





Learning Strategies for Sensitive Content Detection

Daniel Povedano Álvarez [†], Ana Lucila Sandoval Orozco [†], Javier Portela García-Miguel [†]
and Luis Javier García Villalba ^{*,†}

Group of Analysis, Security and Systems (GASS), Department of Software Engineering and Artificial Intelligence (DISIA), Faculty of Computer Science and Engineering, Office 431, Universidad Complutense de Madrid (UCM), Calle Profesor José García Santesmases, 9, Ciudad Universitaria, 28040 Madrid, Spain

* Correspondence: javiergv@fdi.ucm.es; Tel.: +34-91-394-7638

† These authors contributed equally to this work.

Abstract: Currently, the volume of sensitive content on the Internet, such as pornography and child pornography, and the amount of time that people spend online (especially children) have led to an increase in the distribution of such content (e.g., images of children being sexually abused, real-time videos of such abuse, grooming activities, etc.). It is therefore essential to have effective IT tools that automate the detection and blocking of this type of material, as manual filtering of huge volumes of data is practically impossible. The goal of this study is to carry out a comprehensive review of different learning strategies for the detection of sensitive content available in the literature, from the most conventional techniques to the most cutting-edge deep learning algorithms, highlighting the strengths and weaknesses of each, as well as the datasets used. The performance and scalability of the different strategies proposed in this work depend on the heterogeneity of the dataset, the feature extraction techniques (hashes, visual, audio, etc.) and the learning algorithms. Finally, new lines of research in sensitive-content detection are presented.

Keywords: deep learning; digital forensics; image recognition; sensitive content; sexually explicit content detection; video classification



Citation: Povedano Álvarez, D.; Sandoval Orozco, A.L.; García-Miguel, J.P.; García Villalba, L.J. Learning Strategies for Sensitive Content Detection. *Electronics* **2023**, *12*, 2496. <https://doi.org/10.3390/electronics12112496>

Academic Editor: Zhenhua Guo

Received: 30 March 2023

Revised: 25 May 2023

Accepted: 29 May 2023

Published: 1 June 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In recent years, the growth of multimedia devices (in particular mobile devices and tablets) and the sharing of multimedia content on social networks has led to an increase in illegal content on the Internet. In this context, the proliferation of sensitive Internet content is growing by the day, which means that large amounts of illegal and pornographic material are accessible to all users. The detection of adult sites and material has an important place in the prevention of sexual activities and pornography.

With the rapid development of the Internet, illegal and pornographic images/videos can spread more easily and affect the mental health of adolescents. Our dependence on the Internet to facilitate our personal and professional lives has increased especially during the COVID-19 pandemic. In [1], they study the long-term effects of COVID-19 and its socio-economic implications on children in distress, highlighting the growth of psychosocial risks to which they are exposed, during and after the COVID-19 crisis and their blockade situation. In particular, they link the blockade and the resulting economic problems with risks related to sexuality and abuse, such as in situations of child work and child trafficking, and child sexual abuse, which cause short- and long-term mental health effects on children.

In this sense, the same technology that offers us opportunities for growth and connectivity can also pose significant risks to children. Child pornography (and all its forms such as sexting, grooming, cyberbullying, etc.) has become a matter of concern because this market and the confinement of 2020 could result in an increase in more child abuse material and could significantly influence children's physical and mental health during childhood and adulthood, as well as undermine their dignity by depicting them as sexual objects.

In this regard, specific automated applications that fight against child sexual abuse (CSA) and contribute to the identification of perpetrators and victims have garnered considerable interest in law enforcement and forensic activities. In its annual report [2], Europol has provided a review of emerging dangers and critical effects in the field of cyber-crime over the past year, highlighting online child exploitation as an urgent child protection concern. Such specific automatic applications are primarily aimed at detecting sensitive material for adults, able to block uploading or access to sensitive content to certain individuals (e.g., minors) or places (public places, children's schools, etc.), which reduce the workload of authorities, as well as the on-site analysis of seized sensitive material, which could result in the quick apprehension of potential offenders.

The automatic classification of child sexual abuse material (CSAM) is a complicated field of research, as it is illegal to possess such material and should only be accessible to law enforcement agencies. For the detection of this type of material, it is first necessary to carry out automatic detection of sensitive content or adult pornography. Thus, after the detection process, other strategies can be applied for the detection of CSA, such as facial and age recognition, image and video forgery detection, camera model identification, object and background identification, etc., which could help with the automatic filtering of images and videos and their gathering during the course of an investigation.

The aim of this study is to comprehensively review the strategies proposed in the literature for the detection of sensitive content in images and videos in order to identify research gaps and open issues. The rest of the work is organised as follows: Section 2 highlights the differences with other surveys and the key contributions of this work. Section 3 presents an overview of the main strategies for classifying sexually sensitive content. Section 4 details the approaches based on text features such as hashes and keywords. Section 5 details the approaches based on key-frame visual features and describes the strategies based on the image descriptor. The details of motion, audio and multimodal analysis approaches that use a combination of visual and audio features are described in Section 6. Section 7 details the state-of-the-art deep learning techniques for sensitive-content detection. Section 8 discusses the results of the experiments undertaken, and finally, the conclusions of the research are included in Section 9.

2. Difference with Other Surveys

In the literature, to the best of our knowledge, there are only a few studies regarding sensitive-content detection. A detailed review and assessment of the achievements and challenges of CSAM detection research are presented in [3], focusing on political and legal aspects, distribution channels and applications for the detection of this type of material. Regarding applications for sensitive-content detection, the authors focus on four types of detection: and image hash database, webcrawler, detection based on filename and metadata and visual detection. This paper highlights the improvements of deep learning techniques over conventional work but only covers the period up to 2018, without taking into account work using modern convolutional neural networks (CNNs) or the advent of vision transformers (ViTs). In addition, the review of applications for CSAM detection is limited to visual and audio detection methods without reviewing other more conventional methods (such as the use of image descriptors).

Pour et al. [4] carried out a comprehensive survey to examine a VCR system (automatic video content rating). This system classifies a video according to the age group of the audience. Based on current manual rating systems, the VCR system is based on five principles: violence, foul language, nudity, pornography and substance abuse. To this end, they reviewed DL-related works with the relevant VCR themes mentioned above. In this sense, they investigated works based on audio, static and motion visual aspects. Moreover, the authors reviewed the different datasets related to violence and pornography. Other sensitive content-detection methods based on textual features such as hashes, metadata or file names were not taken into account in this work. In addition, the reviewed sensitive content-detection methods are related to the latest developments in ML and DL, leaving

aside other classical techniques such as the segmentation of skin colour regions to characterise nudity. Finally, the reviewed works that used DL techniques for sensitive-content detection go up to 2019, without taking into account the latest developments in computer vision, such as vision transformers.

Cifuentes et al. [5] conducted a comprehensive review of existing sexually explicit video techniques up to the end of 2019 and early 2020. In this study, they analyse strategies mainly focused on works on detecting sensitive material in videos (as well as a few works using image datasets) and work using deep learning (DL) strategies, without considering the latest developments in this field such as the visual attention mechanism. Although the authors perform an exhaustive analysis and discussion of the main works on visual features, they do not consider other techniques such as hashes or filenames or works that use only audio as input features.

Studies published thus far in high-level journals have focused either on visual feature detection strategies, but almost without taking into account auditory features or the latest advances in deep learning such as ViTs (emerging architectures since 2020), or on textual features and commercial tools, but they have not performed a comprehensive comparison between all strategies (commercial and non-commercial implementations, based on textual and visual features, etc.) for the detection of sensitive material. Therefore, this study covers all these gaps comprehensively.

Contribution of This Work

A comprehensive study of learning strategies for the detection of sensitive content is presented with the following highlights.

- The few papers published to date in the field of sensitive-content detection show the overall picture of the research contribution in this field.
- To our awareness, this is the first comprehensive systematic study that brings together valuable research contributions in this field, bringing together content-based strategies on video/image (visual and auditory) and textual (hashes and keywords) features.
- This study is classified according to the methodologies proposed to facilitate comparison between them and the selection of the best ones.
- This review will be useful for new researchers to identify the issues and challenges that the community is addressing in this field. In addition, gaps are discussed that will help future researchers to identify and explore new directions in the field of sensitive-content detection.

3. Types of Strategies to Detect Sexually Sensitive Content Classification

To address the problems described above, different solutions have been developed. As for methods using textual features, they can be further divided into three groups [3]: image hash database, webcrawler, and filename detection. Among the hash-based methods, the best known is Microsoft's PhotoDNA [6], a technology that helps to identify and remove well-known child exploitation images. PhotoDNA generates a unique digital signature of an image that is then compared with hashes of other photos to identify copies of the identical image. PhotoDNA helps to detect, block and notice the dissemination of child exploitation material by comparing the hash of a suspicious image with a database that stores hashes of earlier reported illegal images. However, this tool is not able to detect new CSA content, as it is used to compare with previously existing hashes in a database.

Among webcrawlers, Project Arachnid is worth mentioning. The aim of this project is to crawl web pages in order to enter the contents into databases for verification and indexing needs, as well as to use PhotoDNA technology for the creation of hashes of the collected content. For CSAM detection, the Arachnid [7] project uses the hash lists of several organisations such as The National Center for Missing & Exploited Children (NCMEC), the Royal Canadian Mounted Police (RCMP) and Interpol. On the other hand, detection based on file names and metadata is performed by obtaining information from well-known CSA and non-CSA files. To do this, metadata features are extracted from the files by segmenting

and normalising the text. Approaches using textual features use conventional ML classifiers such as support vector machine (SVM) or logistic regression to separate normal files from those with sensitive content.

In addition, there are commercial tools [8,9] to regulate access to this type of content and specific tools [10,11] to control and mitigate risks by limiting access to CSAM. Some commercial tools provide a content-filtering approach according to whitelists and blacklists based on metadata information. Nevertheless, they are inefficient, as sensible content can be intentionally added to a simple and irrelevant text, rendering these linked text labels insufficient for analysis.

Consequently, the last category (visual and audio information analysis) is essential for the accurate classification of pornographic content.

In this sense, after reviewing the literature on explicit content detection in videos, the main strategies reviewed from the state of the art can be classified into four categories: text analysis, visual detection, motion, audio and multimodal analysis and DL techniques (Figure 1).

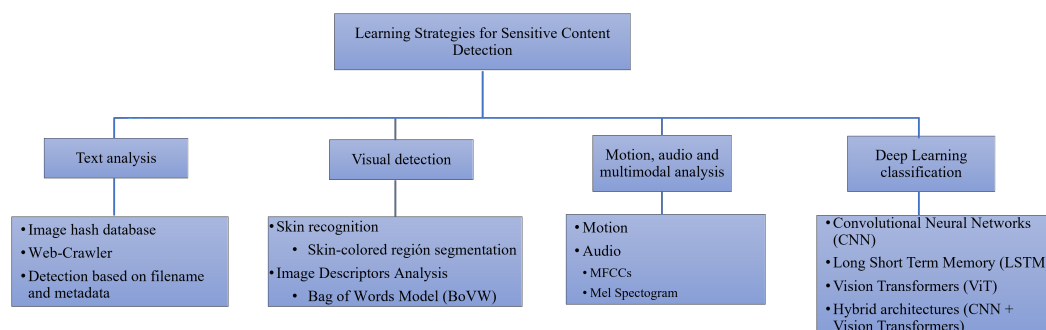


Figure 1. General classification of strategies to automatically detect sexually sensitive content and explicit videos.

Early approaches in sensitive-content detection primarily incorporate the segmentation of skin colour regions to characterise nudity [12–19]. In general, these solutions take such input to identify the set of pixels and spatial distributions describing the naked persons. However, recognising bare skin per se is not a reliable criterion, as large areas of skin are not explicitly pornographic (e.g., swimming costumes and sporting activities such as sumo or boxing), which could lead to many false positives. Furthermore, it is not possible to analyse greyscale images using this method. On the other hand, alternative strategies focus on image descriptors to achieve the identification task. In the case of local descriptors, multiple features are extracted using texture to wrap each region and to characterise the face, and then, a score level (SL) is used to integrate the information obtained from the different descriptors.

In these techniques, local features are extracted and subsequently quantised using a codebook describing bag of visual words (BoVW) models [20–23]. While these techniques have shown very interesting achievements in this field of research, the local descriptors computed throughout the process of analysis are computationally intensive and result in high-dimensionality vectors, requiring the application of dimensionality reduction algorithms such as principal component analysis (PCA). In addition, these methods are especially susceptible to the choice of keywords, the size of the codebook, and the coding and clustering algorithms, indicating a large number of parameters to fit to achieve good performance.

Tian et al. used in [24] the colour feature to describe the local colour of the sex organs and to concatenate them with the histogram of oriented gradients to characterise the sex organs. From the concatenated feature of sex organs using the colour-saliency preserved mixture deformable part model (CPMDPM), they detected pornographic images sequentially with sex organ detectors.

Most of the literature tends to extend the solutions used for images to video, through the analysis of single frames and an application to sensitive samples. However, the spatiotemporal data in videos can show added features to enhance detection accuracy. Therefore, other strategies such as multimodal (vision features and recently audio features) pornography detection techniques have been implemented to find better effective solutions. Among these methods, there are [25–27].

The works [26,27] have shown that this multimodal analysis reaches greater accuracy. In [27], they proposed a late fusion process capable of combining several fragment classifiers to locate sensitive scenes, taking advantage of the multimodal nature of video data (e.g., motionless frames, audio stream, video space–time, etc.) to efficiently determine the frames of interest. To validate the solution, they conducted localisation experiments with pornographic and violent video streams, two of the most common types of sensitive content.

Although these strategies have a higher true positive rate and a lower false positive rate compared to other works, they continue to fail to detect certain situations, such as scenes with a clothed person performing sexual actions within static movements or moaning as masturbation, and there are certain multimodal features, such as audio, that raise the false positive rate, as audio detectors are based on low-level features, while the others are based on high-level features extracted through deep learning.

