that this direction is worth pursuing. The results are not yet forensic-ready but certainly promising. Furthermore, the experiments were conducted in a closed-set scenario, i.e. the models were trained to extract signatures from images coming from a known set of cameras. On the other hand, the problem of open-set camera identification is not addressed in this work and is planned for future work.

1.2 Are RGB values alone sufficient to achieve reliable camera identification?

In forensic investigations, the greater the evidence better would be the outcome. Therefore, other forms of data (such as image meta-data) are certainly helpful. However, to answer the question the analysis is restricted to only the RGB pixel data. We measure to what extent a reliable source camera identification can be performed with just image-pixel data. We observe that camera brand and model can be identified with a high degree of accuracy. Identifying the camera device needs further investigation. This is not the first work to address this research question, however, we improve the state-of-the-art in all our attempts using only RGB image pixel data.

1.3 Do some regions in an image have a greater presence of camera noise when compared to others? Can we leverage them for increased robustness in camera identification?

It is known that the flat frames (i.e images produced when the imaging sensor is exposed to a uniform light source) contain maximum sensor noise. These images can be thought of as a scene representing the homogeneous region. With this background information, a system was designed (in Ch. 3) to extract small patches from the input image and to classify them. In general, the probability of a small patch being homogeneous is much greater than the whole image being homogeneous. We leveraged this observation and built a system to extract and classify homogeneous patches from an image for source camera model identification. Such a system performed better than the state-of-the-art for camera identification on the Dresden benchmark data set.

1.4 With newer camera brands, models, and devices coming under the ambit of LEAs, is it possible to update the existing models with little computational overhead?

ConvNets are often trained to classify several hundreds or even thousands of classes. While this is fine when the number of classes is fixed, however, with the modification of the number of classes the entire system must be re-trained. To minimise this repetitive computation, a hierarchical classification setup is proposed in Ch. 3. The first level is a brand classifier, depending on its prediction outcome, the specific model-level classifier of that brand is used. Subsequently, the predicted model can be used to determine a device-level classifier for the predicted camera model. With such a hierarchical design, the addition/modification of a specific camera device would only affect the relevant classifiers instead of the whole system. Interestingly, such a hierarchical system also achieves slightly better classification accuracy than a traditional flat classifier.

1.5 How does the performance of a Source Camera Identification (SCI) system vary for different social media compression? For instance, if an original video is shared via WhatsApp or YouTube how will it impact the overall performance?

In principle, the video quality gets affected along with the forensic traces. In this work, we explore, even with the modified traces can a system be learnt to recognise the source camera device given a video. Therefore, three compressions were included during both train and evaluation - native, WhatsApp, and YouTube. We observed that, even with WhatsApp and YouTube compressions, the system is able to identify these videos on par with the accuracy achieved on the native videos. This is a promising result for media forensics.

1.6 Do some video frames have more forensic traces than others?

There are three types of video frames (I, P, and B). Among these, I-frames are the least compressed while P and B require I-frames for decoding. In Ch. 5 we explore the role played by I-frames and empirically observed that such frames do indeed result in slightly better camera identification. Frame selection based on scene content was also investigated. In particular, homogeneity of scene content was used as a policy to choose frames. Unlike previous experiments with images, video frames did not result in any improvement with such scene-based frame selection.

1.7 Do constrained convolutions help in suppressing scene content? If so, does such a mechanism integrate well with every ConvNet?

Constrained convolutions for single channel images was proposed by Bayar and Stamm (2018a) for scene content suppression. We extended this for RGB images and experimented with its effectiveness in various settings of ConvNets. It was observed that such a scheme works well for shallow ConvNets such as MISLNet (see Ch. 4) while not quite with deeper ConvNets (refer Ch. 5) such as MobileNet-v3-small and ResNet50.

2. *How can image noise be leveraged for increased robustness of convolutional neural networks?*

A common strategy to improve the robustness of ConvNets is to train the models with additional data. In particular to train the models with corrupted versions of the data that would be present also in the evaluation set. Research on such data augmentation-based methods is quite popular, in comparison to the research on orthogonal methods. In this work, we propose two methods - one is a data pre-processing step while in the other a computational unit to replace the first convolutional layer in a ConvNet is proposed.

