Plug to Place: Indoor Multimedia Geolocation from Electrical Sockets for Digital Investigation

Kanwal Aftab[1], Graham Adams[2], Mark Scanlon[1]

[1]*Forensics and Security Research Group, School of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland*
[2]*School of Engineering, Case Western Reserve University, Cleveland, Ohio, United States*

Abstract

Computer vision is a rapidly evolving field, giving rise to powerful new tools and techniques in digital forensic investigation, and shows great promise for novel digital forensic applications. One such application, indoor multimedia geolocation, has the potential to become a crucial aid for law enforcement in the fight against human trafficking, child exploitation, and other serious crimes. While outdoor multimedia geolocation has been widely explored, its indoor counterpart remains underdeveloped due to challenges such as similar room layouts, frequent renovations, visual ambiguity, indoor lighting variability, unreliable GPS signals, and limited datasets in sensitive domains.

This paper introduces a pipeline that uses electric sockets as consistent indoor markers for geolocation, since plug socket types are standardised by country or region. The three-stage deep learning pipeline detects plug sockets (YOLOv11, mAP@0.5 = 0.843), classifies them into one of 12 plug socket types (Xception, accuracy = 0.912), and maps the detected socket types to countries (accuracy = 0.96 at >90% threshold confidence). To address data scarcity, two dedicated datasets were created: socket detection dataset of 2,328 annotated images expanded to 4,072 through augmentation, and a classification dataset of 3,187 images across 12 plug socket classes. The pipeline was evaluated on the Hotels-50K dataset, focusing on the TraffickCam subset of crowd-sourced hotel images, which capture real-world conditions such as poor lighting and amateur angles. This dataset provides a more realistic evaluation than using professional, well-lit, often wide-angle images from travel websites. This framework demonstrates a practical step toward real-world digital forensic applications. The code, trained models, and the data for this project is open source and can be obtained from [GitHub redacted for blind review].

Keywords:
Multimedia Geolocation, Computer vision, Hotels-50K, Indoor, Multimedia Forensics, Human Trafficking

1. Introduction

Human trafficking is a severe global crime affecting millions across all ages, genders, and backgrounds, causing deep personal, community, and societal harm [40]. It entails the illegal trade of people through deception, violence, or exploitation, resulting in forced labour, sexual abuse, and organ trafficking [11]. Given its severe and long-lasting impact, the fight against human trafficking is explicitly prioritised under 3 of the United Nations Sustainable Development Goals (SDGs) [38, 2]. In addition, child sexual exploitation material (CSEM) investigation is one of the most common case types encountered in digital forensics laboratories within law enforcement agencies worldwide [16].

With rapid technological advancement, the rise of feature-rich smartphones, expanded storage capabilities, widespread internet access, and the growing influence of social media, nearly every facet of modern life has become digital [20, 26]. However, these same technologies are increasingly exploited by offenders to facilitate serious crimes, underscoring the critical importance of digital forensics [13]. As the volume of digital data continues to surge, the analysis and interpretation of digital evidence have become indispensable to modern investigations [13, 29]. The application of artificial intelligence (AI) to digital forensic investigation is still very much in its infancy [7], but nonetheless, practitioners have already identified image/media classification as having the most potential for the future use of AI in their investigations [16]. In parallel, hotel recognition has become a common need for human trafficking investigations, as hotels are frequently used as intermediary stopover locations during the trafficking of victims [23]. Online human trafficking advertisements or the interception of organised crime gang's electronic communications are often the starting points for many human trafficking investigations. Identifying hotel rooms from these photographs is an extremely arduous task. Indeed, police agencies often resort to crowdsourcing their locations. For example, hotel rooms often feature in the Trace an Object projects run by Europol[1] or the Australian Centre to Counter Child Exploitation[2], asking for the general public's help in identifying the hotels during investigations of cases involving child sexual exploitation material. In the DF-Pulse 2024 survey [16], digital forensic practitioners identified image/media classification and CSAM investigation as two of

[1] https://www.europol.europa.eu/stopchildabuse
[2] https://www.accce.gov.au/what-we-do/trace-an-object

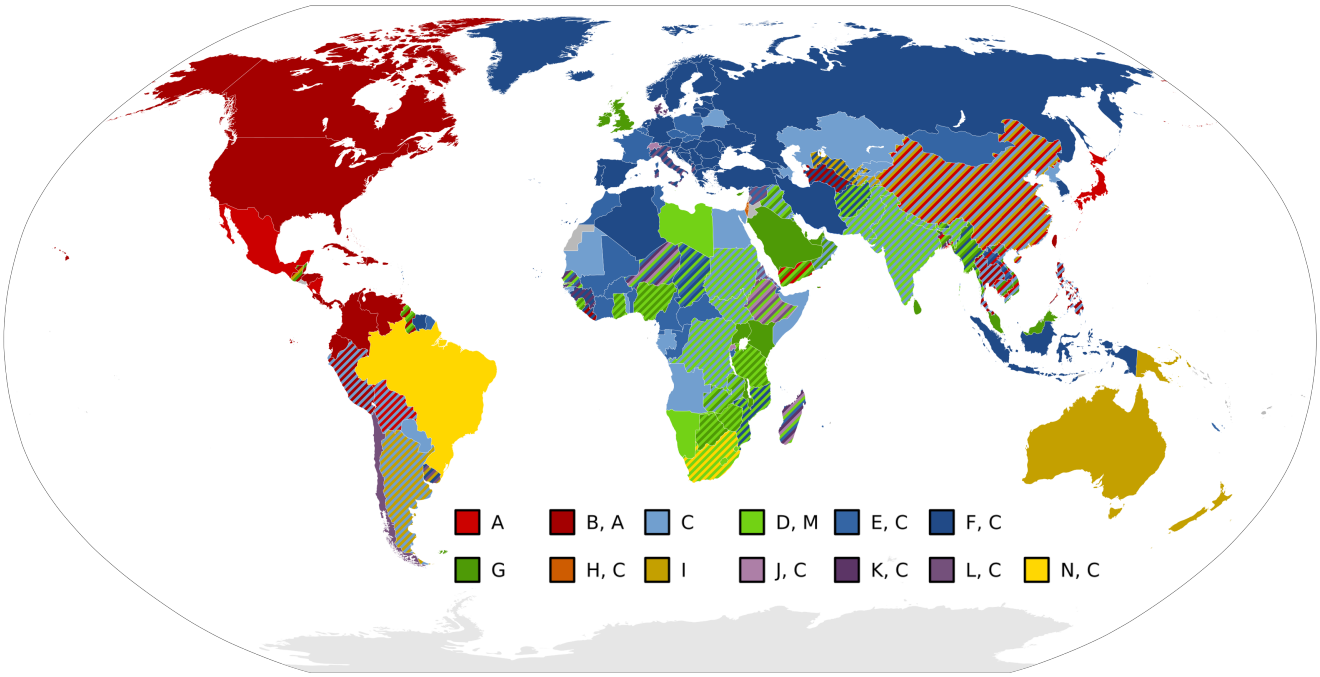


Figure 1: Worldwide plug type distribution map ©

the main areas where artificial intelligence has the potential to assist in their future cases.