Moreover, not only the timbre element reflected by the spectrogram but also chroma, amplitude and other elements are necessary to accurately detect pornographic videos based on audio features alone. Therefore, this methodology has been used to detect pornography in images and videos based on recent successful DL solutions. Thus, the convolutional neural network (CNN) and the recurrent neural network (RNN) architectures (or combinations of the above) have been proposed in this domain [28]. In addition, in recent years, some classical CNN architectures such as AlexNet, VGG, ResNets and other RNN and long short-term memory (LSTM) have been extended by researchers [29–33], and more advanced ones, such as transformers, initially applied in natural language processing (NLP), have shown strong interest [33–35]. The outcomes associated with these works have shown that deep-learning-based architectures can improve the performance in detecting pornography in images and videos compared to the conventional strategies named above.

4. Strategies Based on Text Analysis

This section presents methods that do not use the content of images and videos themselves (e.g., visual and auditory features). These methods use textual features (strings) for sensitive-content detection. These textual features can be hashes, keywords, links, filenames and metadata. Strategies based on text analysis can be divided into three categories: image hash database, webcrawlers and filename and metadata detection.

4.1. Strategies Based on Image Hash Database

Hashing is a tool that allows us to know that a copy of digital information is the same as the original by transforming a series of variable-length digital input data into a fixed-length hexadecimal number, and it is generally used for data verification, password encryption and other highly sensitive data. Today, the main technology for detecting CSAM is image hashing [13,36]. Using this strategy, images earlier recognised as CSAM are compared with the unique hash value. This measure is obtained by means of a mathematical algorithm, resulting in shorter data of fixed length (24 bit hexadecimal code) serving to encrypt and authenticate the image content. The algorithm is designed to ensure that the identical input information generates identical output data every time. Since hash operations have a virtually limitless entry length and a pre-determined length of the output, different entries that are possible will yield the same output hash. This is called a hash collision. Such an event depends on the hash function used [11,36,37].

The best-known methods to automatically detect this type of material are cryptographic hash functions and image hashing algorithms. The best-known cryptographic hash functions are MD5 (message–digest algorithm 5), which produces a 128 bit hash, and SHA

(SecureHash algorithm), which produces a 160 bit hash value. On the other hand, among the best-known image hashing algorithms are aHash, dHash and Phash. Cryptographic hash algorithms, such as MD5 or SHA256, are designed to generate unpredictable results. To this end, they are optimised to change as much as possible with similar inputs. Perceptual hashes are the opposite: they are optimised to change as little as possible for similar inputs. The goal of perceptual hashing is to mimic the human visual system's evaluation of a comparison of two images based on the content of the underlying scene, as opposed to a purely numerical comparison based on pixel values [38].

This is achieved by extracting from the bulk representation of pixel space a concise, distinct and perceptually meaningful signature that resists image modifications, such as compression, colour changes, cropping, rotation, etc., or any other modification of the image that does not fundamentally change the underlying content but that alters pixel values. Depending on the content of the hash database, an image that has been previously identified can be identified more quickly if it reappears on the Internet.

Therefore, it is important that databases are regularly updated with new CSAM cases to increase the likelihood of finding the reappearance of such content elsewhere, such as on the darknet [36]. The use of image hash databases has proven effective to date and is common practice, especially for identifying known images in P2P networks [16].

In recent years and thanks to advances in DL, new algorithms have emerged that combine neural networks and perceptual hashes [39–43]. Perceptual hash functions can use techniques such as CNN to adaptively detect manipulation techniques and features. These techniques rely on deep neural networks to extract unique features from an image and then compute a hash value based on these features, yielding promising results given their ability to differentiate between substantially different images without being fooled by superficial changes [44].

In recent years, new tools have emerged from big technology companies such as Microsoft's PhotoDNA [6], Facebook's PDQ [45], Google (Content Safety API) [46] and Apple [47]. Apple recently announced its new tool called NeuralHash in 2021, a perceptual hashing algorithm for scanning content on devices used by its customers. This tool focuses on identifying CSAM content in user files uploaded to Apple's iCloud service. The tool works in two stages. First, an image is passed to a CNN to generate an N-dimensional feature vector. Second, the vector is passed through a hashing scheme to convert the N floating point numbers into M bits. As an advantage, NeuralHash achieves an excellent level of compression and retains enough information about the image so that the comparison and search between sets of images remain satisfactory. The neural network that generates the feature vector is trained using a self-supervised method. Finally, the images are modified with transformations that keep them perceptually identical to the original, creating an original/modified pair.

New approaches (to be reviewed in the following sections) propose purely metadata-based analysis (without examining the file contents) using filenames or file paths. Combining these techniques together with content hashing could increase the success of detecting new CSAM material. In practice, organisations can use a combination of these and other automated tools to detect CSAM on their network.

However, this approach has some constraints. Firstly, the database only supports searches for CSA images. In addition, other media formats such as videos should be incorporated into CSAM searches. Another weakness is the ineffectiveness of the database to find CSAMs that have not been identified as such in the past. In this sense, it is not possible to detect new CSA content using only image hash databases. Another disadvantage is that criminals could bypass these detection techniques by slightly manipulating the image or file name. These small modifications to the content, such as scaling, transcoding or renaming, could make these files undetectable with this approach [48].

Finally, Dr. NealKrawetz in [49] describes some weaknesses of PhotoDNA. He says that PhotoDNA does not detect flips, reflections, 90-degree rotations and inversions. However, it is supposed to detect visually similar images. Digitally transforming less than 2%

of the image at particular locations can effectively prevent detection. In addition, these edits can be applied to non-protruding regions of the image. Whoever wants to generate false positives only has to modify selective parts of the image. Forcing false positives can be used to justify plausible deniability to a tribunal.

A summary of the methodology of the work using hashes is outlined in Figure 2.



Figure 2. Pipeline of sensitive-content classification using hashes.

4.2. Strategies Based on Web-Crawlers

Web crawlers are developed to automatically scan websites and to gather information based on predefined search criteria. In general, web crawlers download content from crawled websites and enter and index it in a database [11]. New content is identified by visiting the hyperlinks contained in each crawled website. To this end, it is essential to identify the differences between websites that include CSAM and those that do not, in order to design appropriate features (e.g., keywords). In addition, keywords can also be used in the search for CSAM by adding them to web crawlers.

Steel et al. [50] identified and analysed CSAM-related content in 235,513 user queries and 194,444 hits. The study confirmed that a considerable proportion of peer-to-peer exchanges are CSAM. Nevertheless, it is not certain that users will use these keywords to find CSAM on the Internet. Searching for other words describing sexual acts or the age of children will also find CSAM, as they are often included in the title of such files.

The crawled websites are then examined to see whether or not they include the stated keywords. It is therefore crucial to understand how CSAM criminals name CSAM files they use on websites as identification for others seeking CSAM. The main benefit of using keywords is the possibility of finding material that has not been recognised before, unlike hashes. The main drawback of keywords could be too many false positives, including websites containing adult pornography, rather than illegal CSAM. To prevent the number of false positives, keywords should be chosen effectively. Another important disadvantage is the lack of a complete and robust image database, which limits its reliability. In addition, with the continuous expansion of the Internet, it is necessary to re-investigate the search criteria, as there could be a deviation between the criteria with which the crawler has been configured and the criteria at the time of use.

Westlake et al. [51] examined the performance of web crawlers for CSAM detection. Experimental results concluded that web crawlers successfully identify CSAM websites if search criteria (e.g., images and keywords) are appropriately selected, obtaining results consistent with the crawler developed in their previous work [36]. A list of websites where CSAM was detected was compiled. In addition, databases of image hashes are frequently used in combination with other detection approaches, as in [36], where image hashing was combined with keywords for online detection of CSAM.

The keyword-based search for CSAM content was divided into three categories: (1) words used by CSAM traffickers, (2) words that are found in CSAM searches but are also found in non-child abuse contexts and help contextualise the website content, and (3) keywords that are commonly used to search for sexual content that is not associated with minors.

A summary of the methodology of the work using web crawlers is outlined in Figure 3.

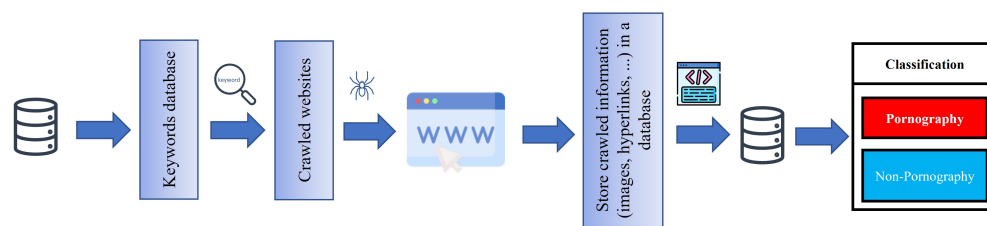


Figure 3. Pipeline of sensitive content classification of web crawlers.

4.3. Strategies Based on Filename and Metadata

Some researchers have focused on complementary cues to the content or scenes of images or videos that aid in the identification of CSAM. Panchenko et al. [52] presented a method to detect CSAM in P2P networks using ML and NLP techniques, being able to recognise author or filename queries in data extracted from P2P networks containing CSAM. The aim of this approach is to discriminate against CSAM and mainstream pornography. This is performed by retrieving information from known CSA files and non-CSA files, by extracting features from the segmented and normalised file metadata. The text classification approach uses conventional statistical machine learning classifiers such as SVM or logistic regression to distinguish normal files from files with CSA content. In addition, the file names complement the frameworks proposed by [13,53,54].

Other techniques include file metadata, queries on major search engines and conversations involving grooming or CSAM sharing [55]. In this context, other work has used textual features to identify CSAM, such as keywords related to website content [36], using NLP techniques, such as conversations [48,56–58]. Previous work has found that attackers tend to use a specific vocabulary to name CSAM-related files. For this reason, the use of file paths (combination of location and filename) is a promising approach for the identification of this type of material.

More recently, Pereira et al. [59] proposed a tool to train and evaluate ML models ready for deployment. Subsequently, they applied the proposed approach to the problem of CSAM detection in metadata-based file storage systems (file path). The resulting model is based on charCNN, achieving an accuracy of 97% and a recall of 94%. Furthermore, the model proposed by the authors is robust against different attacks (such as adversarial attacks). The experiments were performed on a binary dataset labelled as CSA and non-CSA and different ML and DL algorithms. The dataset consists of 1,010,000 real file paths (55,312 unique storage systems) collected by Project VIC International (<https://www.projectvic.org>) (accessed on 20 December 2022).

For the pre-processing stage, the authors performed three different approaches: bag of words (BoW), character N-grams and character quantisation. The authors compared the following classical ML algorithms such as logistic regression, naive Bayes and boosted decision trees against DL architectures such as CNNs and LSTM. The results obtained show that CNN outperformed the rest of the classifiers with 96.8% of accuracy. As for the framework to evaluate the robustness of the model against attacks at test time, the authors conducted two experiments. In the first experiment, the adversary can send file paths to the model but does not have access to the outputs of the model.

The only action the adversary is allowed to take is to make changes to the file paths (the number of modifications in the file path is called the adversarial budget). In this experiment, two types of modifications were made, random substitutions and CSAM lexicon substitutions. For random substitutions and a 15% adversarial budget, they observed a decrease in recovery rates of 0.02% in the bag-of-words models and naive Bayes models, and 0.07% in the CNN model. For a lexicon substitution, in general, selective changes in file paths result in small changes in recall rates. Logistic regression and boosted decision trees exhibit more notable variations than naive Bayes models and deep neural networks.

In order to verify the rate of true positives and false positives and the generalisation of the model, in the second experiment, they evaluated the best-obtained model (charCNN) with a dataset of benign file paths from Common Crawl index CC-MAIN-2021-10. For

decision thresholds above 0.8, the FPR was low for both Linux file paths (0.03%) and Windows file paths (0.001%). The authors did not test other types of architectures more suitable for textual features such as transformers. A summary of the methodology of the work using the filenames and metadata features is outlined in Figure 4.

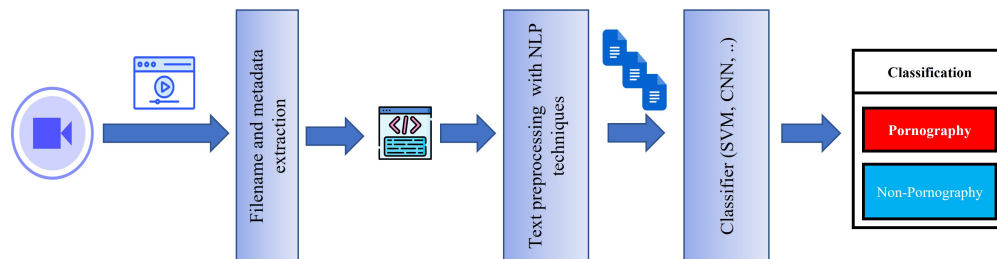


Figure 4. Pipeline of sensitive content classification using filenames and metadata features.

Finally, Table 1 shows the main papers related to text analysis.

Table 1. Most relevant papers related to text analysis.

Reference	Dataset Size	Features	Classification Algorithm	Evaluation Measures
Polastro et al. [53]	330,595 files	File name analysis + Image analysis	SVM	Recall: 95% Precision: 93%
Panchenko et al. [52]	106,350 files	File name analysis + metadata (term extractor + filename normaliser)	C-SVM linear	Acc: 96.97%
Peersman et al. [54]	40,000 CSA file names and 40,000 legal pornographic file names	File name classification	SVM	Precision: 89.9
Bogdanova et al. [57]	Chat logs (5 subsets) from the perverted-justice website [60]	Chat logs (lexicon)	SVM	Recall: 95%
Peersman et al. [48]	330,595 files	File name categorisation (CSA-rel. keywords + Char. n-grams)	SVM	Overall F1-score: 77.75%
Al-Nabki et al. [56]	65,351 files	File name classifier (n-grams)	CNN	F1-score: 85%
Pereira et al. [59]	1,010,000 file paths	File path-based character quantisation	CNN	Acc: 96.8%

5. Strategies Based on Visual Detection

This section describes approaches based on visual features based on skin recognition and image descriptors.

5.1. Strategies Based on Skin Recognition

In the context of image nudity detection, most of the early works have studied human skin recognition. Human skin detection is one of the most interesting topics in the research community. Strategies centred on this approach start from the premise that sensitive adult colour content includes a large part of skin regions. Detection is performed on the basis of low-level features, such as shape colour or global distribution patterns. This strategy has been studied considerably in the field of imaging [61–65].