2.1 *With a focus on data-centric learning, can the noise in the input data be suppressed at the stage of data pre-processing?*

We propose a methodology (refer Ch. 6) that works well with Gaussian and uniform random noise type image corruptions encountered during the test. The proposed methodology can be used in scenarios wherein the high-level shape of the object is of importance, in comparison to the fine texture-level details. Input images are pre-processed using the delineation maps generated by the CORF push-pull inhibition operator. Such a scheme exhibited improved robustness to noisy data during evaluation while achieving comparable results on the noise-free data.

2.2 *With a goal to improve robustness, is it possible to bake robustness directly into the model itself (by means of architectural changes)?*

In Chapter 7 we propose a computational unit namely a push-pull convolutional unit, to replace the traditional convolutional unit in the first layer of the ConvNet. The motivation behind the design of this unit is elucidated in Chapter 7. The proposed methodology was rigorously tested against 15 types of image corruptions and robustness was achieved with little compromise on clean test data.

8.2 Summary

This thesis addresses two important applications of source camera identification and improved robustness of ConvNets while leveraging image noise. The focus of the first part of the thesis is on the forensic analysis of digital images and videos. This helps LEAs to gather additional intelligence in identifying the person behind such content. This work has an important societal application to fighting child sexual abuse. The second part focuses on making ConvNets robust to unseen image corruptions, the techniques of which can be used for various applications. Both these parts are concerned with the noise in the input image, wherein, in the former case the presence of noise is crucial while in the latter case, it is suppressed.

The first part of the thesis focuses on source camera identification, where the noise in the input image is leveraged to our advantage for the identification of the source camera from images or video frames. In comparison to traditional approaches (Goljan, 2008), recent methods based on ConvNets (Cozzolino and Verdoliva, 2019) have shown significant progress. To this end, this thesis is limited to the study of ConvNets for source camera identification. Furthermore, this work is restricted to identifying the source camera based on only the pixel data (without considering the accompanying image meta-data). This was done to safeguard the system from relying on meta-data, the tampering of which may go undetected.

Chapters 2 and 3 deal with source camera identification from images. Forensic investigators could be faced with a situation to ascertain if two images come from the same camera device. In order to answer this question, a system design is proposed (Ch. 2) and validated with experiments for device-based image matching. This consists of a two-part network, the first one extracts forensic signature from an input image and the second network takes a pair of signatures to compute the similarity score between them. The experiments on the *jpeg* subset of the Dresden data set demonstrated the potential of this approach with scope for interesting future work. Chapter 3 poses the question, "Do some regions in an image have a greater presence of camera noise when compared to others"? Based on the observations that the sensor noise can be extracted when it is exposed to a uniformly lit scene and the scene content distorting the sensor noise, it was hypothesized to prefer homogeneous regions in an image for forensic analysis. It was shown that when such input data is trained in a hierarchical fashion, it results in a classifier that is computationally efficient, modular and more effective than a flat (single classifier) approach. By means of thorough experiments on the *natural* subset of the Dresden data set, we achieve the best-ever classification accuracy of 99.01% for camera model identification.

Chapters 4 and 5 concerns source camera identification from videos. Analogous

to camera identification from images, video frames were used for the source classification of a video. Among the three types of video frames, I-frames were found to contain more forensic traces (as they would be unaffected by video stabilization). An extended constrained convolutional layer was proposed for scene content suppression from RGB images (refer Ch. 4) for shallow ConvNets. It was realized that such a scene-suppression scheme is counter-productive for sophisticated ConvNets. The impact of video compression on forensic traces was also studied. In particular, the performance of the system remains consistent even when the native videos are subjected to YouTube and WhatsApp compression. Extensive experiments were conducted on the VISION and the QUFVD data sets along with various sophisticated ConvNets to demonstrate the effectiveness of the methods.

Although machine learning systems were traditionally evaluated by keeping the test set distribution the same as the train, in practice, the deployed systems experience distributional shifts due to several uncontrolled factors. Therefore, in the second part of the thesis, noise suppression for robust ConvNets is investigated. In particular, two independent directions were explored. In the first direction, a pre-processing step to improve model robustness is proposed. The second approach consists of making architectural changes for improved model robustness to image corruptions.