In terms of automated geolocation, indoor geolocation specifically faces significant challenges. Hardware camera sensors are often unreliable indoors, and image metadata is frequently stripped during online or instant messaging sharing – making it difficult to trace an image's origin [4]. While outdoor environments typically provide clear geolocation cues, such as landmarks and infrastructure [43, 8, 25], indoor settings are far more complex. Similar layouts, recurring furniture, and inconsistent lighting make it challenging to reliably differentiate between locations [5]. Amid these challenges, it is therefore crucial to identify consistent and distinctive cues for indoor geolocation. One such feature of indoor environments is the presence of electric sockets, which can serve as distinctive and geographically informative visual markers. Each country adheres to standardised socket designs governed by national or regional electrical regulations [18]. Using Computer Vision (CV), detecting and classifying these plug sockets can provide a reliable cue for narrowing down the search space, as illustrated in Fig. 1.

1.1. Contribution of this Work

This paper makes the following key contributions:

- Dataset Creation and Release. Two novel CV datasets have been curated and publicly shared: (i) a socket detection dataset for identifying sockets in indoor room images and (ii) a socket classification dataset containing 12 socket categories. These datasets provide valuable benchmarks for advancing research in fine-grained indoor object recognition.
- Algorithmic Pipeline for Detection and Classification. A comparative study of state-of-the-art detection models and Convolutional Neural Network (CNN) classifiers has been conducted. Based on this analysis, a pipeline was designed to select the most effective combination of detection and classification models for socket recognition.
- Evaluation of the Proposed Approach on Real-World Data. The approach is evaluated on the Hotels-50K dataset, specifically the TraffickCam subset, demonstrating its practical utility for law enforcement by narrowing down search spaces in real-world human trafficking investigations. Furthermore, this work aims to lay the foundation for a universal socket detector, enabling broader applications beyond the law enforcement/investigative domains.

2. Related Work

To address the challenge of indoor multimedia geolocation, specifically hotel recognition, researchers have employed a variety of techniques, ranging from hand-crafted feature extraction for colour-based image retrieval (CBIR) and image classification to neural networks for automated feature extraction, as well as more advanced approaches such as object-based similarity, image embeddings, and semantic scene understanding. Regardless of the methodology employed, the fundamental building block remains CV. In this context, high-quality data remains essential. However, obtaining such data can be particularly challenging in sensitive scenarios, such as human trafficking investigations. Recognising these ethical and operational challenges, the European Union formalised AI regulations through

the AI Act [14], aiming to maximise the benefits of AI while enforcing rigorous ethical and safety standards.

In the context of data collection and processing, the creation of specialised datasets has been significant. The Hotels-50K dataset [35] contains over one million labelled images from 50,000 hotels worldwide, sourced from both travel websites and the TraffickCam mobile application, and is designed specifically for hotel recognition research. The Hotel-ID dataset [23] provides a similarly large-scale resource for the same domain. Both datasets serve as benchmarks for evaluating recognition approaches.

Building on this extensive data, Tseytlin and Makarov [37] framed large-scale hotel recognition as a deep metric learning task. Traditional methods using Contrastive and Triplet losses often fail to capture all intra-batch similarity information. To address this, the authors proposed the Contrastive-Triplet loss, combining both losses with no added computational cost. Their main result is that, on Hotels-50K, this approach boosts mAP@R to 0.039 versus 0.031 (Contrastive) and 0.029 (Triplet), and lifts P(1) to 0.300 from 0.268 (Contrastive) and 0.255 (Triplet). On CUB200, it matches the Contrastive loss in mAP@R (0.260) and yields a more stable hyperparameter performance, underlining its robustness across datasets.

Deep learning models are often black-box systems, and their complex structures make it difficult to interpret how visual inputs lead to an output, obscuring the specific cues that drive similarity judgments. To address this, Black et al. [6] introduced a paired image similarity visualisation technique tailored for Transformer architectures. Their method enables the inspection of attention maps to reveal which regions contribute most to a given similarity score. When comparing ResNet and Vision Transformer (ViT) on datasets such as Hotels-50K, Clean Google LandmarksV2, and Stanford Online Products, the study found that ViT's attention focused on meaningful cues, such as bed covers, runners, and tiled shower walls. These results emphasise the need for explainable embeddings in visual retrieval. Similarly, Wazzan et al. [42] explored how context affects object matching, finding that a moderate amount of context improves annotation efficiency and retrieval accuracy, while excessive context complicates recognition in ambiguous scenes.

Beyond these pixel-level and embedding-based methods, some research now focusses on object-centric retrieval. This approach represents images as groups of distinct objects rather than just overall textures, aiming to bridge the gap between simple image features and what people actually see. For instance, Kim et al. [24] introduced a method to detect and segment the main object in an image, as subjects are often centred in photos. The authors used relevance feedback on these segments to improve retrieval, though their reliance on colour made it less effective in busy scenes. Later, Pradhan et al. [30] used visual saliency maps to identify object regions and create feature vectors for retrieval. This method worked well for images with a clear main object but was less effective for complex scenes. As Wadhai and Kawathekar [39] points out, CBIR systems based only on shape are quick and simple but struggle to handle significant visual differences between object categories.

Object-centric modelling works well for hotel recognition in sensitive investigations, such as human trafficking. Bhavanasi and Stylianou [5] proposed an ensemble approach that focusses on important objects in images, achieving 23.5% top-1 accuracy compared to just 3.9% for full-image models. This shows the value of pinpointing distinctive features, such as furniture, fixtures, or wall patterns. Additionally, semi-automated labelling pipelines [42] demonstrate that using less context speeds up labelling without hurting performance, suggesting that focussing on main objects and a moderate amount of context best balances interpretability and efficiency.

Besides these semantic methods, feature-level CBIR focusses on capturing broad visual qualities such as colour, texture, and layout [9]. Of these, colour is the most widely used and efficient descriptor [15, 1]. While RGB is common, colour spaces like CIELAB, Munsell, and fuzzy-based models often yield more meaningful results [34]. Using a mix of colour spaces, such as RGB, YCbCr, and Lab*, can also boost precision [32]. Specific to hotel room identification, Herrmann et al. [17] evaluated CBIR systems using colour features on Hotels-50K, achieving over 95% Top-50 accuracy with just two descriptors, thus supporting faster and more reliable investigative workflows. More recently, Bamigbade et al. [3] combined major colour palettes and simple histograms with deep metric learning and classification to raise top-20 retrieval accuracy by 17%. These findings show that blending handcrafted features with deep embeddings can improve both clarity and results.

Even with sophisticated methods, systematic reviews expose key gaps in using CV for social good. Dimas et al. [11] note that Operations Research and Analytics efforts for anti-human trafficking mostly target sex trafficking and prosecution, with less attention to labour trafficking, prevention, or victim protection. Similarly, Bamigbade et al. [4] highlighted the benefits of using CV based geolocation in investigations, calling for more data types and clearer deep learning models to extract useful information.