Thus, conventional techniques developed to classify sensitive video content (e.g., pornography, CSAM) involve extracting visual features based on keyframes. Then, Ref. [17] separated the video stream into shots and keyframes, for the detection of nude regions of the human body. In particular, they used the skin colour pattern distribution feature to propose the corresponding Gaussian model in YCbCr colour space by categorising the skin pixels using a Bayesian method, obtaining an accuracy of 89.2% and 90.3% for the long-shot and short-shot videos, respectively. Lee et al. presented an approach containing two visual features using a support vector machine (SVM) for classification.

The first feature calculates a Gaussian probability of a skin colour pixel on a frame. The second is a group of frames (GoF) that characterises the aggregate representation of colour-based features for several colours of several frames. The method consists of quantifying the hue saturation value (HSV) space into 256 intervals, aggregating from several video frames and calculating the mean of the related interval values. They used 1200 video files to achieve results with 100% and 96.6% accuracy in training and validation.

Eleuterio et al. [16] carried out the detection of child pornography in videos based on a framework named NuDetective and an additional technique for frame segmentation. NuDetective is capable of automatically detecting nudity in images using a threshold ratio of the RGB colour space. In addition, they employed an algorithm from [66], where a threshold is defined for regions where the skin is present. The segmentation method focuses on a function representing the adaptive sampling of the video. For a study of 149 videos, the authors obtained an accuracy of 85.9% for 170 s, which is 0.2% more accurate and 44.8% faster in relation to previous results published in [13].

On the other hand, Ref. [67] presents a python porn image detector (nudity) that uses an HSV colour histogram and other SIFT descriptors. It uses Scikit-Learn and Opencv for feature extraction and classification and achieves 85% tagging accuracy with 1500 positive and 1500 negative samples. On the other hand, Ref. [68] proposed a pornographic image recognition approach based on the oriented FAST and rotated BRIEF (ORB) method. The recognition process is divided into two parts: coarse detection and fine detection. Coarse detection can quickly identify non-pornographic images with few or no skin colour regions and facial images.

For the remaining images, which contain many more skin colour regions, a fine detection is performed, which includes three steps: (1) extract ORB descriptors from the skin colour areas and present them based on the BoVW model, (2) generate the feature vector by combining the ORB feature with the 72-dimensional HSV colour feature of the whole image, and (3) train the classification model with SVM. The authors obtained a recognition accuracy of 93.03% and drastically reduced the average time cost to one-fourth of the SIFT-based process. Finally, Garcia et al. [19] classified files as naked if they meet the threshold conditions of the proposed skin regions in pre-processed multimedia files. In this research, frame extraction was performed using the PHP video toolkit [69]. The skin and non-skin levels are then calculated for each frame. The performance of their proposal was measured on a dataset of 1239 files from the web (986 images; 253 videos) and achieved an accuracy of 80.23%.

A summary of the methodology of the work using the skin recognition approach is outlined in Figure 5.

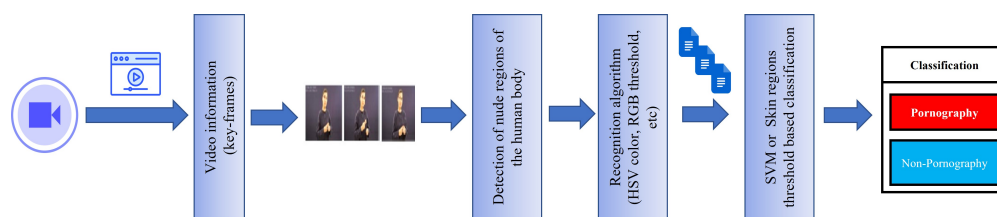


Figure 5. Pipeline of sensitive content classification using skin recognition algorithms.

Table 2 shows an overview of the research strategies based on visual feature analysis.

Table 2. Most relevant papers related to skin recognition.

Reference	Dataset Size	Frame Extraction Algorithm	Recognition Algorithm	Features	Classification Algorithm	Evaluation Measures
Wang et al. [17]	112	Colour difference [70]	Gaussian model in YCbCr	Skin Colour Skin texture morphology	Bayesian	Precision: 90.3% Recall: 91.5%
Lee et al. [14]	1200	Uniform sampling	Single: Gaussian Global: HSV colour discriminant	Skin colour	SVM	Precision: 96.6% Recall: 86.19%
Castro Polastro et al. [13]	149	Uniform sampling	RGB threshold	Skin colour	Skin regions threshold [66]	Precision: 85.7% Recall: 84.9%
Silva Eleuterio et al. [16]	149	Logarithmic function	RGB threshold	Skin colour	Skin regions threshold [66]	Precision: 85.9% Recall: 87.3%
Li Zhuo et al. [68]	19,000		ORB descriptor extraction [71], HSV and BoVW	Skin colour	SVM	Precision: 93.03%
Garcia et al. [19]	253	Uniform sampling [69]	YCbCr threshold Gaussian Low-pass filter	Skin colour Texture skin	Skin regions threshold [66]	Precision: 90.33%

5.2. Strategies Based on Image Descriptor

Due to the drawbacks associated with low-level visual features, an alternative approach named BoVW has been used extensively. The aim is to reduce the difference between low-level features (e.g., pixel colour) and high-level features of sensitive content. This is achieved by using an adaptation of the BoVW model, which is used in NLP tasks such as information retrieval, text classification, etc. The procedure named before is implemented using local patch histograms. In this way, prior to feature extraction, a visual codebook is created using a clustering algorithm in each patch.

From the above results, the feature vector used is constructed from the histogram generated from the frequency with which each visual word appears in the image [72]. This approach has been widely applied in the literature for the detection of sensitive content [18,22,68,73,74]. In this line of research, Ref. [22] used a BoVW-based approach using hue scale invariant feature transform (HueSIFT) to detect nudity in videos.

In this sense, the frames are mapped using a bag of visual features (BoVF) and are finally classified using an SVM (linear) classifier. A voting scheme is implemented based on the obtained results to improve the classification of the video segments, achieving in the best case, a classification accuracy of 93.2%.

In [75], the authors used local spatiotemporal descriptor codebooks (space–time interest points), employing the space–time interest descriptor (STIP) [76] together with a BoVW model, applying a random code sampling method for codebook representation. In the pornography application, the experiments included a 77-h dataset for 800 videos with an overall accuracy of 91.9%.

Avila et al. developed two mid-level representations using keyword-distance distributions, BOSSA [77] and BossaNova [78]. In both investigations, the representation was enhanced by a histogram generated from the calculated distances between the descriptors extracted from the codebook and the image descriptors. They applied a majority voting scheme to classify videos using their frame predictions in both works. The proposal was evaluated with the pornography-800 dataset [78], achieving an accuracy rate of 87.1% and 89.5%, for the BOSSA- and BossaNova-based techniques, respectively.

In addition, Ref. [20] proposed to use a combination of local binary descriptors together with BossaNova. One of the main strengths of this approach is its independence from arbitrary shape or skin detectors to classify sensitive material. The authors analysed a pornography dataset, consisting of almost 80 h of 400 pornographic and 400 non-pornographic videos, previously divided into 16,727 video keyframes and achieving an accuracy rate of 90.9%.

Caetano et al. [21] extended the aforementioned BoVW model by preserving visual information more accurately. In particular, they presented a combination of two video descriptors, BossaNova video descriptor (BNVD) and BoVW video descriptor (BoVW-VD). As for the experimental results, they improved the detection performance by adding the mid-level information of all video frames into a single representation. The experimental results showed an accuracy of 92.4% on the pornography-800 dataset.

In general, spatiotemporal features have not been studied in depth in pornographic-video-detection algorithms. One such approach has been implemented by [72], fusing a BoVW model with MPEG-4 motion vectors to detect pornographic videos. The experiments were conducted on 932 real-world adult web videos and 2663 clips of harmless material collected from YouTube. The inclusion of motion analysis reduced the equality error from 9.9% to 6.0% over traditional BoVW approaches.

In this regard, Ref. [26] presented a spatiotemporal descriptor of points of interest named temporal robust features (TRoF), which exploits the most important motion features in the video. In particular, they used the local information obtained by TRoF in a mid-level representation using Fisher vectors. The performance of this approach is compared to the BoVW solutions described above. Experimental results with the pornography-2k [26] dataset, consisting of 2000 web videos collected from the Internet (1000 pornographic and 1000 non-pornographic), achieved an accuracy of 95%, outperforming the BoVW-based techniques mentioned above.

On the other hand, in [24], the authors used the colour attribute to describe the local colour of sexual organs, concatenating it with the oriented gradient histogram. From the concatenated feature of sex organs using CPMDPM, they detected pornographic images sequentially with sex organ detectors using feature descriptors, histogram of oriented gradients (HoG) and colour attributes (CA). This method preserves colour saliency as well as shape information based on the gradient of the sex organs. Since the false positive rates of sex organ detectors are very low, sex organ detectors are sequentially fused for pornographic image detection. Experimental results demonstrate the superiority of their method compared to the best-known methods, obtaining 80% accuracy, 82% precision and 81% F1-score, and significantly improving the results compared to BoVW and Hue-SIFT techniques. The performance obtained is better than the detector based on shape features, which is outstanding versus the methods based on low-level features of skin regions, which are the BoVW model and scale invariant feature transform (SIFT) features that are embedded in colour.

The feature extractions performed by some of the previous studies, such as GLCM, YCbCr and RGB, were sensitive to image size changes, translation changes or image rotations. In this regard, Hartatik et al. [23] used SIFT and sped up robust features (SURF) techniques, as they are invariant to changes in two-dimensionality, rotation, translation, illumination and size and can extract many key features and better show object descriptors. The training dataset consisted of 7997 pornographic and non-pornographic images taken from Yahoo (NSW dataset) [79], while 984 images were used as test data. The descriptors of each image were clustered using the K-means algorithm. Then, image representation was performed using BoVW, and finally, the classification was carried out using the K-nearest neighbours (KNN) algorithm. In terms of performance, the SURF method achieved better accuracy and time values compared to the SIFT method. The highest accuracy rate of the SURF method was obtained with a dictionary of size 300 with an accuracy of 82.26%.

More recently, Chen et al. [80] proposed an automatic detection system for porn and gambling websites using visual and textual content-based decision mechanisms (PG-VTDM). Initially, Doc2Vec was used with the aim of learning textual features that can map the textual content of the HTML source code of websites. In this way, the traditional BoVW was enhanced by studying local spatial relationships of feature points to better represent the visual features of the website. After that, from these two types of features, they trained two classifiers, a text classifier and an image classifier. In the decision mechanism, a logistic regression (LR)-based data fusion algorithm was designed to obtain the final prediction

result by measuring the contribution of the two classification results to the final category prediction. To evaluate the performance of the proposed procedure, several sets of comparative experiments were performed on a set of real data from the Internet. They collected the data traffic between the user and the web server from the outgoing router mirror and analysed it using protocol analysis tools. The two-feature approach outperformed the single-feature method and some of the more advanced ones, with accuracy, precision and F1-score over 99%. A summary of the methodology of the work using the image descriptors approach is outlined in Figure 6.

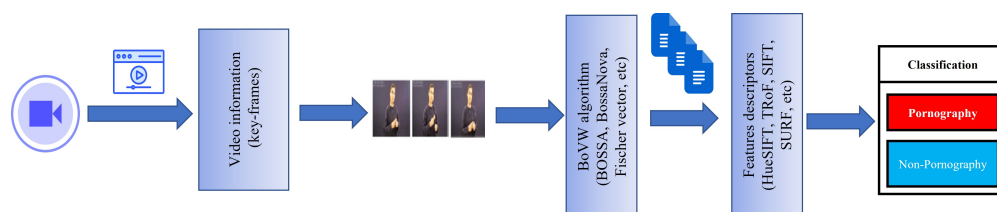


Figure 6. Pipeline of sensitive content classification using image descriptors.

Table 3 summarises the most relevant papers, showing the algorithms and evaluation measures of the papers reviewed in this section.

Table 3. Most relevant papers related to image descriptor analysis.

Reference	Dataset Size	BoVW Type Algorithm	Features Descriptor	Classification Algorithm	Evaluation Measures
Lopes et al. [22]	179	Standard	HueSIFT	SVM (linear kernel)	Acc: 93.2%
Avila et al. [77]	800 800	BOSSA	HueSIFT	SVM (Nonlinear kernel)	Acc: 87.1%
Avila et al. [78]	4900	BossaNova	HueSIFT	SVM (Linear kernel)	Acc: 89.5%
Tian et al. [24]	800	CPMDPM	HoG and CA	Latent SVM	Precision: 80% Recall: 82% F1-score: 81%
Caetano et al. [20]	800	BossaNova	Binary descriptors	SVM (Nonlinear kernel)	Acc: 90.9%
Caetano et al. [21]	800	BossaNovaVD	Binary descriptors	SVM (Nonlinear kernel)	Acc: 92.4%
Jansohn et al. [72]	3595	Standard	Motion vectors	SVM(Not specified kernel)	Equal error: 6.04%
Valle et al. [75]	800	Standard	Motion vectors	SVM (Nonlinear kernel)	Acc: 91.9%
Souza et al. [81]	800	Standard	colour STIP	SVM (Linear kernel)	Acc: 91.0%
Li Zhuo et al. [68]	19,000	Standard	ORB Descriptor Extraction [71] HSV	SVM (Nonlinear kernel) (RBF kernel)	Acc: 93.03%
Moreira et al. [26]	2000	Fisher vector	TRoF	SVM (Linear kernel)	Acc: 95.0%
Hartatik et al. [23]	8981	Standard	SIFT and SURF	KNN	Acc: 82.26%

6. Strategies Based on Motion, Audio and Multimodal Analysis

Most research on sensitive-content detection has been dedicated to examining low- to mid-level features; however, these approaches do not take into account complementary information, such as motion and audio, through which classification performance could be improved. In this context, Rea et al. [82] proposed a multimodal approach, using skin colour estimation together with periodic patterns in the audio of a video with sensitive content. The periodicity of the audio is obtained by localising the maxima and minima in the autocorrelation of the respective energy signal. The method used was evaluated on one test film only, which excludes the feasibility of obtaining overall classification findings.