Chapter 6 focuses on image pre-processing for robust image classification. A transformation step has been proposed that attempts to enhance the generalization ability of the ConvNets. Concretely, the delineation maps of given images are determined using the CORF push-pull inhibition operator. Such an operation transforms an input image into a space that is more robust to noise before being processed by a Convolutional Neural Network (ConvNet). Our experiments on the Fashion MNIST data set with AlexNet showed that the proposed CORF-augmented pipeline exhibits substantially higher generalization ability for additive Gaussian and uniform noise than a conventional AlexNet without the CORF transformation step. While achieving comparable results on noise-free images.

Chapter 7 investigates out-of-distribution robustness for ConvNets by making architectural changes. A computational layer namely PushPull-Conv is proposed to replace the traditional convolutional one, in the first layer of Convolutional Neural Network (ConvNet). The design of PushPull-Conv is brain-inspired and consists of excitatory and inhibitory (or push and pull) computation paths. This methodology helps achieve robustness against high-frequency corruptions with a little compromise on clean test data. We conducted extensive analysis and experiments on CIFAR10-C and ImageNet-C data sets to validate our hypothesis.

In summary, this work leverages image noise and proposes better solutions for two vital applications namely source camera identification and improved robustness of convolutional neural networks. The proposed methods advance the state-of-the-art in the respective research problems and help address vital societal problems.

8.3 Outlook

The following research directions could be considered for future investigations:

It is important to point out that, the methods proposed in this thesis for source camera identification have been rigorously tested on the limited publicly available data sets. However, for its acceptability in the courts of law, it is important to have methods that are explainable. Since ConvNets are a class of black-box methods, instead of explainability, large-scale validation is absolutely necessary for their acceptability. This is an ongoing debate between law practitioners and forensic researchers and it is our opinion that ConvNets with high accuracy and extensive validation should replace traditional approaches with explainability but offer relatively lower accuracy. This aspect of large-scale validation must be considered in all future works of source camera identification.

Device-based image matching introduced in this thesis (refer to Ch. 2) addresses a practical situation faced by forensic experts. This methodology needs to be developed further, for instance, to be able to match the source similarity of two images both of which are not in the database. This is also known as open-set identification. This will allow the investigator first to gather all photos captured by the same camera device, without needing to have a reference camera fingerprint of that device in the database. The technique adopted in this thesis addresses the problem of closed-set identification, i.e. when a reference fingerprint is already present in the database.

The methodology adopted in this thesis took a focused approach to consider only RGB pixel values for source camera identification. It would be interesting to design a system combining features from structural meta-data and RGB pixel values. The term meta-data in this thesis refers to the 'values' of some specific headers (or blocks) in the EXIF header of an image file which can be easily modified without traces. The structure of this meta-data, i.e. the arrangement of these blocks when the image file is read in bytes, is referred to as structural meta-data. It requires niche expertise to modify the structure of the meta-data without leaving behind traces. Moreover, the structure of the meta-data is almost always unique for each camera model and every editing software modifies this structure in a unique way. Therefore, structural meta-data can be used for forensic

investigations and identifications of the camera model. When combined with RGB data (i.e. sensor noise), this has the potential to become a robust system for camera device identification.

Most of the data sets concerning source camera identification have a fixed set of scenes that are captured with each device present in the data set. Therefore, the same scene could be present in two different images captured from two different devices. This could lead to two different scenarios while classifying images to identify the source device: (1) in the ideal case, the classifier learns to extract the sensor noise and ignores the scene content, and (2) in the worst case, the classifier learns to classify the scene and would lead to misleading results. To overcome the second scenario it is important to de-bias the scene content during training. For instance, a simple strategy to perform de-biasing would be to set up multi-task classification (camera identification and scene recognition) and to decrease the loss for camera identification while increasing the loss for scene recognition.

Concerning scene suppression in videos, this thesis explored the role of constrained convolutions. Another strategy for scene suppression could be to exploit the temporal information. In almost static parts of the video, two consecutive video frames share a high degree of the same scene content. By subtracting such adjacent frames from each other, scene content could be largely removed leaving behind the residual noise. This strategy could aid in the extraction of camera fingerprints from a video. Moreover, videos are subjected to stabilization to account for shakes and minor movements of the camera. This could result in displacement or shearing of camera noise. Therefore, methods to destabilize the video frames could be adopted as a pre-processing step before the extraction of the camera noise for robust camera identification.