Overall, these studies show a shift in hotel recognition research: from data-driven deep metric learning to more comprehensible, object-focused, and feature-blended CBIR systems. This change not only improves technical performance but also supports the growing demands for ethical, transparent, and socially responsible AI in sensitive investigations.

3. Experimental Setup and Analysis

The CV methodology proposed as part of this paper is a three-tiered process, as detailed in Fig. 2. In the first stage, YOLO object detection identifies and localises electric sockets in an image, generating cropped regions of interest (ROI). In the second stage, a CNN classifies each cropped socket ROI by type. Finally, in the third stage, the detected socket type is mapped to potential countries, which narrows the list of likely locations and supports law enforcement investigations. Sections 4 to 6 provide a detailed discussion of each stage, including dataset preparation for every stage, the methodologies used, and a comparative analysis of different algorithms along with their results.

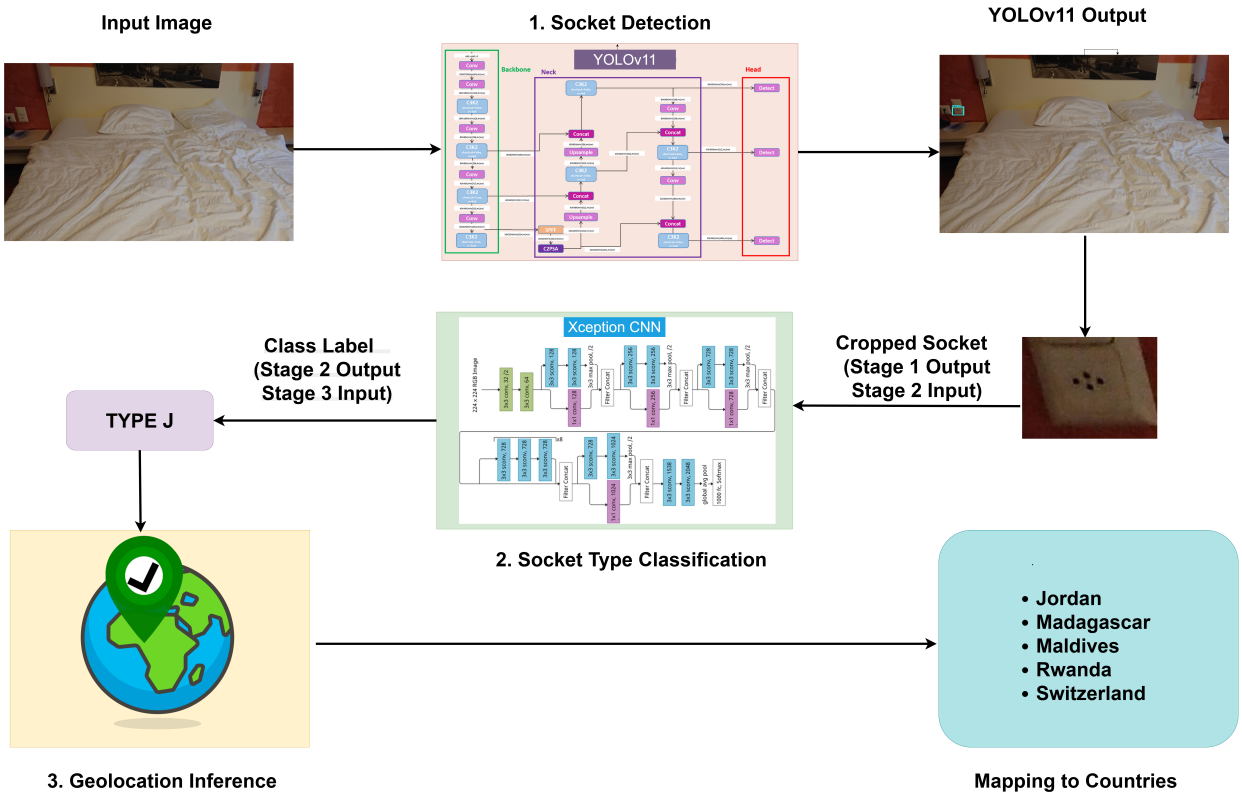


Figure 2: Architecture of the proposed three-stage pipeline: (1) Socket detection, (2) Socket type classification, and (3) Geolocation.

4. Stage 1: Socket Detection

Detecting electric sockets provides an automated method for inferring locations from visual data. This is useful for law enforcement when analysing photos from hotel rooms or other indoor locations where victims may have been exploited. Visual cues that help determine locations indoors are often difficult to find. However, electric sockets are consistent and recognisable markers. Each country or region uses specific socket types, identified by their pin configurations. These features reliably indicate a victim's geographic location. CV object detection automates this step. Modern algorithms can accurately identify objects and their positions in digital images. This enables large-scale and efficient analysis to aid investigations.

4.1. Dataset Preparation

High-quality and diverse datasets are critical in CV, as both the quantity and quality of training data directly affect model accuracy and generalisation [22]. For this study, a total of 2,328 socket images were compiled, of which 1,525 were obtained from publicly available Roboflow [31] socket datasets licenced under CC 4.0, while the remaining images were cropped from hotel room scenes in the Hotels-50K dataset. This original dataset (Dataset A) was partitioned into training (70%, 1,629 images), validation (20%, 455 images), and test (10%, 244 images) subsets. All images were annotated with bounding boxes using the Roboflow platform and manually classified into two categories: class 0 (NA) for switchboards and class 1 for sockets. Although the initial focus was exclusively on sockets, the frequent misclassification of visually similar switchboard buttons necessitated the inclusion of the NA class to better differentiate non-target elements from actual sockets. This confusion was heavily compounded by the small sizes of the objects, their low resolution in cropped images, and the similar colouring and high concentration of buttons and sockets on the switchboards.

To enhance dataset diversity and improve model generalisation, data augmentation was applied exclusively to the training set, doubling its size from 1,629 to 3,258 images. Augmentation techniques included random cropping (0–20%), rotation (–15° to +15°), grayscale conversion (15% of images), hue adjustment (–24° to +24°), and brightness adjustment (–19% to +19%). After additional images were added, the dataset resulted in training (80%, 3,258 images), validation (10%, 409 images), and test (10%, 407 images) subsets, for a total of 4,074 images. The validation and test sets were not augmented; however, an additional 163 images cropped from the Hotels-50K dataset were added to ensure an unbiased evaluation of model performance.

4.2. Methodology

Recent advances in CV have positioned object detection as a fundamental task, achieving performance that, in some cases, rivals or even surpasses human capabilities [27]. A wide range of state-of-the-art algorithms are available for object detection,

Table 1: Test results for all classes across models. Bold indicates the best value per column.