Another work [83] also used auditory features for pornography detection based on a Gaussian mixture model (GMM) using a 13-dimensional feature vector (12 Mel frequency cepstral coefficients—MFCC—plus an energy term) from the audio signal of the video samples. In order to improve recognition, they used a combination of a generalised contour-based algorithm and a Bayes classifier. The performance was 92.3% on a dataset composed

of 352 blue videos and 537 normal videos. In this study, the periodicity of the audio is not taken into account, and the spacing between frames is very short.

In this sense, the methods named above lack an accurate representation of audio semantics and show little focus on pornographic audio features. Yizhi et al. [84] present a novel approach to fusing audio vocabulary with visual features for pornographic video detection. The novelty of their strategy consists of three aspects: (1) representation of audio semantics by means of an energy envelope unit (EEU) and BoVW, a periodicity-based audio segmentation algorithm, and a periodicity-based video decision algorithm. The first one, called the EEU+BoVW method, aims to describe the audio semantics by means of a vocabulary. The audio vocabulary is generated by k-means clustering of EEU. The latter two aspects complement each other to take full advantage of the periodicities of pornographic audio. Before fusion, two SVMs are applied for the methods based on audio vocabulary and visual features. Finally, the result fusion is performed by selecting a key frame from each EEU based on the initial and last positions, to implement a weighted scheme and a periodicity-based video classification algorithm to achieve the final outcomes.

To evaluate their proposal, they collected videos from the Internet and formed a training and test set, consisting of 48 pornographic and 300 benign training videos and 50 pornographic and 150 benign test videos. The final results show how their approach improves on the traditional approach, which is based solely on visual features. The TPR reaches 94.44%, while the FNR is 9.76% on the dataset collected by themselves. Furthermore, the authors of this paper do not use the motion features of the videos for classification.

Kim et al. [85] use the motion vectors of each frame to detect global motion. In the presence of local motion recognition, the algorithm performs skin region detection using invariant moment features. Lastly, the classification is carried out using the shape matching algorithm described in [86]. During the validation process, 2275 conventional and 980 sensitive videos were analysed, with a final accuracy of 96.5%.

The authors of [87] proposed motion periodicity recognition to capture recurrent patterns in pornographic videos. This detection was performed by spectrally calculating the dominant motion for 16-second intervals of the video sequence. The final results are only approximate, and given the restricted video material used for validation, no general conclusions can be drawn.

In another work, following a similar line of research, Ref. [88] extracted motion vectors from MPEG video sequences and smoothed them using a median filter. This methodology calculates the direction and strength of the motion vector, and the detection of pornography is carried out based on a threshold. Experimental results were carried out using 30 pornographic and 70 conventional videos with an overall accuracy of 90%. The weakness of the developed approach is that it cannot correctly identify pornographic videos with large or no global movements.

In a later work [89], a multimodal method is proposed that includes several features such as skin colour, motion histograms, the coefficients of the discrete cosine transform of image patches using a BoVW model and audio features employing cepstral coefficients. The experiments showed how the multimodal approach combining all features improves the accuracy considerably, reducing the equal error by 36–56% compared to the best unimodal system. The dataset analysed in this work is composed of 1000 pornographic clips and 2300 YouTube clips.

Behrad et al. [90] proposed to identify the largest section of skin colour by extracting six motion-based features employing the Fourier transform of the inter-frame autocorrelation. The classification of pornographic videos is performed using an SVM classifier, obtaining an accuracy of 95.44% over 2000 episodes of pornographic videos and 2000 episodes of conventional videos. On the downside, the skin-recognition method may not work well in some cases.

On the other hand, Jung [91] uses spatiotemporal motion patterns to perform pornographic video detection. In this work, features such as the magnitude and frequency of periodic motion are extracted, as well as features such as skin colour, obtained from satura-

tion and hue. The experiments were carried out on 1500 video segments (500 adult videos and 1000 documentaries), while the test set consisted of 18,313 scenes (1103 scenes with at least one sample of sensitive content). The results obtained were better than the work of [89,90], being slightly less efficient in detecting pornographic material that contains a tiny portion of skin colour in each frame. In addition, this approach performs better in the absence of skin colour, unlike methods that do not use skin colour as a part of the classification, as described in [87].

In this context, Schulze et al. [92] proposed a multimodal approach for automatic CSAM detection using visual and audio features. For this purpose, they employed low-level features such as skin features, the mid-level sentiment feature SentiBank for images, colour correlograms, visual words as well as audio words for videos. The dataset used consisted of three classes, namely global (non-offensive), adult pornographic and CSA content. For the image experiments, they collected a total of 60,000 samples, 20,000 for each class. For the video experiments, they collected a total of 3000 samples, 1000 for each class. Subsequently, they applied a separate SVM classifier for each feature mentioned above.

In terms of the results achieved, the colour-correlogram was the best performer among all the low-level features in image classification, particularly for distinguishing CSAM files from adult pornography. Skin patterns are not the best suited for correctly classifying adult pornography, and they only work well for classifying adult content material from normal media, but not for CSAM detection. As for video classification, adding audio information significantly enhances detection accuracy. The multimodality of the features improves the final classification and reduces the error rate compared to unimodal models (error rate 10% for images and 8% for videos).

Recently, Liu [93] proposed an approach that combines visual saliency and audio periodicity for adult video detection. Thus, after analysing the periodic patterns and the salient regions, respectively, in audio and visual frames, a multimodal model formed by the combination of both features is obtained. The experiments show that this approach achieves remarkable results and improves on the aforementioned methods that rely solely on visual features.

A summary of the methodology of the work using the multimodal approach is outlined in Figure 7 (some works use the visual characteristics extraction pipeline, some use the audio characteristics extraction pipeline, and some use both types of features).

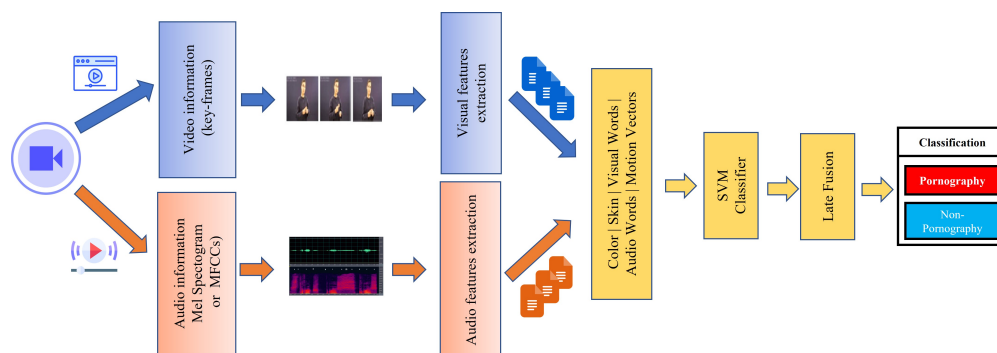


Figure 7. Pipeline of sensitive content classification using motion, audio features and multimodal features.

Table 4 presents a summary of the main strategies based on motion and audio analysis for the problem of pornographic video detection.

Table 4. Most relevant papers related to motion, audio and multimodal analysis.

Reference	Dataset Size	Features Analysed	Classification Algorithm	Evaluation Measures
Zuo et al. [83]	889 videos	12 MFCC and energy term body contour	GMM Bayes classifier	Precision: 92.3% Recall: 98.3%
Kim et al. [85]	3255 videos	Motion vectors moments of shape	Shape matching [86]	Acc: 96.5%
Endeshaw et al. [87]	750 videos	Motion vectors	Spectral estimation threshold	TPR > 85% FNR < 10%
Zhiyi et al. [88]	100 videos	Motion vectors (strength and direction)	Two-motion features threshold	Acc: 90.0%
Ulges et al. [89]	3300	Motion vectors, audio features, skin colour	SVM (RBF kernel)	Equal error: 5.92%
Behrad et al. [90]	4000 videos	Motion and periodicity features	SVM (Linear kernel)	Acc: 95.44%
Schulze et al. [92]	60,000 images	Colour-correlograms + skin features, visual pyramids + visual words + SentiBank mid-level sentiment feature	SVM RBF (for each feature) + late fusion	Equal error: 10%
	3000 videos	Colour-correlograms + skin features, visual pyramids + visual words + audio words	SVM RBF (for each feature) + late fusion	Equal error: 8%
Liu et al. [84]	558 videos	Periodicity-based video	SVM (RBF kernel) BoVW	Acc: 94.44% FPR: 9.76%
Liu et al. [93]	548 videos	Audio periodicity and visual saliency colour moments	SVM (RBF kernel)	TPR: 96.7% FPR: 10%

7. Strategies Based on Deep Learning

Considering the difficulty of establishing correct thresholds for sensitive-content detection in the above-mentioned strategies, such as skin- and motion-based ones and the growth of deep learning, the latter techniques have been considered as important solutions in the area of sensitive-content detection, taking advantage of the capability of these networks for automatic feature extraction.

7.1. Early Approaches Based on CNNs

Since the emergence of CNNs to solve computer vision tasks, they have recently been successfully employed in sensitive-content detection [94–98].

Moustafa et al. [29] was the first to evaluate the use of DL to deal with the problem of sensitive-content detection in videos. The authors proposed a hybrid approach using two different CNN architectures (GoogLeNet [99] and AlexNet [100]), pre-trained with the ImageNet dataset [101] and tuned with pornographic data, which was proposed for pornography detection. Experiments were carried out with the dataset described in [78], containing 400 adult videos and 400 normal videos. For the classification of the videos, keyframes were selected and tested separately. The last decision used majority voting to obtain the final score. The method presented achieved an accuracy of 94.1%, improving the results obtained in [20,78]. Despite the good results, this work does not employ motion features.

Perez et al. [30] also proposed a strategy based on CNNs using, in this case, only GoogleNet. In terms of the type of features used, they employed motion vectors and optical flow (motion information) and static features (raw frames). The motion vectors consisted of the extraction of motion data from the video and the associated image, feature extraction using the CNN model, a concatenation of the vertical and horizontal (dy) and (dx) descriptions, an average clustering of the descriptions, and finally an SVM classifier. Experiments were performed on the pornography-800 and pornography-2k datasets presented in [78] and [26], obtaining an overall accuracy of 97.9% and 96.4%, respectively.

More recently, Ref. [98] presented an intelligent filtering method using modern CNNs. The authors presented a new architecture that combines the AlexNet and LeNet [102] architectures. Using this strategy, the chosen parts of each layer are like AlexNet, and the

other domains are like LeNet, in which each convolutional layer is immediately followed by a clustering layer. For the experiments, they used the [78] dataset on which they obtained an accuracy of 95.83%. The results obtained using the CNN network architecture are superior to the results presented in the previous sections.

7.2. New Approaches Based on Fusion Models (CNNs and RNN)

To enhance the performance obtained in the strategies using CNNs, other researchers opted to use temporal features (changes between video frames) and the combination of several kinds of features (multimodal approach) extracted by different types of neural networks.

Wehrmann et al. [103] presented a novel strategy combining GoogleNet and ResNet [104] architectures for static visual feature extraction from keyframes and the LSTM network for obtaining the final classification score. The presented method, called Adult Content Recognition (ACORDE), was evaluated on the pornography-800 [78] dataset, obtaining an accuracy of 95.6% with the ResNet-101 model.

Song et al. [105] proposed a DL-based approach using multimodal features, such as image descriptor features, visual features of each frame, audio features extracted from video, and motion features using optical flow. For image descriptor extraction, as for the detector using video descriptor features, the authors employed the pre-trained VGG-16 model. First, the video data were processed into suitable data forms, such as frame image, motion, frame sequence, and audio. These multimodal features were used to create each independent detector. At that time, the image-based detector performed the integration of results to make the final decision. Finally, the stacking model combined the results of each detector to produce the final decision result.

As for the detector based on audio features, they divided the audio data of the video file by a unit of 10 s. Subsequently, they extracted the audio features from the audio frame that constituted the audio clip, using the Mel-scale spectrogram. Finally, the features of each audio frame were aggregated into a single audio descriptor using the pooling method. The dataset used for training and SVM testing was pornography-2K. The authors performed a 10-fold cross-validation but did not follow the methodology recommended by the authors of the dataset; thus, the results are not fairly comparable with the state of the art. As for the results, they obtained an accuracy of 88.3% for the motion texture-based detector and 80% for the audio feature-based detector. For the multimodal detector, they only obtained an accuracy of 67% and a high false positive rate, although they obtained a true positive rate (TPR) of 100%.

On the other hand, Silva and Marana [28] used two spatiotemporal 3D CNNs: CNN VGG-C3D [106] and CNN ResNet R(2+1)D [107] for the detection of pornography in videos. The CNN ResNet R(2+1)D architecture allows for adding non-linear activation functions such as rectified linear unit (ReLU) between 2D and 1D convolution, obtaining a higher number of non-linearities. They used the pornography-800 dataset for the experiments, obtaining an accuracy of 95.1% and 91.8% for VGG-C3D CNNs and R(2+1)D CNNs from residual neural networks (ResNet), respectively. On the contrary, despite using video recognition tools such as 3D CNNs, they did not improve on the results of the recent works for this dataset.

In this line of research, Singh et al. proposed in [108] a specific approach to detect unsafe content for children, called KidsGUARD. The proposed strategy consists of using an LSTM-based autoencoder to extract the representative video features from a CNN VGG16. The experimental results were validated with the pornography-800 [78] dataset while adding animated videos containing short nude clips. The best classification results for the [78] dataset had 89% accuracy and 85% recall values.

More recently, Mallmann et al. [31] proposed a framework called private parts censor (PPCensor) based on the Faster R-CNN network for real-time detection of pornographic content. The network was trained on the Private Parts Object dataset created by the authors themselves. The network can detect pornography based on the private parts detected in the images. PPCensor also shows similar results to other techniques in terms of FNR,

reaching 2.34% on the pornography-2k [26] dataset, which is an increase of only 0.31% compared to the best CNN architecture. However, the detection performance of PPCensor is significantly better in object detection than the other CNN architectures. The authors did not follow the evaluation methodology for the pornography-2k dataset; thus, the results cannot be fairly compared.