Chapter 6 presented a data pre-processing technique to extract CORF push-pull contour maps for training, to make the ConvNets robust to unseen corruptions. One can for instance include these maps as additional channels to the input image. This will enable neural networks to extract features from the entire input space instead of just the CORF-based contour maps. This could help the network to also learn the texture details along with an emphasis on the contours.

Proper data augmentation is known to improve model robustness (Yun et al., 2019; Cubuk et al., 2019; Hendrycks et al., 2019b, 2021). The robustness techniques presented in this thesis are independent of data augmentation. Therefore, it would be interesting to study the performance of a system when non-data-augmentation-based methods are used in conjunction with data-augmentation-based methods. We speculate that introducing robustness at both input data and to the model, would help improving the robustness of the overall system. Data augmentation would help the model learn a fixed set of corruptions and model-level robustness would

help improving robustness to unseen data.

In conclusion, a part of this work was done to address the problem of source camera identification from digital images or videos while leveraging image noise. Instead of conventional approaches, this work makes use of ConvNets and pushes the state-of-the-art further for both image and video-based source camera identification on the publicly available Dresden, VISION, and QUFVD data sets. As mentioned earlier, in addition to these publicly available data sets, large-scale validation on privately held-out data sets is necessary to rigorously test these methods for acceptability in the courts of law. The second part of this thesis explores two different directions to improve model robustness to unseen data during the training. This is yet another challenge that is encountered with every deployed AI system (based on ConvNets). Both the proposed methods are novel and improve model robustness to a certain category of image corruption, pushing further the state-of-the-art to out-of-distribution robustness.

Appendices

8.A Summary of thesis in Spanish

Esta tesis aborda dos aplicaciones importantes de la identificación de la cámara de origen y la mejora de la robustez de ConvNets aprovechando el ruido de la imagen. La primera parte de la tesis se centra en el análisis forense de imágenes y vídeos digitales. Esto ayuda a las LEAs a reunir inteligencia adicional para identificar a la persona que se esconde tras dichos contenidos. Este trabajo tiene una importante aplicación social en la lucha contra el abuso sexual infantil. La segunda parte se centra en hacer que una ConvNets sea robusta frente a corrupciones invisibles de imágenes, cuyas técnicas pueden utilizarse para diversas aplicaciones. Ambas partes se ocupan del ruido en la imagen de entrada, siendo crucial en el primer caso la presencia de ruido, mientras que en el segundo se suprime.

La primera parte de la tesis se centra en la identificación de la cámara de origen, donde el ruido en la imagen de entrada se aprovecha en nuestro beneficio para la identificación de la cámara de origen a partir de imágenes o fotogramas de vídeo. En comparación con los enfoques tradicionales (Goljan, 2008), los métodos recientes basados en ConvNets (Cozzolino and Verdoliva, 2019) han mostrado avances significativos. Por ello, esta tesis se limita al estudio de ConvNets para la identificación de cámaras de origen. Además, este trabajo se limita a identificar la cámara de origen basándose únicamente en los datos de píxeles (sin tener en cuenta los metadatos de la imagen que la acompañan). Esto se ha hecho para evitar que el sistema dependa de los metadatos, cuya manipulación podría pasar desapercibida.

Los capítulos 2 y 3 tratan sobre la identificación de la cámara de origen a partir de imágenes. Los investigadores forenses podrían necesitar determinar si dos imágenes proceden del mismo dispositivo de cámara. Para responder a esta pregunta, se propone un diseño de sistema (Capítulo 2) y se valida con experimentos para la comparación de imágenes basada en dispositivos. Consiste en una red de dos partes, la primera extrae la firma forense de una imagen de entrada y la segunda red toma un par de firmas para calcular la puntuación de similitud entre ellas. Los experimentos con el subconjunto *jpeg* del conjunto de datos de Dresde demostraron el potencial de este enfoque, con margen para interesantes trabajos futuros. El capítulo 3 plantea la siguiente pregunta: "¿Algunas regiones de una imagen tienen una mayor presencia de ruido de cámara en comparación con otras?" Basándose en las observaciones de que el ruido del sensor se puede extraer cuando se expone a una escena uniformemente iluminada y el contenido de la escena distorsiona el ruido del sensor, se planteó la hipótesis de

preferir regiones homogéneas en una imagen para el análisis forense. Se demostró que cuando tales datos de entrada se entrenan de forma jerárquica, se obtiene un clasificador que es modular y más eficaz que un enfoque plano (clasificador único). Mediante experimentos exhaustivos en el subconjunto *natural* del conjunto de datos de Dresde, logramos la mejor precisión de clasificación de la historia del 99,01% para la identificación de modelos de cámara.