Version	Type	Epochs	Precision (P)	Recall (R)	mAP@0.5	mAP@0.5:0.95
YOLOv8	s	180	0.843	0.709	0.793	0.529
YOLOv8	s	250	0.834	0.709	0.790	0.524
YOLOv8	m	set=300 (early stopping at 233)	**0.883**	0.750	0.809	**0.540**
YOLOv11	s	180	0.847	0.734	0.804	**0.540**
YOLOv11	m	250	0.846	**0.791**	**0.832**	0.539
YOLOv11	l	set=250 (early stopping at 234)	0.851	0.764	0.806	0.534
YOLOv12	s	180	0.720	0.650	0.748	0.483

and the choice of an appropriate model for custom datasets depends heavily on the specific requirements of the application. Object detection algorithms are broadly classified as two-stage or single-stage detectors [36]. Two-stage models are more accurate but slower, while single-stage models offer faster inference with only a slight accuracy trade-off [10]. To handle large volumes of images efficiently, speed was prioritised, rendering single-stage object detectors a more practical option. The YOLO family was chosen for its combination of high speed and competitive accuracy [41], which is suitable for socket detection as a first-stage step before potential refinement.

Although newer YOLO versions introduce architectural improvements, the latest release is not always stable [19]. A comparative analysis was then conducted. The experimental outcomes, as outlined in Table 1, demonstrate the performance of each model across diverse metrics, starting with YOLOv8 as a benchmark, followed by YOLOv11 and YOLOv12. Each release provides multiple variants (nano, small, medium, large). This study predominantly focused on small and medium models, and in one case, the large model was also considered to balance computational cost with accuracy. Hyperparameter tuning, particularly the number of training epochs, was evaluated to minimise underfitting and overfitting, and the effect of data augmentation on model performance was also assessed.

4.3. Comparative Analysis and Results

This study evaluated the performance of various YOLO models for socket detection using Dataset A, the original dataset, in the first phase, and subsequently assessed the effect of data augmentation in the second phase on the augmented data, Dataset B. All models were trained with a batch size of 16 and an input resolution of 640 pixels. The optimiser was set to AdamW, which automatically tuned hyperparameters and overrode the default learning rate and momentum values, resulting in an effective learning rate of 0.001667 with momentum fixed at 0.9.

The experiment began with YOLOv8, starting with the small variant (YOLOv8s) to assess the effect of increasing the number of training epochs from 180 to 250. The focus then shifted to the medium variant (YOLOv8m) to evaluate the trade-off between model capacity and performance. Although the number of epochs was initially set to 300, training was stopped early at 233 epochs due to the early stopping mechanism designed to prevent overfitting.

The analysis then progressed to YOLOv11, beginning with the small variant (YOLOv11s) trained for 180 epochs, followed by the medium variant (YOLOv11m) trained for 250 epochs. For the large variant (YOLOv11l), training was scheduled for 250 epochs but was stopped early at 234 epochs due to early stopping. Finally, YOLOv12 was evaluated; however, it achieved comparatively lower performance metrics.

Specifically, YOLOv12 underperformed, achieving mAP@0.5 of 0.748 and mAP@0.5:0.95 of 0.483, with a precision of 0.720 and a recall of 0.650. These results indicate weaker socket detection performance, highlighting that cutting-edge models do not always outperform more established versions such as YOLOv8 and YOLOv11. Newly released models are often unstable and improve over time; therefore, it is generally advisable to allow them to mature before deployment in CV applications [21].

Validation accuracy should not be solely relied upon, as it may indicate overfitting on the training data and not necessarily reflect performance on unseen test data [33]. Therefore, test accuracy is considered more decisive than validation accuracy. On the test set, YOLOv8m (early stopping at 233 epochs) achieves the highest precision (0.883), while both YOLOv8m and YOLOv11s (180 epochs) achieve the highest mAP@0.5:0.95 (0.540). This indicates that these models are highly accurate in correctly detecting sockets, though their recall varies (0.750 for YOLOv8m, 0.734 for YOLOv11s). In contrast, YOLOv11m (250 epochs) attains the highest recall (0.791) and the highest mAP@0.5 (0.832), indicating strong overall detection coverage and good localisation.

To further assess the impact of data augmentation, YOLOv11m was trained on Dataset B with augmentation, while YOLOv11s was trained using both augmentation and 5-fold cross-validation. Table 2 summarises the results. Adding only augmentation slightly altered performance, with precision decreasing from 0.846 to 0.783, recall increasing from 0.748 to 0.765, mAP@0.5 decreasing from 0.832 to 0.766, and mAP@0.5:0.95 decreasing from 0.539 to 0.485. These results suggest that augmentation improves recall at a modest cost to precision and overall localisation accuracy.

In contrast, combining K-fold cross-validation with augmentation further enhances model robustness, reducing variance between folds and providing more reliable generalisation to unseen images. Specifically, average precision increased from 0.847 to 0.872, and recall increased from 0.734 to 0.756, demonstrating that K-fold training effectively mitigates overfitting on smaller datasets and improves overall detection performance. Based on this analysis, YOLOv11s with augmentation and K-fold cross-validation was identified as the best performer in Stage One socket detection. Fig. 3 and Fig. 4 show the visual

Table 2: Comparison of YOLOv11m variant with data augmentation and k fold cross validation on and Dataset B. Bold indicates the best value per dataset.

Dataset	Setting	Precision (P)	Recall (R)	mAP@0.5	mAP@0.5:0.95
Test (Dataset B)	Yolov11m without Aug	0.805	0.748	0.792	0.498
	Yolov11m with Aug	0.783	0.765	0.766	0.485
K fold cross validation	Yolov11s with Aug	**0.8675**	**0.7990**	**0.8427**	**0.5771**

socket detection results of this best-performing model.

5. Stage 2: Socket Type Classification

Across the world, there are 14 recognised types of domestic electric sockets, labelled A through N by the International Electrotechnical Commission [18]. While a unified global standard for electricity would be desirable, the reality is rather different: variations exist in plug shapes, socket designs, voltages, and frequencies. In the context of this study, however, the focus is solely on the visual characteristics of plug socket types.

As shown in Fig. 5, there are 14 standard socket designs, and their corresponding global distribution is illustrated in Fig. 1. Types A and B are prevalent in North and Central America and Japan, featuring flat pins, either grounded or ungrounded, while Type C (Europlug), with two round pins, is widespread across Europe and is often compatible with multiple socket types. Types D, M, and N are found mainly in India, South Africa, and Brazil, whereas Types E and F (Schuko) dominate continental Europe, featuring round pins and grounding mechanisms. Type G, used in the United Kingdom and several other regions, is distinguished by three rectangular pins and a built-in fuse, while Types H, I, J, K, and L are more region-specific, occurring in Israel, Australia, Switzerland, Denmark, and Italy, respectively [28]. More recently, Type O was introduced in Thailand; however, as it is not yet widely adopted, it is not included in this study. This diversity in socket types highlights the importance of accurate classification to enable reliable geolocation based on plug design.