Recently, Papadamou [109] proposed a classifier for inappropriate YouTube videos targeting young children using an ensemble approach based on DL. The final model is made up of four different components, where each consider a different type of feature: tags, title, thumbnail, and statistical and style features. Finally, the result of each component is concatenated to build a fully connected two-layer ANN (artificial neural network) that fuses the outputs to produce the final classification. The dataset used consisted of 4797 videos downloaded from YouTube marked as age-restricted. The developed classifier achieved an accuracy of 84.3%. The model performance was constrained by the small training size and the imbalanced dataset.

Song et al. [32] proposed a multimodal stacking scheme for fast and accurate online detection of pornographic content on the Internet. To accurately detect sensitive content, visual and auditory features were extracted using a VGG-16 with a bidirectional RNN to reflect the patterns of signal change over time within each input. Combining both features, a fusion classifier was also constructed. These three component classifiers were stacked in the improved ensemble scheme (fusion classifier, video classifier and audio classifier) with the objective of reducing false negative errors. The dataset used was pornography-2k [26].

For model building, the sensitive contents of the selected dataset contained noxious factors in both visual and auditory elements. Finally, 8000 segment instances were randomly selected (5000 for training and 3000 for testing), each containing 4000 segments of sensitive and non-sensitive content. Since samples labelled as sensitive from each content segment may differ in terms of visual and auditory elements, the harmfulness of the corresponding element was used to select the data used for training each classifier component. The experimental results yielded an accuracy of 92.33%, higher than previous studies and better than the results achieved by the other two classifiers, visual and auditory, separately (95.33% and 89.16%, respectively).

Chaves et al. [110] compared the speed and accuracy of three popular DL-based face detectors on the WIDER Face [111] and UFDD [112] datasets on various CPUs and GPUs. Their method for CSEM detection is based on a combination of face detection, age estimation and pornography detection. The experiments were conducted on a GNU/Linux machine running Ubuntu 18.04 (Cuda 9 and CuNDD 7) with the objective of comparing the accuracy-to-speed ratio on GPUs and CPUs of the publicly available implementations of the MTCNN, PymaridBox and DSFD face detectors using input comprising four relative sizes (100%, 75%, 50% and 25%) of the original sizes and evaluating the performance of the built models of the face detectors on a given image with specific hardware.

The results confirmed that the use of resized images speeds up the face detection phase but reduces accuracy. In addition, they found how the speedup previously achieved resizing and GPU, depending on the complexity of the face detector used. The results obtained with multiple linear regression models are able to predict the performance of face detection with an MAE of 0.113.

Lee et al. [113] proposed an approach to assess integrity by analysing the visual features of digital content using the rate of change of adjacent frame features and then testing whether the video has been tampered with. The results obtained on the test set showed a detection rate of 97%, superior to existing methods. Their method consisted of extracting a facial image from the frame (using an MTCNN model), then extracting visual features and finally calculating the difference between the frames. Three datasets were used to evaluate their proposal. The Face2Face and FaceSwap datasets are provided by FaceForensics++ [114]. This set consists of more than 1000 videos. On the other hand, the Deepfake Detection Challenge (DFDC) dataset contains more than 470 GB from Kaggle [115]. Finally,

their classification method detected a deep fake using a DNN that obtains the variance of a given number of frames from the pre-processed data.

In this context, Aldahoul et al. [116] presented a method to address the high FNR in the various previously presented works on pornography detection. To do so, their proposal addressed the shortcoming of current convolutional approaches that focus their visual attention on the expected nudity regions within the frames to reduce the FNR. They used the You Only Look Once (YOLO) object detector for pornography and nudity detection to detect people as regions of interest (ROI), which were applied to CNN and SVM for nude/normal classification. They carried out several experiments to compare the performance of various CNNs and classifiers using their own dataset (Multimedia Malaysia University).

In the first experiment, they observed that the ResNet50 architecture provided on average 95.04% better accuracy than the 92.43% obtained in [103]. The high FNR rate was due to the misclassification of the nude frames, as they have small-scale nude regions with complex backgrounds. In the second experiment, the existing ResNet50-only method, performing fine-tuning with the pornography-2k dataset, was compared with the proposed YOLO-ResNet50-SVM method. Furthermore, in this second experiment, they explored the limitations of convolutional approaches that apply CNNs directly to full frames of videos and the limitations of the pornography-2k dataset on the small-scale nudity detection task. In this sense, if the nudity-detection model detects at least one naked person, then the whole frame is considered as sensitive, whereas if all persons are predicted as safe, the whole frame is considered normal. The YOLO+ResNet50 method outperforms ResNet50 in accuracy and F1-score in the test phase.

Finally, in the last experiment, the goal was to combine YOLO3 with other CNNs and classifier architectures. Therefore, they combined the six CNN feature extractors (e.g., AlexNet, VGG16, GoogleNet, Inception3, ResNet50, ResNet101) with the six conventional classifiers (e.g., LSVM, GSVM, PSVM, KNN, RF, ELM) to have 36 models in total. To evaluate the performance of the models, 25,983 human images detected from 8579 test images were tested for each feature extractor and classifier. It was observed that ResNet101 with random forest as the final classifier outperformed the other models with an F1-score of 90.03% and an accuracy of 87.75%.

In addition, an ablation study was performed to demonstrate the impact of adding YOLO before CNN, resulting in an increase from 85.5% to 89.5%. A major limitation of the authors' proposed method is related to the performance of the human detector. If the detector is not able to detect a human in the frame, the naked frames are misclassified. The authors did not test their model for the detection of pornography in videos (only in images).

More recently, Lovenia et al. [117] proposed an audio-based approach for pornographic detection. This audio-based method allows for the filtering of sensitive content by exploiting different spectral features. To do this, they explored models using alternative audio features and neural architectures. They found that a CNN trained on the log mel spectrogram reaches the highest performance on the pornography-800 dataset than MFCCs features. The results of their experiment also show that the log mel spectrogram provides better features for models to recognise pornographic sounds. Finally, to classify whole audio waveforms instead of segments, they employed a segment-to-audio voting technique which produces the best audio results. The authors obtained test F1-scores of 94.89% on the segment and 92.02% on the audio level with CNNs using log mel spectrograms as features.

A recent work [118] proposed a frame sequence ConvNet pipeline based on ResNet-18 for feature extraction and to analyse the frame feature map using a proposed ConvNet intended for frame sequence classification, thus encapsulating motion information by encoding changes in the ResNet output feature vector.

To do this, in the first instance, they performed human detection to select those frames with a human presence in the scene. Subsequently, the selected frames were introduced into the ResNet-18 network to extract and combine each frame's 512 feature map. Finally, the feature vector fed the sequence classifier ConvNets, consisting of several convolutional

layers and a fully connected network, and they performed prediction using the softmax function on each class for each feature map of the video. This network has the objective of encapsulating sequence dependencies and detecting the kind of motion that is present in the feature map of the sequence.

In addition, they employed data augmentation techniques, normalisation with Image Net statistics across each RGB channel, colour-jitter, random rotation at 45°, and random horizontal-flip or vertical-flip, so that the network would learn the generalised features and would not over-fit. With the proposed model, they achieve a state-of-the-art accuracy of 98.25% in classifying pornographic videos in the pornography-800 dataset and a state-of-the-art accuracy of 97.15% in classifying videos in the pornography-2k dataset.

As a counterpart, they did not use all the frames of each video of the dataset in the training, but they used the frames of the medium after a previous analysis of where the most representative (e.g., pornographic) frames are located. Therefore, some frames that can be used to detect pornography may be left out. In addition, human detection may not be useful in the case of occlusion of the face or even part of the body, which may reduce the hit rate for CSAM data in real environments.

A summary of the methodology of the work using CNN architectures and motion and audio features is outlined in Figure 8 (some works use the visual characteristics extraction pipeline, some use the audio characteristics extraction pipeline, and some use both types of features).

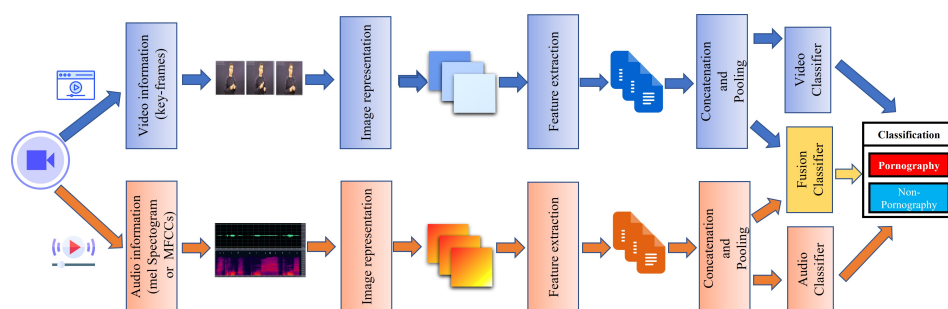


Figure 8. Pipeline of sensitive content classification using motion and audio features.

7.3. Vision Attention

Attention-based architectures, such as transformers [119], have become the default model in NLP. Inspired by the successes of NLP, multiple works have attempted to merge CNN-like architectures with the self-attention mechanism in the field of computer vision [120,121], and others have replaced CNNs altogether [122–124]. This process is shown in Figures 9 and 10.

Chen et al. [34] proposed a three-component approach for conventional pornography detection: (1) a deep one-class with attention to pornography (DOCAPorn) network architecture to address the problem of inadequate negative datasets and inaccurate detection of samples that are not in the training set, (2) pre-processing for compressing and reconstructing (PreCR), a pre-processing approach to minimise small perturbations that may exist in pornographic images and that re-generate corrected images in order to ignore adversarial attacks in the area of sensitive-content detection, and (3) the scale constraint pooling (SCP) scheme to obtain a fixed output size for different input sizes of pornographic images. For the experimental results, the authors compared their method with pornographic image recognition methods based on human skin, image descriptors (BoVW) and neural networks (ResNet-based, Xception-based [125]), obtaining 95.63% accuracy on the pornography-800 dataset [78] and 98.42% accuracy with the attention mechanism and 96.35% accuracy without the attention mechanism on their own dataset. On the negative side, the authors could have tested their proposal with a larger and more challenging dataset, such as the pornography-2K dataset (composed of 2000 videos with scenes that can give a high false positive rate, such as sumo, swimming, etc.).

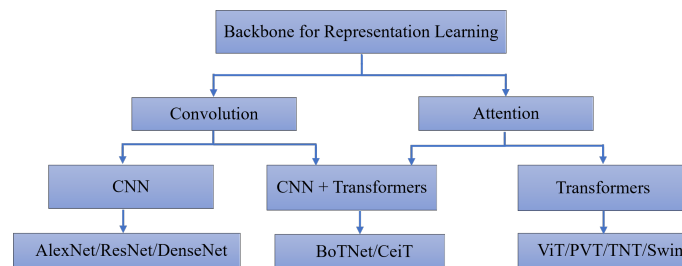


Figure 9. Backbones using convolution and attention.

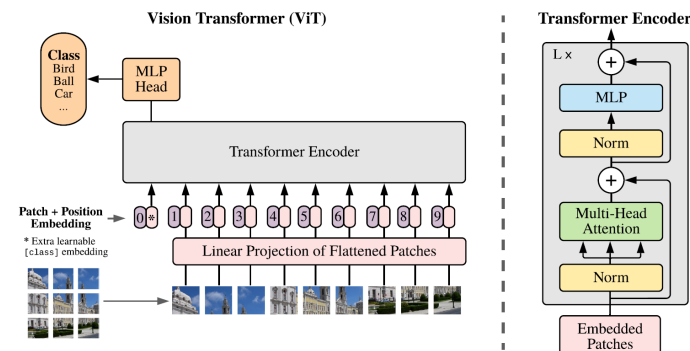


Figure 10. Diagram of ViT transformer (image from [123]).

In another recent work, Lin et al. [126] presented a strategy employing transfer learning with multi-feature fusion. For sensitive-content detection, they fused four separately trained DenseNet-121 [127] models by freezing different surface layers using 120,000 images divided into five classes: porn, sexy, hentai, cartoon and neutral. Compared to a single DenseNet121 network, the authors obtained an improvement of about 1%. For the experiments, they first obtained training data from an open dataset called Not Safe For Work (NSFW) [79] with more than 120,000 images and the set pornography-800 [78]. The images are classified into different levels in terms of the sensitivity of their content. In addition, they proposed a fusion approach using multiple transfer learning models in the inference stage to enhance accuracy in the test set by inserting four attention blocks between each dense block and its next layer. The model training setup included transformations such as random rotation and flips horizontally or vertically. The authors obtained 94.3% on the pornography-800 dataset and 94.96% on the NSFW dataset.

Another work that uses the attention mechanism for pornography detection is Gangwar et al. [35]. In this paper, they split automatic CSAM detection into two subtasks: (1) detection of pornographic content and (2) age group classification of a person as underage or an adult. They introduced a CNN architecture with an innovative attention mechanism and metric learning mechanism, called AttM-CNN. Furthermore, they fused the pornography and age-group classification method for CSAM detection using two different strategies: decision-level fusion for binary CSAM classification and score-level fusion for suspicious image reorganisation. On the other hand, they introduced two new datasets: (1) Pornographic-2M, consisting of two million pornographic images, and (2) Juvenile-80k, which includes 80,000 manually labelled images with recognisable facial age. Experiments related to age and pornography showed similar or better results than state-of-the-art systems on several benchmark datasets for both tasks, respectively.

The three highlights of their AttM-CNN architecture are: (1) visual attention to highlight the most important regions of the image, (2) metric learning combining central loss metric and softmax function to learn better discriminative representation, and (3) concatenation of CNNs, especially residual connections and inception modules, for better model learning while maintaining computational cost and fewer parameters. As for the experiments, first, they trained the AttM-CNN network with 1.2 million training images and 50,000 validation images from the Imagenet set (the network weights are initialised using Kaiming initialisation).

Once the network is trained, the AttM-CNN-Porn model is initialised (except for the last layer). For training, they used one million pornographic images (pornography-2M) and one million images of people from the Google open image set for the non-pornographic class. Finally, to test the constructed pornography-detection model, they used the pornography-2k video dataset, obtaining 97.10% accuracy. In addition, the authors validated their approach with a real CSAM set provided by the police, adult pornography images and secure images (including images of minors). Their model obtained an accuracy of about 92.72% in the binary classification of CSAM. They also showed that 80% of CSAM images can be found among the top 8.5% of images in a classified list created using the CSAM score obtained from their approach. The main limitation of their proposal was the failure in detecting faces in pornographic or CSAM images, which leads to an erroneous estimation of the age group when the face is not visible.