Los capítulos 4 y 5 se centran en la identificación de cámaras fuente a partir de vídeos. De forma análoga a la identificación de cámaras a partir de imágenes, se utilizaron fotogramas del vídeo para clasificar la fuente de un vídeo. Entre los tres tipos de fotogramas de vídeo analizados, se observó que los fotogramas tipo "I" contenían más rastros forenses (ya que no se veían afectados por la estabilización del vídeo). Se propuso una capa convolucional restringida ampliada para la supresión del contenido de la escena a partir de imágenes RGB (véase capítulo 4) para ConvNets de poca profundidad. Se observó que dicho esquema de supresión de escenas es contraproducente para ConvNets sofisticadas. También se estudió el impacto de la compresión de vídeo en los rastros forenses. En particular, el rendimiento del sistema se mantiene constante incluso cuando los vídeos nativos se someten a la compresión de YouTube y WhatsApp. Se realizaron amplios experimentos con los conjuntos de datos VISION y QUFVD junto con varias ConvNets sofisticadas para demostrar la eficacia de los métodos.

Aunque los sistemas de aprendizaje automático se evaluaban tradicionalmente manteniendo la distribución del conjunto de prueba igual a la del de entrenamiento, en la práctica, los sistemas desplegados experimentan cambios distribucionales debido a varios factores no controlados. Por lo tanto, en la segunda parte de la tesis, se investiga la supresión de ruido para ConvNets robustos. En concreto, se exploraron dos direcciones independientes. En la primera dirección, se propone preprocesamiento previo para mejorar la robustez del modelo. El segundo enfoque consiste en realizar cambios en la arquitectura para mejorar la robustez del modelo frente a las corrupciones de la imagen.

El capítulo 6 presenta una técnica de preprocesamiento de datos para extraer mapas de contorno CORF push-pull para el entrenamiento, que permite hacer las ConvNets robustas a corrupciones no vistas previamente. Por ejemplo, se pueden incluir estos mapas como canales adicionales a la imagen de entrada. Esto permite a las redes neuronales extraer características de todo el espacio de entrada en lugar de sólo los mapas de contorno basados en CORF. De esta manera, se ayuda a la red a aprender también los detalles de textura prestando también especial atención a los contornos.

Se sabe que un aumento adecuado de los datos mejora la robustez del modelo (Yun et al., 2019; Cubuk et al., 2019; Hendrycks et al., 2019b, 2021). Las técnicas de

robustez presentadas en esta tesis son independientes del aumento de datos. Por lo tanto, sería interesante estudiar el rendimiento de un sistema cuando se utilizan métodos no basados en el aumento de datos junto con métodos basados en el aumento de datos. Especulamos que la introducción de robustez tanto en los datos de entrada como en el modelo ayudaría a mejorar la robustez del sistema global. El aumento de datos ayudaría al modelo a aprender un conjunto fijo de corrupciones y la robustez a nivel de modelo ayudaría a mejorar la robustez ante datos no vistos.

En conclusión, una parte de este trabajo aborda el problema de la identificación de la cámara de origen a partir de imágenes digitales o vídeos aprovechando el ruido de la imagen. En lugar de basarse en enfoques convencionales, este trabajo hace uso de ConvNets y lleva más allá el estado del arte para la identificación de la cámara de origen basada tanto en imágenes como en vídeos en los conjuntos de datos disponibles públicamente Dresden, VISION y QUFVD. Como se ha mencionado anteriormente, además de estos conjuntos de datos de acceso público, es necesario una validación a gran escala en conjuntos de datos privados para probar rigurosamente la aceptabilidad de estos métodos en los tribunales de justicia. La segunda parte de esta tesis explora dos direcciones diferentes para mejorar la robustez del modelo ante datos no vistos durante el entrenamiento. Este es otro reto que se encuentra con cada sistema de IA desplegado (basado en ConvNets). Los dos métodos propuestos son novedosos y mejoran la robustez del modelo frente a una determinada categoría de corrupción de imágenes, llevando aún más lejos el estado del arte de la robustez fuera de distribución.