In this experiment, CV is used to classify socket types, a task that can substantially narrow the global geographical search space and thereby assist law enforcement in criminal investigations. CNNs are well suited to this application, as they excel at extracting visual features for image classification [12]. Since the focus is exclusively on visual characteristics, the study considers 12 classes instead of the full set of 14. Types D and M have been merged into a single class (DM) because, despite differences in pin size, their layouts are visually indistinguishable. Type M closely resembles Type D but features larger pins. Similarly, Types J and N are merged into a single class (JN). Both sockets use three round pins with nearly identical configurations, differing only in the precise offset of the earth pin – with Type J, it is offset by 5mm, and with Type N,it is offset by 3mm [18]. This small structural difference makes them electrically incompatible, but the variation was deemed too subtle to be reliably distinguished visually. All other socket types present clear visual differences and are, therefore, treated as separate classes. In addition, a noise class was introduced to exclude non-socket objects, such as light switches, thermostats, or low-quality regions of interest detected by YOLO, thereby further enhancing accuracy.

5.1. Dataset Preparation

The dataset for socket type classification was constructed from two primary sources. The first source was publicly available datasets from Roboflow, consisting of plug socket images licenced under CC 4.0. These images were downloaded, cleaned, and merged into a consistent collection. The second source was the Hotels-50K dataset, from which socket regions were automatically detected, cropped, and assigned to their respective classes. After processing, the final dataset comprised 12 socket type classes, with the number of images per class summarised in Table 3. In total, the dataset contains 3,187 images, which were partitioned into training, validation, and test sets using a 70:15:15 split.

Table 3: Socket Types with Image Counts and Corresponding Country Usage Count.

Socket	Images	Countries
A	192	46
B	302	28
C	305	65
D / M	300	12 / 9
E	304	24
F	303	35
G	304	32
H	138	1
I	268	11
J / N	262	5 / 4
K	291	6
L	222	9
Total	3,187	-

5.2. Methodology

A transfer learning approach for multi-class image classification using five state-of-the-art CNN architectures was implemented: VGG16, InceptionV3, Xception, ResNet50, and ResNet101. The dataset, consisting of 3,187 images across 12 classes (socket types), was split into training, validation, and test sets, with 2,224 images for training, 473 images for validation, and 490 images for testing. Images were preprocessed and augmented with normalisation and horizontal flipping to improve generalisation. Each model was initialised with ImageNet pre-trained weights, and the convolutional base was frozen to leverage pre-learned features while training a new classification head for the target classes. Models were trained independently using categorical cross-entropy loss and the Adam optimiser.

(a) (b)

Figure 3: YOLOv11m detection results on room images (a–b), showing bounding boxes for socket classes

(a) (b)

Figure 4: YOLOv11m detection results on bathroom images (a–b), showing bounding boxes for socket classes

5.3. *Evaluation and Comparative Analysis*

The performance of the model was evaluated using standard metrics, including accuracy, precision, recall, F1-score, and a confusion matrix to provide detailed insights into class-wise predictions. Among the evaluated models, VGG16, InceptionV3, and Xception achieved the highest accuracies. Xception attained the best overall performance with an accuracy of 91.22%, consistently demonstrating high precision, recall, and F1-scores across all 12 classes, as summarised in Table 4. In comparison, VGG16 achieved 82.65% accuracy, while InceptionV3 reached 89.80%, highlighting the performance differences among the individual models.

To further validate the robustness of Xception, the impact of K-fold cross-validation was evaluated by modifying the dataset. A new Noise class, consisting of 304 images representing non-socket objects potentially missed by Step 1 YOLO, was added to the original 3,187 images, resulting in a total of 3,495 images across 13 classes (12 socket types + Noise). These images were organised into 13 folders for training and validation in a 5-fold cross-validation setup, while an additional 175 unseen images, spanning all classes, were reserved for testing. This setup allowed us to assess whether cross-validation improves model generalisation.

Table 4: Performance Summary of Different Models

Model	Accuracy	Precision	Recall	F1-score
VGG16	0.827	0.846	0.816	0.819
InceptionV3	0.898	0.907	0.900	0.901
Xception	**0.912**	**0.914**	**0.910**	**0.911**
ResNet50	0.492	0.599	0.477	0.466
ResNet101	0.443	0.634	0.429	0.433

As summarised in Table 5, the 5-fold cross-validation strategy increased accuracy from 85.4% to 87.7%, precision from 87.8% to 89.4%, recall from 85.3% to 88.4%, and F1-score from 85.5% to 88.1%. These results demonstrate that K-fold cross-validation provided a modest but consistent improvement in performance over the single-run Xception model. Note that the test dataset differs from the split of 70:15:15.

Table 5: Performance Summary of Xception Models

Model	Accuracy	Precision	Recall	F1
Xception	0.854	0.878	0.853	0.855
Xception (5-Fold CV)	**0.877**	**0.894**	**0.884**	**0.881**

Therefore, Xception with 5 cross-validation was identified as the best performer in stage two socket type classification.

6. Stage 3: Inferring Geolocation through Socket Type–Country Mapping

The final stage of the proposed pipeline focusses on using socket detection results to infer the geolocation of hotel rooms, highlighting the practical applicability of the proposed approach.

6.1. *Test Dataset Preparation*

The test dataset for this experiment was derived from the Hotels-50K TraffickCam dataset, available from its official

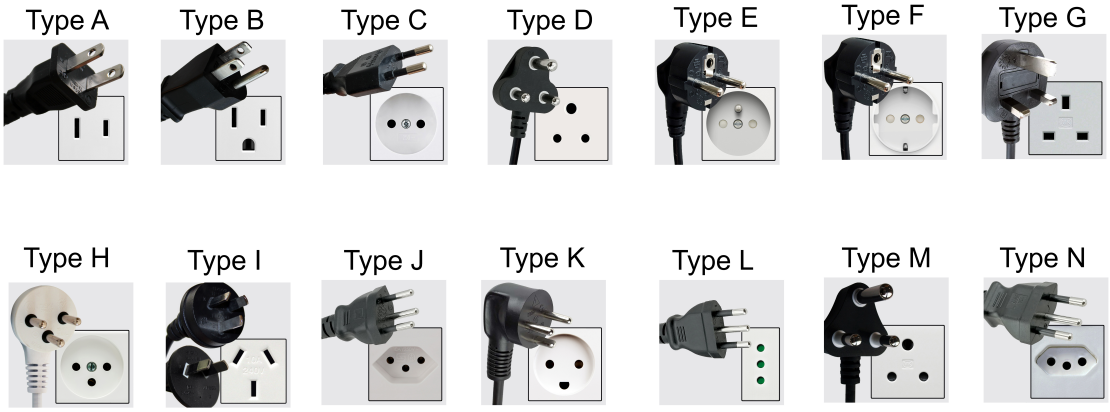


Figure 5: Plug and Socket Types from Type A to Type N [28]

GitHub repository. To prepare the data, the dataset was obtained by modifying the download script to preserve the original image resolution. This was a crucial step because the target object, electric sockets, is often small within the larger image, and resizing could lead to a loss of detail essential for accurate detection.