In the context of attentional mechanisms, unlike conventional techniques that rely only on visual features without considering auditory ones, Fu et al. [33] presented a unified DL system called PornNet that integrates dual subnetworks for pornographic video detection. In particular, audio tracks and frames extracted from pornographic videos are respectively provided to two deep networks for features extraction. To classify pornographic frames, they used a local network that takes image context into account when capturing key contents, while they leveraged an attention network that can capture temporal information to recognise pornographic audio.

For this purpose, they proposed a network (DCNet) to recognise pornographic frames by classifying video frames through simple voting. For audio, they used VGGish, yielding audio feature embeddings, which are log mel spectrograms and image representations of the audio. Subsequently, for the recognition of audio feature embeddings, the authors used the RANet network, subsequently generating video–audio results. Finally, in the video–audio fusion algorithm, a function is predefined to aggregate the video–audio recognition result.

The performance of the proposal was evaluated on a recently collected dataset, showing how the proposed method performs well, reaching an accuracy of 93.4% on the in-house dataset that includes 1000 porn samples along with 1000 normal videos and 1000 sexy videos. The authors did not evaluate the proposal with other pornographic datasets; thus, comparison with other proposals becomes difficult.

Finally, in the context of detecting and classifying sensitive content from YouTube videos, Yousaf et al. [128] presented a study on the integration of an automatic real-time video media filtering system for social media platforms. To do so, they employed a pre-trained CNN ImageNet (EfficientNet-B7) to extract video features. Subsequently, these features feed a bidirectional short-term memory (BiLSTM) to learn efficient video representations and to perform multiclass classification. All evaluation experiments were performed using an auto-generated dataset of 111156 YouTube cartoon videos. The authors integrated an attention mechanism after BiLSTM to apply the probability distribution of attention in the network. The experimental results showed that the EfficientNet-BiLSTM model without the attention mechanism performed better (95.66%) than the same model with the attention mechanism (95.30%). In addition, traditional ML-based classifiers (e.g., KNN, SVM, etc.) performed worse than DL-based classifiers.

Moreover, performance analysis with the more advanced methods showed how the BiLSTM architecture on CNN better captures contextual content information from video features, resulting in better results in detecting video content inappropriate for children.

Finally, different experiments were carried out that indicated the differences between other pre-trained models and the characteristics of the selected model compared to the other classifiers. The advantages of their framework can help to filter inappropriate content in real time. The authors did not take into account the time flow and did not perform cross-validation and data augmentation to train the model, which would make the results more reliable.

In total, the results and characteristics of the sixteen main deep learning models developed thus far for sensitive-content detection are shown in Table 5.

Table 5. Most important papers related to deep learning techniques.

Reference	Dataset Size	DL Architecture	Classification Algorithm	Evaluation Measures
Moustafa [29]	800 videos [78]	Fusion (AlexNet and GoogleNet CNN)	Majority voting	Acc: 94.1%
Perez et al. [30]	800 videos [78] 2000 videos [26]	GoogleNet-based CNN	SVM (linear)	Acc: 97.9% Acc: 96.4%
Wehrmann et al. [103]	800 videos [78]	Fusion (ResNet and GoogleNet CNN)	LSTM-RNN	Acc: 95.6%
Song et al. [105]	2000 videos [26]	Fusion video (VGG-16) + motion (VGG-16) + audio (mel-scaled spectrogram)	Multimodal stacking ensemble	Acc: 67.6% TPR: 100%
Silva and Marana [28]	800 videos [78]	VGG-C3D CNN ResNet R(2+1)D CNN	SVM (Linear) Softmax classifier	Acc: 95.1% Acc: 91.8%
Singh et al. [108]	800 videos [78] + Animated videos with nudity	VGG16 + LSTM autoencoder	LSTM classifier	Pre: 89.0% Rec: 85.0%
Papadamou et al. [109]	4797 videos	Inception-V3 CNN (thumbnail)	2 LSTM RNN + dense layer	Acc: 84.3% Rec: 89.0%
Song et al. [32]	2000 videos [26]	Fusion video (VGG16 + Bi-LSTM) + audio (multilayered dilated Conv.)	Multimodal stacking ensemble	Acc: 92.33% FNR: 4.6%
Chen et al. [34]	800 videos [78] + Custom Dataset (1,000,000 images)	DOCAPorn (VGG19 modification + visual attention)	Softmax classifier	Acc: 95.63% Acc: 98.42%
AlDahoul et al. [116]	2000 videos [26]	YOLOv3 + ResNet-50	Random forest	Acc: 87.75% F1-score: 90.03%
Lin et al. [126]	120,000 [79] + Pornography-800 [78]	Fusion DenseNet121(4) + visual attention	Softmax classifier	Acc: 94.96% Acc: 94.3%
Ganwar et al. [35]	Training: Pornography2M [35] + 1M Google Open Dataset [129] Testing: 2000 videos [26]	CNN + Inception + inception reduction + inception-ResNet + attention	Softmax classifier + centre loss	Acc: 97.1% F2: 97.45%
Fu et al. [33]	30,000 videos	ResNet-50 + BiFPN + ResNet-attention network (RANet) + VGGish	Softmax classifier	Acc: 93.4%
Yousaf et al. [128]	111,561 videos	EfficientNet-B7 + BiLSTM	Softmax classifier	Acc: 95.66% F1-score: 92.67%
Lovenia et al. [117]	800 videos [78]	CNN (audio features)	Voting segment-to-audio alg.	Acc: 95.75%
Gautam et al. [118]	800 videos [78] 2000 videos [26]	ResNet-18 + sequence classifier ConvNets + faster RCNN-inception ResNet V2	Softmax classifier	Acc: 98.25% Acc: 97.15%

8. Results, Challenges and Open Issues

The works analysed in this study show a great diversity of combinations of feature inputs and algorithms for the recognition or automatic detection of sexually explicit videos. As described in Section 1, the most commonly used features in the automatic detection of sensitive content are visual, although recently, text and auditory features are also being used.

8.1. Results

The strategies using text features, summarised in Section 4, aim to use other types of features either generated as hashes or associated with the images as file names, path files, and metadata. On the other hand, some works combine visual and textual features (including chat logs between users exchanging this type of material). In terms of hashes, more advanced methods have emerged in recent years that use DLs, such as in [44,45,47], improving for example the robustness against images that undergo small changes, resulting in fewer FP. On the other hand, some researchers have recently discovered how some methods based on hashes, such as [6], have some limitations and vulnerabilities [49]. Among these limitations, the use of equalisation for scaling the sum of gradients increases the likelihood of a false negative for any smaller edit. Regarding vulnerabilities, Ribosome [130] inverts PhotoDNA hashes using machine learning. His demonstration uses provocative images to show that approximate body shapes and faces can be recovered from the PhotoDNA hash. Like any other lossy function, the PhotoDNA hash is not perfectly invertible, but the hash leaks a lot of information about the original input, as these image recreations demonstrate. Because of the above, and because the use of hashes always involves a comparison with a database of previously identified materials (hashes), they do not recognise new cases of CSAM; thus, organisations using this approach alone must adapt other methods to make detection more efficient.

With regard to webcrawlers, Westlake et al. [51] found that specialised databases (e.g., hash values) are valid criteria for identification. However, their reliability, and thus their usefulness, depends on the completeness of the database. Moreover, subsequent research cannot rely solely on the results of previous research to choose the selection criteria. Of the 27 keywords from previous research such as in [50], more than half (16) had a minimal presence on any given website. Even code keywords that persist are not particularly useful if used alone, as they are found in other types of websites.

Regarding filename, metadata and other textual features (e.g., chat logs), successful results have been achieved [48,52] by using ML algorithms such as SVM and by applying CNNs on text features [56,59]. The use of this kind of approach involves manual pre-processing with NLP techniques, which means having good prior knowledge of the way CSAM files are named or the way (lexicon) of communicating in chats, etc. The best result was obtained in [59], employing file paths and using CNN (one dimension) with an accuracy of 96.8%.

Textual feature-based approaches have the advantage of not working directly with CSA photos or videos, thus providing a media-agnostic classifier. In such approaches, it has not been possible to compare the different results due to the lack of a common dataset (file names and metadata features). As a main disadvantage, it is necessary to know, in depth, the behaviour of the offenders, e.g., how they name the files, in order to enhance the generalisation of the final model.

The strategies using visual features, summarised in Section 5, mainly use skin colour and skin texture to perform recognition or detection of sensitive content, by defining a threshold using a specific colour model. Considering that the results obtained in sensitive content classification are highly influenced by the quality and size of the dataset, it is hard to make a comparison between the different methods without a common dataset in the validation phase. In any event, regarding the largest dataset, the best results were obtained in [14] with an accuracy of 96.6%, while employing an HSV colour space and an SVM algorithm for the classification stage.

The strategies described in Section 5 show acceptable results, but as they are highly dependent on skin colour count, this can lead to errors in scenes with many people or with objects that have colours similar to human skin. In this sense, skin colour-based algorithms are not able to generalise correctly for colour changes from different ethnic groups or from those caused by lighting variations. To address these shortcomings, several studies focus on an alternative methodology using the analysis of image descriptors, as shown in Section 5.2. In this sense, most of the works that used these features validated their results with common public databases, facilitating the analysis and comparison of the best model configurations, with SVM being the ML algorithm with the better performance. This ML approach has been considered a simpler strategy, as it relies on fewer parameters for the definition of the classification model.

The kernel used in the SVM algorithm with which the best results were obtained was the linear one. Caetano et al. [21] recorded the best accuracy values (92.4%) for the dataset described in [78] (800 videos) using the BossaNovaVD algorithm and a binary feature descriptor. The best results obtained using descriptors were by Moreira et al. [26], who used a combination of Fisher vectors and TRoF descriptors to obtain 95% over the dataset and who published a larger dataset (2000 videos) created in their work.

More recently, Hartatik et al. [23] obtained 82.26% with SIFT and SURF descriptors using KNN as a classifier on a dataset consisting of 8981 photos and videos. One of the main drawbacks associated with this approach is that it still requires manual feature extraction and the best possible configuration, dimensionality reduction before model training and high complexity due to a large number of algorithms to generate the code words. Another approach explored in this study is the analysis of motion and audio features to enhance the detection of videos with sensitive content, as shown in Section 6. In order to improve the performance of classifiers, several researchers started to use features such as motion vectors and audio periodicity to represent motion. The algorithms employed in this work include a thresholding setup or ML-based algorithms such as Bayes or SVM. Even though this strategy implies the analysis of supplementary information to the conventional visual features, there are promising results (95.44% accuracy) according to [21]. Recently, the authors of [93] used audio periodicity in their multimodal approach and obtained a TPR of 96.7% and an SVM classifier. These strategies did not use public datasets; thus, it is difficult to evaluate their results, besides having to extract the features manually, as in the aforementioned strategies.

In recent years, several different DL architectures have been developed in this field, given their recent success in image and video analyses in other tasks. In this line of investigation, all strategies employ CNN models for the keyframe feature extraction stage, although recently, with the explosion of transformers in NLP, attention mechanisms have been used for high-level feature extraction. The attention mechanism provided by transformers together with CNN help to model long-range dependencies without compromising computational and statistical efficiency. Regarding the final classification algorithm, we can discuss two different approaches: those using LSTM recurrent neural networks, which is a DL model widely used to analyse temporal sequences, and those using the softmax function.

The best performance was by Chen et al. [34], who achieved an accuracy of 98.42% on their own dataset composed of 1 million images using modified VGG19 and visual attention. As for the best result obtained on one of the most commonly used datasets, [78], it was recorded by Gautam et al. [118], outperforming Perez et al. [30] (accuracy: 98.25%). On the other hand, the dataset in [26] as well as Gautam et al.'s dataset [118], with an accuracy of 97.15%, both outperform Ganwar et al.'s dataset [35] (accuracy: 97.1) by using a combination of CNNs (ResNet-18, sequence classifier ConvNets, faster RCNN-inception and ResNet-V2). The best-performing DL architectures were combinations of CNNs such as ResNet and Inception, plus the addition of attention mechanisms. DL-based strategies work well with large datasets, but if these are limited, other strategies such as transfer learning could be used to enhance the results.

To highlight the growth of artificial intelligence and DL in particular, and based on the keywords of the principles referenced in this work (e.g., pornography, sexual content, obscene content, deep learning, neural networks and convolutional neural networks), Figure 11 below shows the number of times the keywords are used. As can be seen, since approximately 2017, the number of references to the above-mentioned keywords in the literature has been increasing year by year.

Despite the popularity of transformers, the increase in performance in sensitive-content detection has been small compared to recent advances in NLP, such as in GPT3 or BERT. Visual transformers (VT) do not appear to outperform CNNs by a wide margin at the moment. In this regard, recent works such as [33,35], use transfer learning to initialise it with the weights being pre-trained on ImageNet.

Another disadvantage of taking into account DL models is the lack of interpretability, considered a black box in most of their implementations. This makes it impossible to detect possible biases that could be transferred from the data to the classifiers. Current interpretability methods in other domains, such as in health care and in those applicable to sensitive-content detection in adults, are suggested for the examination of DL-based systems.

In view of our review of the existing literature on learning strategies for sensitive-content detection techniques, the following section identifies the main challenges and open issues.

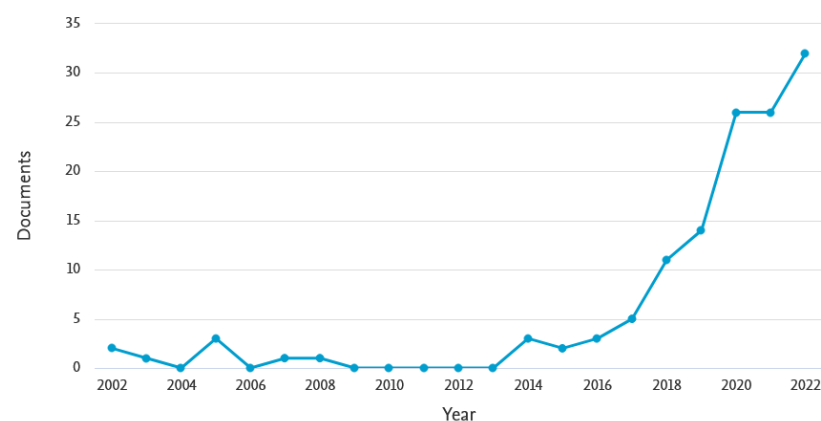


Figure 11. Documents per year referencing keywords related to this field.