The Hotels-50K dataset comprises two subcategories: TraffickCam and travel website images. TraffickCam consists of crowd-sourced photos submitted by travellers worldwide, reflecting real-world conditions. After retaining and downloading the images in their original resolution of 1024×768, the dataset provides images in this resolution, with sockets typically appearing in small regions of approximately 130×87 pixels.

In contrast, travel website images are professionally captured under ideal lighting and angles, often with colour correction and photo editing. These images are also provided at a lower resolution (350×233), making the sockets barely visible. As a result, they were excluded from testing. The TraffickCam subset, by depicting more realistic, non-professionally taken and/or edited photos, is considered to provide a better representation of conditions encountered in practical investigations and was therefore used exclusively for evaluation.

To establish a ground truth for the experiment, the dataset was processed to associate each image with its corresponding country. This involved a multistep process: Merging Metadata: the various CSV files from the original dataset were consolidated to create a unified file containing image IDs, hotel IDs, and geographic coordinates (latitude and longitude). Geolocation: Using the `geopy.geocoders` library, the geographic coordinates were converted into country names. Standardisation: To ensure consistency, the country names were standardised using the `pycountry` library, as the raw geolocation output sometimes returned names in native languages. Illegal characters were removed, and spaces were replaced with underscores to create valid directory names.

The final dataset was restructured into a clean directory, with images organised into subfolders named after their respective countries. This structure, along with a consolidated CSV file containing all relevant metadata, streamlined the subsequent country-specific analysis. The reason images were arranged in folders instead of being directly taken and compared from the CSV file was to facilitate visual inspection, ensuring that the code functioned correctly and that each country was accurately represented with its respective socket type.

6.2. *Data Analysis, Evaluation, and Results*

The model's performance was evaluated by assessing its ability to detect a plug socket in an image and then classify it to determine its corresponding country. The process involved two main tasks for each image in the test dataset: socket detection, which checks for the presence of a plug socket, and socket classification, which identifies its type. A predefined mapping associates specific plug types with the countries in which they are most common. This mapping serves as the ground truth. For each image in which a socket was successfully detected and classified, it was checked whether the predicted plug type and the image's actual country formed a valid pair according to this mapping. A correct match was assigned a score of 1, an incorrect match received a score of -1, and the noise class was assigned a neutral score of 0. These scores were used to calculate key performance metrics, such as the confusion matrix, precision, recall, and F1 score. This approach provided a comprehensive assessment of the model's accuracy. Finally, the results were compiled into a detailed summary report. They were presented with graphical visualisations, such as bar charts, to provide a clear and intuitive interpretation of the model's overall performance.

A total of 44,630 TraffickCam images were processed through the algorithmic pipeline. In the first stage, YOLO detected 3,759 potential sockets. To enhance detection accuracy and eliminate false positives, a second-stage classifier was employed to identify and remove noise. Specifically, instances where non-socket objects (e.g., switchboards) were incorrectly detected as sockets in the first stage were classified as noise. This step identified 1,393 noisy detections, leaving 2,366 valid sockets. These valid detections were subsequently passed to the socket classification stage, where only those with a confidence threshold above 70% were retained. The results are summarised in Table 6.

Table 6: Classifier Threshold Analysis for Socket Detection

Class Confidence	Correct	Wrong	Total	Accuracy (%)
≥ 70%	1595	146	1741	91.61
≥ 80%	1421	95	1516	93.73
≥ 90%	1167	45	1212	96.29

When considering different socket classification confidence thresholds, the performance varies. Without setting any socket confidence threshold, 1,967 predictions were correct, and 399 were incorrect, resulting in an accuracy of 83.08%. At a threshold above 70%, 1,595 predictions were correct, and 146 were incorrect, resulting in an accuracy of 91.61%. Increasing the threshold above 80% slightly reduced the number of correct detections to 1,421, while incorrect detections decreased to 95, yielding an improved accuracy of 93.73%. At the highest threshold of above 90%, correct detections further decreased to 1,167, with only 45 incorrect predictions, resulting in the highest accuracy of 96.29%. These results illustrate the trade-off between confidence and accuracy: lower thresholds capture more sockets but result in more false positives, whereas higher thresholds reduce errors at the cost of missing some detections, as shown in Fig. 6.

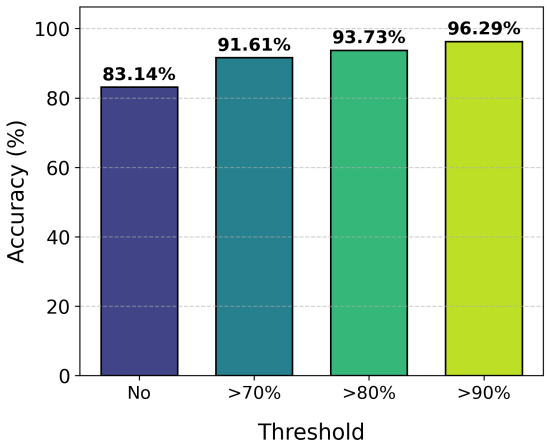
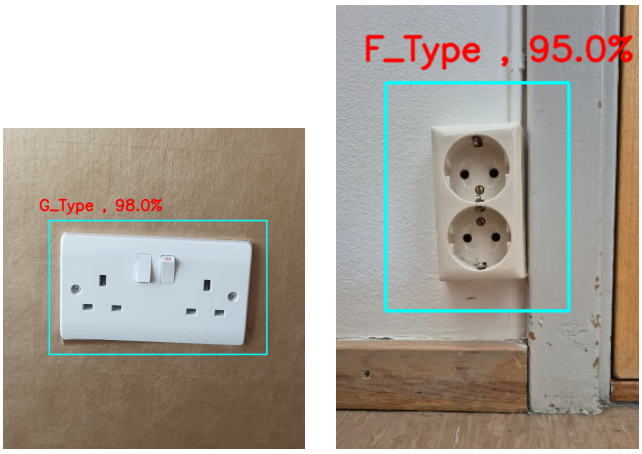


Figure 6: Country prediction accuracy by varying socket type confidence threshold values

7. Discussion

The proposed pipeline is designed to be generalisable, capable of detecting sockets in images beyond the Hotels-50K dataset. The prototype was tested on images captured by a camera, and the algorithm produced accurate detections, as shown in Fig. 7. This demonstrates that the model is not limited to the Hotels-50K dataset; it can detect sockets in any image, making the application practical and universally applicable for geolocation purposes.

Socket detection is inherently challenging due to the small size and frequent low-resolution imagery. Despite this, the proposed method performs strongly when sockets are present, with geolocation inference achieving 96.29% accuracy. Future



(a) (b)

Figure 7: Visual results of proposed pipeline images captured outside the Hotels-50K dataset

work will focus on enhancing the robustness of the geolocation pipeline to better handle cases in which sockets are not visible or are occluded. Of course, not every indoor photo contains a detectable socket. Future research includes the exploration of complementary approaches, such as image similarity analysis. By generating perceptual hashes for images, visually similar scenes can be grouped together, even in the absence of sockets. This enables indirect geolocation inference through contextual or visual correlation with previously classified images. This clustering could also help identify patterns across large datasets and link unknown or unlabelled images to geographically known locations. Furthermore, integrating multimodal cues, such as architectural styles, interior designs, or visible textual information, could further strengthen geolocation inference when electrical fixtures are unavailable. Incorporating these techniques within a unified framework would make the system more scalable and applicable to real-world forensic and investigative applications. Ultimately, this could bridge the gap between AI-based contextual geolocation and digital investigations.

8. Conclusion

This paper presents a universal pipeline for socket detection and classification, motivated by its potential to enhance geolocation capabilities in investigative contexts where conventional cues, such as metadata, outdoor landmarks, and sensor data, are unavailable. By curating two new datasets, benchmarking multiple detection and classification models, and evaluating performance on the Hotels-50K TraffickCam dataset, this study demonstrates both the feasibility and the challenges associated with socket-based indoor geolocation. Despite challenges such as small object size and low-resolution imagery, the results demonstrate strong detection accuracy and validate the concept's practical potential, enabling investigators to automate geolocation across large volumes of digital evidence and convert it into actionable intelligence.

Beyond technical performance, this work contributes to the emerging field of AI-driven multimedia forensics, where visual scene elements are utilised to support digital investigations. Socket detection offers a unique, region-specific forensic cue that can aid law enforcement agencies in narrowing search regions, corroborating other forms of evidence, and identifying the possible origins of illicit or trafficking-related imagery. The proposed framework, therefore, represents an important step towards scalable, context-aware forensic tools that bridge the gap between CV and real-world investigative practice.

Future work will focus on integrating this approach with perceptual hashing and multimodal analysis techniques to cluster visually similar scenes and strengthen geolocation inference when sockets are not visible. By combining socket-based detection with broader contextual cues, the framework could evolve into a robust and general-purpose forensic system capable of supporting a wide range of investigative and humanitarian applications.

References

[1] Abdullah, M.N., Shukran, M.A.M., Isa, M.R.M., Ahmad, N.S.M., Khairuddin, M.A., Yunus, M.S.F.M., Ahmad, F., 2021. Colour Features Extraction Techniques and Approaches for Content-Based Image Retrieval (CBIR) System. Journal of Materials Science and Chemical Engineering 9, 29–34. doi:10.4236/msce.2021.97003.

[2] Bales, K., Hesketh, O., Silverman, B., 2015. Modern slavery in the UK: How many victims? Significance 12, 16–21. URL: https://nottingham-repository.worktribe.com/output/5091819, doi:10.1111/j.1740-9713.2015.00824.x.

[3] Bamigbade, O., Scanlon, M., Sheppard, J., 2025. Improving image embeddings with colour features in indoor scene geolocation. IEEE Access 13, 79860–79870. doi:10.1109/ACCESS.2025.3564496.

[4] Bamigbade, O., Sheppard, J., Scanlon, M., 2024. Computer Vision for Multimedia Geolocation in Human Trafficking Investigation: A Systematic Literature Review. URL: https://arxiv.org/abs/2402.15448, arXiv:2402.15448.

[5] Bhavanasi, S.S., Stylianou, A., 2023. Hotel Recognition Using Object Ensembles, in: 2023 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), pp. 1–8. doi:10.1109/AIPR60534.2023.10440661.

[6] Black, S., Stylianou, A., Pless, R., Souvenir, R., 2022. Visualizing paired image similarity in transformer networks, in: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1534–1543. doi:10.1109/WACV51458.2022.00160.

[7] Breitinger, F., Hilgert, J.N., Hargreaves, C., Sheppard, J., Overdorf, R., Scanlon, M., 2024. DFRWS EU 10-year review and future directions in Digital Forensic Research. Forensic Science International: Digital Investigation 48, 301685. URL: https://www.sciencedirect.com/science/article/pii/S2666281723002044, doi:https://doi.org/10.1016/j.fsidi.2023.301685. DFRWS EU 2024 - Selected Papers from the 11th Annual Digital Forensics Research Conference Europe.

[8] Brejcha, J., Čadík, M., 2017. State-of-the-art in visual geo-localization. Pattern Analysis and Applications 20, 613–637. doi:10.1007/s10044-017-0611-1.

[9] Cao, B., Araujo, A., Sim, J., 2020. Unifying deep local and global features for image search, in: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX, Springer-Verlag, Berlin, Heidelberg. p. 726–743. doi:10.1007/978-3-030-58565-5_43.

[10] Demetriou, D., Mavromatidis, P., Robert, P.M., Papadopoulos, H., Petrou, M.F., Nicolaides, D., 2023. Real-time construction demolition waste detection using state-of-the-art deep learning methods; single–stage vs two-stage detectors. Waste Management 167, 194–203. URL: https://www.sciencedirect.com/science/article/pii/S0956053X23003872, doi:https://doi.org/10.1016/j.wasman.2023.05.039.

[11] Dimas, G.L., Konrad, R.A., Maass, K.L., Trapp, A.C., 2022. Operations research and analytics to combat human trafficking: A systematic review of academic literature. PLOS ONE 17, 1–24. URL: https://EconPapers.repec.org/RePEc:plo:pone00:0273708.

[12] Elngar, A.A., Arafa, M., Fathy, A., Moustafa, B., Mahmoud, O., Shaban, M., Fawzy, N., 2021. Image Classification Based On CNN: A Survey. Journal of Cybersecurity and Information Management 6, 18–50. doi:10.5281/zenodo.4897990.

[13] Faqir, R.S.A., 2023. Digital Criminal Investigations in the Era of Artificial Intelligence: A Comprehensive Overview. International Journal of Cyber Criminology 17, 77–94. doi:10.5281/zenodo.4766706.

[14] Gstrein, O., Haleem, N., Zwitter, A., 2024. General-purpose AI regulation and the European Union AI Act. Internet Policy Review 13, 1–26. doi:10.14763/2024.3.1790.

[15] Hadid, M., Hussein, Q.M., Al-Qaysi, Z., Ahmed, M., Salih, M.M., 2023. An Overview of Content-Based Image Retrieval Methods And Techniques. Iraqi Journal For Computer Science and Mathematics 4, 6. doi:10.52866/ijcsm.2023.02.03.006.

[16] Hargreaves, C., Breitinger, F., Dowthwaite, L., Webb, H., Scanlon, M., 2024. DFPulse: The 2024 digital

forensic practitioner survey. Forensic Science International: Digital Investigation 51, 301844. URL: https://www.sciencedirect.com/science/article/pii/S2666281724001719, doi:https://doi.org/10.1016/j.fsidi.2024.301844.