8.2. Challenges and Open Issues

Based on the strategies and results obtained by the different works described in the previous sections, the challenges and open problems that can be identified in this review are summarised below:

- Given the success of deep learning methodologies, research on combinations of supervised, unsupervised and, more recently, self-supervised deep network architectures is expected to continue in order to achieve the most optimal configuration in the field of sensitive-content detection.
- Studying other variations of deep networks, such as variational autoencoders [131], capsule network [132] architectures or focal modulation networks [133], and other learning methods such as self-supervised or semi-supervised, should be further explored for this field of application.
- Contrarily, there is an anticipation that additional video features, including the more recent audio features, alongside static (keyframes) and dynamic (motion vectors) features, will be assessed in this domain, alongside the outcomes of deep learning methodologies. Within this framework, it is crucial to evaluate the various approaches using the same dataset to ensure an impartial assessment of the aforementioned strategies.

- The use of audio features should be further explored if the number of false negatives can be reduced, either with spectrograms or with transcription to text (e.g., whisper [134]) and subsequent classification with NLP techniques.
- For textual features, the main problem is the frequency with which the search pattern for keywords and filenames needs to be updated. Using only textual features is a challenge, as it would require more frequent re-training of the built models than models based on visual features. However, these features can be incorporated together with visual and auditory features to improve the final ranking.
- In the realm of current state-of-the-art attention mechanisms and approaches, models are designed to consider both global (ViT) and local (CNN) contexts, which play a crucial role in identifying the difficulty of detecting certain images and videos with and without sexual content that may be ambiguous. Consequently, the latest CNN and ViT-based architectures can assign a higher pornographic score to images featuring semi-naked individuals within a context that suggests sexual interest, such as erotic or provocative poses. In contrast, safe images containing semi-nude individuals, such as a girl in a bikini or boys in swimming costumes, receive a low score. However, when it comes to images with sensitive content where individuals are clothed or show minimal skin exposure, and where body exposure is partial or no genitalia are depicted, automatic evaluation systems tend to falter. In contrast, humans find it relatively easy to discern pornographic context, often due to facial expressions.
- As mentioned above, most of the strategies perform well in detecting sensitive content but fail in certain cases. This could be because the number of such images labelled as pornographic in the training set is very low. In addition to focusing on improving the architecture of the neural networks, it is important to have a robust dataset and to perform the relevant pre-processing correctly so that the proposed models can generalise successfully. For this, state-of-the-art image-generation models (text-to-image or image-to-image) such as stable diffusion could be used to improve the dataset [135].
- To the previous point, one of the main challenges in the context of sensitive-content detection is to create a large dataset labelled by experts that is as heterogeneous as possible (different categories of sexual images, poses, etc.), taking into account the diversity of ethnicities, genders, etc., which serves to evaluate and compare the different DL models created by the scientific community in a satisfactory way.
- Recent research on semantic analysis, such as object and background detection in videos, has yielded excellent results. In this regard, it would be interesting to detect patterns between backgrounds and objects in scenes with sexual content, aiming not only to improve detection performance by avoiding false negatives in scenes with minimal nudity but also to create a database that provides more context for CSAM/CSEM video scenes.
- Performing an analysis of the robustness of the built model (e.g., against adversarial attacks) and, whenever feasible, employing algorithms to analyse the explainability of the model will facilitate in understanding the decision-making process of the black-box algorithm and will enable evaluation of the model to enhance its performance and to identify areas where it may be failing.
- As the DL-based models developed detect new sensitive content, they should automatically generate hashes of newly detected explicit content and other textual and contextual features to update international databases, so that major NGOs can check available material against the hashes and other extracted features. To do this, researchers, organisations and Big Tech that have access to the databases must agree to move in the same direction in order to increase the success rate.

9. Conclusions and Future Work

This study summarises, analyses and interprets the state of the art of achievements and challenges of sensitive-content research. In this sense, this paper reviews the main algorithms for detecting sensitive content in images and videos. Due to the large volume of pornographic content accessible on the web and the increase in child pornography exploitation cases, effective automatic detection of sensitive content in multimedia databases has emerged as a critical issue in forensic analysis.

As described in the previous sections, different strategies have been considered to deal with this problem. In the context of text features, the results are promising, but searches for files, metadata and keywords are very domain knowledge dependent. In addition, it is necessary to pre-process the data in a manual format by updating the keywords and names with which offenders name the CSA files. Since all the papers used different datasets, it is not possible to compare them on a specific dataset. One of the papers by Pereira et al. [59] achieved a 96.8% accuracy rate using a dataset comprising 1,010,000 file paths. This work utilised adversarial attack techniques to analyse the model's robustness and employed a real dataset obtained from Project VIC International for model development.

In the exploration of visual features, most detection strategies focus on a thresholding approach to skin-colour features. The dissimilarities consist of the colour space used for the analysis. In this context, some work also uses texture and shape morphology, with good results. As for the image descriptors, most of the researchers used the same [78] dataset. In this context, for the 800 datasets, the best result was in [21], reaching 92.4% accuracy. In the case of the 2000 dataset [26], Moreira et al. [26] obtained 95% accuracy. The works that employed this strategy mostly used SVM by changing only the kernel structure while combining various types of features as input to the SVM by changing only the kernel structure, except for [23], which used a KNN classifier. The downside of this type of algorithm is the manual extraction of a certain feature and the application of different thresholds for classification. The main advantage of these algorithms with respect to DL models is the degree of explainability, as these algorithms are white boxes and are therefore easier to understand the decisions that are taken.

With regard to works that use motion, audio and multimodal analyses, despite good results, it is difficult to compare the performance between the different methods due to the large differences in the dataset used.

Strategies using DL techniques, compared to the conventional approaches described above, detect videos with sensitive content with higher performance. In particular, the research developed by Gautam et al. [118] shows the best performance thus far for a CNN-based system in the domain of sensitive-content detection. In this sense, most of the works that have employed DL as a classification strategy used the same dataset, which makes it easier to compare all research works.

Finally, although the performance of the latest algorithms is quite high for automatically detecting sensitive content in videos, further actions should be made to classify child pornography and to give feedback on the tools created by researchers from LEAs. Automatic classification of sensitive content is a challenge for the scientific community, and in particular CSAM, due to the inaccessibility of the data, which are private and exclusively held by law enforcement. In particular, CSAM detection is a major challenge for researchers and organisations due to the inaccessibility of the data, as they are private and sensitive and in the restricted possession of law enforcement. To help stakeholders extract information from the hidden data and to securely provide further insight into CSAM images, Laranjeira et al. [136] provided an analytical template that goes beyond the statistics of the dataset and its corresponding labels. They extracted automatic signals, provided both by pre-trained machine learning models, e.g., pornography detection and object categories, and by image features such as luminance and sharpness. Only added statistics were provided to ensure the anonymity of the victims, which is one of the key applications of these results.

In addition, many of the reviewed works focused on improving and fine-tuning deep learning architectures rather than emphasising the data (data-centric). Furthermore, the

majority of these works focused on improving accuracy on a dataset (unfortunately, a relatively outdated one) and tended to overfit the model to that particular dataset.

In lack of an appropriate and robust dataset, tools should be implemented that combine all possible detection algorithms, either through the analysis of content (visual and audio features) or through textual features (keywords, metadata and hashes), in order to have a set of tools with which to cover all the limitations of each of the separate sensitive-content detection strategies.

Author Contributions: Conceptualization, D.P.Á., A.L.S.O., J.P.G.-M. and L.J.G.V.; methodology, D.P.Á., A.L.S.O., J.P.G.-M. and L.J.G.V.; investigation, D.P.Á., A.L.S.O., J.P.G.-M. and L.J.G.V.; writing—original draft preparation, D.P.Á., A.L.S.O., J.P.G.-M. and L.J.G.V.; writing—review and editing, D.P.Á., A.L.S.O., J.P.G.-M. and L.J.G.V. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the HEROES project (<http://heroes-fct.eu>, accessed on 4 February 2023). This project has received funding from the European Union’s Horizon 2020 research and innovation programme under Grant agreement No.101021801. The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ramaswamy, S.; Seshadri, S. Children on the brink: Risks for child protection, sexual abuse, and related mental health problems in the COVID-19 pandemic. *Indian J. Psychiatry* **2020**, *62*, S404. [\[CrossRef\]](#)
2. Europol. *Internet Organised Crime Threat Assessment (IOCTA) 2021*; Publications Office of the European Union: Luxembourg, 2021; p. 12. [\[CrossRef\]](#)
3. Lee, H.E.; Ermakova, T.; Ververis, V.; Fabian, B. Detecting child sexual abuse material: A comprehensive survey. *Forensic Sci. Int. Digit. Investig.* **2020**, *34*, 301022. [\[CrossRef\]](#)
4. Khaksar Pour, A.; Chaw Seng, W.; Palaiahnakote, S.; Tahaei, H.; Anuar, N.B. A survey on video content rating: Taxonomy, challenges and open issues. *Multimed. Tools Appl.* **2021**, *80*, 24121–24145. [\[CrossRef\]](#)
5. Cifuentes, J.; Sandoval Orozco, A.L.; García Villalba, L.J. A survey of artificial intelligence strategies for automatic detection of sexually explicit videos. *Multimed. Tools Appl.* **2022**, *81*, 3205–3222. [\[CrossRef\]](#)
6. PhotoDNA. Microsoft. Available online: <https://www.microsoft.com/en-us/photodna> (accessed on 17 December 2022).
7. Canadian Centre for Children Protection. Project Arachnid. 2022. Available online: <https://projectarachnid.ca/> (accessed on 19 December 2022).
8. Media Detective—Software to Detect and Remove Adult Material on Your Home Computer. Available online: <https://www.mediadetective.com/> (accessed on 5 January 2023).
9. Hyperdyne Software—Detect and Remove Adult Files with Snitch Porn Cleaner. Available online: <https://hyperdynesoftware.com/> (accessed on 5 January 2023).
10. Thorn Research: Understanding Sexually Explicit Images, Self-Produced by Children. Available online: <https://www.thorn.org/blog/thorn-research-understanding-sexually-explicit-images-self-produced-by-children/> (accessed on 4 January 2023).
11. NetClean. Bright Technology for a Brighter Future. Available online: <https://www.netclean.com/> (accessed on 4 January 2023).
12. Choi, B.; Han, S.; Chung, B.; Ryou, J. Human body parts candidate segmentation using laws texture energy measures with skin colour. In Proceedings of the International Conference on Advanced Communication Technology, Gangwon-Do, Republic of Korea, 3–16 February 2011; pp. 556–560.
13. Polastro, M.D.C.; Eleuterio, P.M.D.S. A statistical approach for identifying videos of child pornography at crime scenes. In Proceedings of the 7th International Conference on Availability, Reliability and Security, Prague, Czech Republic, 20–24 August 2012; pp. 604–612. [\[CrossRef\]](#)
14. Lee, H.; Lee, S.; Nam, T. Implementation of high performance objectionable video classification system. In Proceedings of the 8th International Conference Advanced Communication Technology, Gangwon-Do, Republic of Korea, 20–22 February 2006; Volume 2, pp. 959–962. [\[CrossRef\]](#)
15. Beyer, L.; Izmailov, P.; Kolesnikov, A.; Caron, M.; Kornblith, S.; Zhai, X.; Minderer, M.; Tschannen, M.; Alabdulmohsin, I.; Pavetic, F. FlexiViT: One Model for All Patch Sizes. *arXiv* **2022**, arXiv:2212.08013.
16. Eleuterio, P.; Polastro, M. An adaptive sampling strategy for automatic detection of child pornographic videos. In Proceedings of the Seventh International Conference on Forensic Computer Science, Brasília, Brazil, 26–28 September 2012; pp. 12–19. [\[CrossRef\]](#)
17. Wang, D.; Zhu, M.; Yuan, X.; Qian, H. Identification and annotation of erotic film based on content analysis. *Electron. Imaging Multimed. Technol. IV* **2005**, 5637, 88. [\[CrossRef\]](#)