[17] Herrmann, J., Bamigbade, O., Sheppard, J., Scanlon, M., 2024. Perceptual Colour-based Geolocation of Human Trafficking Images for Digital Forensic Investigation, in: 2024 Cyber Research Conference - Ireland (Cyber-RCI), IEEE. pp. 1–8. doi:10.1109/Cyber-RCI60769.2024.10941203.

[18] International Electrotechnical Commission, 2015. Plugs and socket-outlets for domestic and similar general use standardized in member countries of IEC. URL: https://webstore.iec.ch/publication/23628. Technical Report.

[19] Jegham, N., Koh, C.Y., Abdelatti, M., Hendawi, A., 2025. YOLO Evolution: A Comprehensive Benchmark and Architectural Review of YOLOv12, YOLO11, and Their Previous Versions. URL: https://arxiv.org/abs/2411.00201, arXiv:2411.00201.

[20] Ji, R., Gao, Y., Liu, W., Xie, X., Tian, Q., Li, X., 2015. When location meets social multimedia: A survey on vision-based recognition and mining for geo-social multimedia analytics. ACM Transactions on Intelligent Systems and Technology 6. doi:10.1145/2597181.

[21] Jiang, T., Zhong, Y., 2025. Odverse33: Is the new YOLO version always better? A multi domain benchmark from YOLO v5 to v11. CoRR abs/2502.14314. doi:10.48550/ARXIV.2502.14314, arXiv:2502.14314.

[22] Joshi, S., Jain, A., Payani, A., Mirzasoleiman, B., 2024. Data-efficient contrastive language-image pretraining: Prioritizing data quality over quantity, in: Dasgupta, S., Mandt, S., Li, Y. (Eds.), Proceedings of The 27th International Conference on Artificial Intelligence and Statistics, PMLR. pp. 1000–1008. URL: https://proceedings.mlr.press/v238/joshi24a.html.

[23] Kamath, R., Rolwes, G., Black, S., Stylianou, A., 2021. The 2021 Hotel-ID to Combat Human Trafficking Competition Dataset. URL: https://arxiv.org/abs/2106.05746.

[24] Kim, S., Park, S., Kim, M., 2003. Central object extraction for object-based image retrieval, in: Proceedings of the 2nd International Conference on Image and Video Retrieval, Springer-Verlag, Berlin, Heidelberg. p. 39–49.

[25] Luo, J., Joshi, D., Yu, J., Gallagher, A., 2011. Geotagging in multimedia and computer vision—a survey. Multimedia Tools and Applications 51, 187–211. URL: https://doi.org/10.1007/s11042-010-0623-y, doi:10.1007/s11042-010-0623-y.

[26] Manovich, L., 2018. 100 Billion Data Rows per Second: Media Analytics in the Early 21st Century. International Journal of Communication 12, 473–488. URL: http://ijoc.org.

[27] Mayo, D.I., 2019. Understanding object recognition performance at scale in machines and humans. Ph.D. thesis. Massachusetts Institute of Technology.

[28] McGregor, C.H., 2025. Plug & socket types. URL: https://www.worldstandards.eu/electricity/plugs-and-sockets/. https://www.worldstandards.eu/electricity/plugs-and-sockets/ Accessed: 2025-09-18.

[29] Moore, D.M., 2024. Algorithmic Exploitation in Social Media Human Trafficking and Strategies for Regulation. Laws 13. doi:10.3390/laws13030031.

[30] Pradhan, J., Pal, A.K., Banka, H., 2016. A prominent object region detection based approach for CBIR application, in: 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), pp. 447–452. doi:10.1109/PDGC.2016.7913237.

[31] Roboflow, 2025. Roboflow: Computer Vision Tools for Developers and Enterprises. URL: https://roboflow.com/. accessed: 2025-09-18.

[32] Sangeetha, S.K.B., Mathivanan, S.K., Pandi, T., Arivu selvan, K., Jayagopal, P., Teshite Dalu, G., 2022. An Enhanced Triadic Color Scheme for Content-Based Image Retrieval. Mathematical Problems in Engineering 2022, 5736630. URL: https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/5736630, doi:https://doi.org/10.1155/2022/5736630.

[33] Santos, C.F.G.D., Papa, J.a.P., 2022. Avoiding overfitting: A survey on regularization methods for convolutional neural networks. ACM Computing Surveys 54. URL: https://doi.org/10.1145/3510413, doi:10.1145/3510413.

[34] Shamoi, P., Sansyzbayev, D., Abiley, N., 2022. Comparative Overview of Color Models for Content-Based Image Retrieval, in: 2022 International Conference on Smart Information Systems and Technologies (SIST), pp. 1–6. doi:10.1109/SIST54437.2022.9945709.

[35] Stylianou, A., Xuan, H., Shende, M., Brandt, J., Souvenir, R., Pless, R., 2019. Hotels-50K: A Global Hotel Recognition Dataset, in: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI Press. pp. 726–733. URL: https://doi.org/10.1609/aaai.v33i01.3301726, doi:10.1609/aaai.v33i01.3301726.

[36] Tian, Z., Shen, C., Chen, H., He, T., 2019. Fcos: Fully convolutional one-stage object detection, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9626–9635. doi:10.1109/ICCV.2019.00972.

[37] Tseytlin, B., Makarov, I., 2021. Hotel Recognition via Latent Image Embeddings, in: Rojas, I., Joya, G., Català, A. (Eds.), Advances in Computational Intelligence, Springer International Publishing, Cham. pp. 293–305. doi:10.1007/978-3-030-85099-9_24.

[38] United Nations Office on Drugs and Crime, 2025. Human Trafficking and the SDGs. https://www.unodc.org/unodc/human-trafficking/sdgs.html.

[39] Wadhai, S.A., Kawathekar, S.S., 2019. Techniques of Content Based Image Retrieval: A Review. IOSR Journal of Computer Engineering (IOSR-JCE) , 75–79.

[40] Walby, S., Francis, B., 2025. Improving the Estimate of Trafficking in Human Beings and Modern Slavery by Integrating Data From ILO/Walk Free/IOM and UNODC. Social Indicators Research 176, 669–693. doi:10.1007/s11205-024-03474-w.

[41] Wang, X., Li, H., Yue, X., Meng, L., 2023. A comprehensive survey on object detection YOLO, in: The 5th International Symposium on Advanced Technologies and Applications in the Internet of Things (ATAIT 2023), pp. 77–89.

[42] Wazzan, A., Ahmad, I., Macneil, S., Souvenir, R., 2024. Context or Clutter? Efficiently Matching Objects Across Scenes, in: Proceedings of the 2024 International Conference on Multimedia Retrieval, Association for Computing Machinery, New York, NY, USA. p. 404–413. doi:10.1145/3652583.3658090.

[43] Zhang, X., Wang, L., Su, Y., 2021. Visual place recognition: A survey from deep learning perspective. Pattern Recognition 113, 107760. URL: https://www.sciencedirect.com/science/article/pii/S003132032030563X, doi:https://doi.org/10.1016/j.patcog.2020.107760.