18. Ulges, A.; Stahl, A. Automatic detection of child pornography using colour visual words. In Proceedings of the IEEE International Conference on Multimedia and Expo, Barcelona, Spain, 11–15 July 2011; pp. 3–8. [\[CrossRef\]](#)
19. Garcia, M.B.; Revano, T.F.; Habal, B.G.M.; Contreras, J.O.; Enriquez, J.B.R. A Pornographic Image and Video Filtering Application Using Optimized Nudity Recognition and Detection Algorithm. In Proceedings of the 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management, Baguio City, Philippines, 29 November–2 December 2018; pp. 1–4, ISBN 978-1-5386-7767-4. [\[CrossRef\]](#)
20. Caetano, C.; Avila, S.; Guimarães, S.; De Araújo, A.A. Pornography detection using BossaNova video descriptor. In Proceedings of the European Signal Processing Conference, Lisbon, Portugal, 1–5 September 2014; pp. 1681–1685.
21. Caetano, C.; Avila, S.; Schwartz, W.R.; Guimarães, S.J.F.; Araújo, A.D.A. A mid-level video representation based on binary descriptors: A case study for pornography detection. *Neurocomputing* **2016**, *213*, 102–114. [\[CrossRef\]](#)
22. Lopes, A.P.B.; De Avila, S.E.; Peixoto, A.N.; Oliveira, R.S.; Coelho, M.D.M.; Araújo, A.D.A. Nude detection in video using bag-of-visual-features. In Proceedings of the 22nd Brazilian Symposium on Computer Graphics and Image Processing, Rio de Janeiro, Brazil, 11–15 October 2009; pp. 224–231. [\[CrossRef\]](#)
23. Hartatik.; Setyanto, A.; Kusri, K.; Made Artha Agastya, I. Comparison of SIFT and SURF methods for porn image detection. In Proceedings of the 4th International Conference on Information Technology, Information Systems and Electrical Engineering, Yogyakarta, Indonesia, 20–21 November 2019; pp. 281–285. [\[CrossRef\]](#)
24. Tian, C.; Zhang, X.; Wei, W.; Gao, X. Colour pornographic image detection based on colour-saliency preserved mixture deformable part model. *Multimed. Tools Appl.* **2018**, *77*, 6629–6645. [\[CrossRef\]](#)
25. Wang, H.; Schmid, C. Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3551–3558. [\[CrossRef\]](#)
26. Moreira, D.; Avila, S.; Perez, M.; Moraes, D.; Testoni, V.; Valle, E.; Goldenstein, S.; Rocha, A. Pornography classification: The hidden clues in video space–time. *Forensic Sci. Int.* **2016**, *268*, 46–61. [\[CrossRef\]](#)
27. Moreira, D.; Avila, S.; Perez, M.; Moraes, D.; Testoni, V.; Valle, E.; Goldenstein, S.; Rocha, A. Multimodal data fusion for sensitive scene localization. *Inf. Fusion* **2019**, *45*, 307–323. [\[CrossRef\]](#)
28. Da Silva, M.V.; Marana, A.N. Spatiotemporal CNNs for pornography detection in videos. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Cham, Switzerland, 2019; Volume 11401, pp. 547–555. [\[CrossRef\]](#)
29. Moustafa, M. Applying deep learning to classify pornographic images and videos. *arXiv* **2015**, arXiv:1511.08899.
30. Perez, M.; Avila, S.; Moreira, D.; Moraes, D.; Testoni, V.; Valle, E.; Goldenstein, S.; Rocha, A. Video pornography detection through deep learning techniques and motion information. *Neurocomputing* **2017**, *230*, 279–293. [\[CrossRef\]](#)
31. Mallmann, J.; Santin, A.O.; Viegas, E.K.; dos Santos, R.R.; Geremias, J. PPCensor: Architecture for real-time pornography detection in video streaming. *Future Gener. Comput. Syst.* **2020**, *112*, 945–955. [\[CrossRef\]](#)
32. Song, K.; Kim, Y.S. An enhanced multimodal stacking scheme for online pornographic content detection. *Appl. Sci.* **2020**, *10*, 2943. [\[CrossRef\]](#)
33. Fu, Z.; Li, J.; Chen, G.; Yu, T.; Deng, T. PornNet: A unified deep architecture for pornographic video recognition. *Appl. Sci.* **2021**, *11*, 3066. [\[CrossRef\]](#)
34. Chen, J.; Liang, G.; He, W.; Xu, C.; Yang, J.; Liu, R. A Pornographic Images Recognition Model based on Deep One-Class Classification With Visual Attention Mechanism. *IEEE Access* **2020**, *8*, 122709–122721. [\[CrossRef\]](#)
35. Gangwar, A.; González-Castro, V.; Alegre, E.; Fidalgo, E. AttM-CNN: Attention and metric learning based CNN for pornography, age and Child Sexual Abuse (CSA) Detection in images. *Neurocomputing* **2021**, *445*, 81–104. [\[CrossRef\]](#)
36. Westlake, B.; Bouchard, M.; Frank, R. Comparing methods for detecting child exploitation content online. In Proceedings of the European Intelligence and Security Informatics Conference, Odense, Denmark, 22–24 August 2012; pp. 156–163. [\[CrossRef\]](#)
37. Stallings, W. *Cryptography and Network Security*; Pearson: London, UK, 2017; p. 767, ISBN 978-1-2921-5858-7.
38. Farid, H. An Overview of Perceptual Hashing. *J. Online Trust. Saf.* **2021**, *36*, 1405. [\[CrossRef\]](#)
39. Liong, V.E.; Lu, J.; Wang, G.; Moulin, P.; Zhou, J. Deep hashing for compact binary codes learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–15 June 2015; pp. 2475–2483. [\[CrossRef\]](#)
40. Zhao, F.; Huang, Y.; Wang, L.; Tan, T. Deep Semantic Ranking Based Hashing for Multi-Label Image Retrieval. *arXiv* **2015**, arXiv:1501.06272.
41. Liu, H.; Wang, R.; Shan, S.; Chen, X. Deep Supervised Hashing for Fast Image Retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2064–2072.
42. Wu, D.; Lin, Z.; Li, B.; Ye, M.; Wang, W. Deep Supervised Hashing for Multi-Label and Large-Scale Image Retrieval. In Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval ICMR '17, Bucharest, Romania, 6–9 June 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 150–158. [\[CrossRef\]](#)
43. Wang, H.; Yao, M.; Jiang, G.; Mi, Z.; Fu, X. Graph-Collaborated Auto-Encoder Hashing for Multi-view Binary Clustering. *arXiv* **2023**, arXiv:2301.02484.
44. Jiang, C.; Pang, Y. Perceptual image hashing based on a deep convolution neural network for content authentication. *J. Electron. Imaging* **2018**, *27*, 043055. [\[CrossRef\]](#)
45. Facebook. Open-Sourcing Photo- and Video-Matching Technology to Make the Internet Safer | Meta. Available online: <https://about.fb.com/news/2019/08/open-source-photo-video-matching/> (accessed on 17 January 2023).

46. Google. Content Safety API. Available online: <https://protectingchildren.google/tools-for-partners/#learn-about-our-tools> (accessed on 17 January 2023).
47. Apple. CSAM Detection. 2021. Available online: https://www.apple.com/child-safety/pdf/CSAM_Detection_Technical_Summary.pdf (accessed on 17 January 2023).
48. iCOP: Live forensics to reveal previously unknown criminal media on P2P networks. *Digit. Investig.* **2016**, *18*, 50–64. [CrossRef]
49. Krawetz, N. PhotoDNA and Limitations—The Hacker Factor Blog. Available online: <https://www.hackerfactor.com/blog/index.php?archives/931-PhotoDNA-and-Limitations.html> (accessed on 4 February 2023).
50. Steel, C.M. Child pornography in peer-to-peer networks. *Child Abus. Negl.* **2009**, *33*, 560–568. [CrossRef]
51. Westlake, B.; Bouchard, M.; Frank, R. Assessing the Validity of Automated Webcrawlers as Data Collection Tools to Investigate Online Child Sexual Exploitation. *Sex. Abus. J. Res. Treat.* **2017**, *29*, 685–708. [CrossRef]
52. Panchenko, A.; Beaufort, R.; Fairon, C. Detection of child sexual abuse media on p2p networks: Normalization and classification of associated filenames. In Proceedings of the LREC Workshop on Language Resources for Public Security Applications, Istanbul, Turkey, 27 May 2012; pp. 27–31.
53. Polastro, M.D.C.; Da Silva Eleuterio, P.M. NuDetective: A forensic tool to help combat child pornography through automatic nudity detection. In Proceedings of the Workshops on Database and Expert Systems Applications, Bilbao, Spain, 30 August 2010; pp. 349–353. [CrossRef]
54. Peersman, C.; Schulze, C.; Rashid, A.; Brennan, M.; Fischer, C. ICOP: Automatically identifying new child abuse media in P2P networks. In Proceedings of the IEEE Security and Privacy Workshops, San Jose, CA, USA, 18–21 May 2014; IEEE Computer Society, 1730 Massachusetts Ave., NW: Washington, DC, USA, 2014; pp. 124–131. [CrossRef]
55. Gov.UK. New AI Technique to Block Online Child Grooming Launched—GOV.UK. Available online: <https://www.gov.uk/government/news/new-ai-technique-to-block-online-child-grooming-launched> (accessed on 4 December 2022).
56. Al-Nabki, M.W.; Fidalgo, E.; Alegre, E.; Aláiz-Rodríguez, R. File name classification approach to identify child sexual abuse. In Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods ICPRAM, Valletta, Malta, 22–24 February 2020; pp. 228–234. [CrossRef]
57. Aldahoul, N.; Karim, H.A.; Abdullah, M.H.L.; Fauzi, M.F.A.; Ba Wazir, A.S.; Mansor, S.; See, J. Exploring high-level features for detecting cyberpedophilia. *Comput. Speech Lang.* **2014**, *28*, 108–120. [CrossRef]
58. Peersman, C. Detecting Deceptive Behaviour in the Wild: Text Mining for Online Child Protection in the Presence of Noisy and Adversarial Social Media Communications. Ph.D. Thesis, Lancaster University, Lancaster, UK, 2018. [CrossRef]
59. Pereira, M.; Dodhia, R.; Anderson, H.; Brown, R. Metadata-Based Detection of Child Sexual Abuse Material. *arXiv* **2020**, arXiv:2010.02387.
60. Perverted Justice Foundation. The Largest and Best Anti-Predator Organization Online. Available online: Perverted-Justice.com (accessed on 22 February 2023).
61. Carlsson, A.; Eriksson, A.; Isik, M. Automatic Detection of Images Containing Nudity. Ph.D. Thesis, IT University of Goteborg, Gothenburg, Sweden, 2008.
62. Fleck, M.M.; Forsyth, D.A.; Bregler, C. Finding naked people. In Proceedings of the 4th European Conference on Computer Vision, Cambridge, UK, 14–18 April 1996; Volume 1065, pp. 594–602. [CrossRef]
63. Jones, M.J.; Reh, J.M. Statistical colour models with application to skin detection. *Int. J. Comput. Vis.* **2002**, *46*, 81–96. [CrossRef]
64. Kakumanu, P.; Makrogiannis, S.; Bourbakis, N. A survey of skin-colour modeling and detection methods. *Pattern Recognit.* **2007**, *40*, 1106–1122. [CrossRef]
65. Platzer, C.; Stuetz, M.; Lindorfer, M. Skin sheriff: A machine learning solution for detecting explicit images. In Proceedings of the 2nd International Workshop on Security and Forensics in Communication Systems, Kyoto, Japan, 3 June 2014; Association for Computing Machinery: New York, NY, USA, 2014; pp. 45–55. [CrossRef]
66. Ap-apid, R. An Algorithm for Nudity Detection. In Proceedings of the 5th Philippine Computing Science Congress, University of Cebu (Banilad Campus), Cebu City, Philippines, 4–5 March 2005; pp. 201–205.
67. Ozinov, F. GitHub—Bakwc/PornDetector: Porn Images Detector with Python, Tensorflow, Scikit-Learn and Opencv. Available online: <https://github.com/bakwc/PornDetector> (accessed on 12 December 2022).
68. Zhuo, L.; Geng, Z.; Zhang, J.; Guang Li, X. ORB feature based web pornographic image recognition. *Neurocomputing* **2016**, *173*, 511–517. [CrossRef]
69. Lillie, O. PHP Video. 2017. Available online: <https://github.com/buggedcom/phpvideotoolkit-v2> (accessed on 27 December 2022).
70. Zhu, M.-L. Video stream segmentation method based on video page. *J. Comput. Aided Design. Comput. Graph.* **2000**, *12*, 585–589.
71. El-Hallak, M.; Lovell, D. ORB an efficient. *Arthritis Rheum.* **2013**, *65*, 2736. [PubMed]
72. Jansohn, C.; Ulges, A.; Breuel, T.M. Detecting pornographic video content by combining image features with motion information. In Proceedings of the ACM Multimedia Conference, with Co-located Workshops and Symposia, Beijing, China, 19–24 October 2009; pp. 601–604. [CrossRef]
73. Deselaers, T.; Pimenidis, L.; Ney, H. Bag-of-visual-words models for adult image classification and filtering. In Proceedings of the International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008.
74. Zhang, J.; Sui, L.; Zhuo, L.; Li, Z.; Yang, Y. An approach of bag-of-words based on visual attention model for pornographic images recognition in compressed domain. *Neurocomputing* **2013**, *110*, 145–152. [CrossRef]

75. Valle, E.; de Avila, S.; da Luz, A.; de Souza, F.; Coelho, M.; Araújo, A. Content-Based Filtering for Video Sharing Social Networks. *arXiv* **2011**, arXiv:1101.2427.
76. Laptev, I. On space-time interest points. *Int. J. Comput. Vis.* **2005**, *64*, 107–123. [[CrossRef](#)]
77. Avila, S.; Thome, N.; Cord, M.; Valle, E.; De Araújo, A. BOSSA: Extended bow formalism for image classification. In Proceedings of the International Conference on Image Processing, Brussels, Belgium, 11–14 September 2011; pp. 2909–2912. [[CrossRef](#)]
78. Avila, S.; Thome, N.; Cord, M.; Valle, E.; De Araújo, A. Pooling in image representation: The visual codeword point of view. *Comput. Vis. Image Underst.* **2013**, *117*, 453–465. [[CrossRef](#)]
79. Kim, A. NSFW Dataset. 2019. Available online: https://github.com/alex000kim/nsfw_data_scraper (accessed on 22 November 2022).
80. Chen, Y.; Zheng, R.; Zhou, A.; Liao, S.; Liu, L. Automatic detection of pornographic and gambling websites based on visual and textual content using a decision mechanism. *Sensors* **2020**, *20*, 3989. [[CrossRef](#)]
81. Souza, F.; Valle, E.; Camara-Chavez, G.; De Araujo, A. An Evaluation on Colour Invariant Based Local Spatiotemporal Features for Action Recognition. In Proceedings of the Conference on Graphics, Patterns and Images, Ouro Preto, Brazil, 22–25 August 2012; pp. 1–6.
82. Rea, N.; Lacey, G.; Lambe, C.; Dahyot, R. Multimodal periodicity analysis for illicit content detection in videos. In Proceedings of the IET Conference Publications, Leela Palace, Bangalore, India, 26–28 September 2006; pp. 106–114. [[CrossRef](#)]
83. Zuo, H.; Wu, O.; Hu, W.; Xu, B. Recognition of blue movies by fusion of audio and video. In Proceedings of the IEEE International Conference on Multimedia and Expo, Hannover, Germany, 23–26 June 2008; pp. 37–40.
84. Liu, Y.; Yang, Y.; Xie, H.; Tang, S. Fusing audio vocabulary with visual features for pornographic video detection. *Future Gener. Comput. Syst.* **2014**, *31*, 69–76. [[CrossRef](#)]
85. Kim, C.Y.; Kwon, O.J.; Kim, W.G.; Choi, S.R. Automatic System for Filtering Obscene Video. In Proceedings of the 10th International Conference on Advanced Communication Technology, Phoenix Park, Korea, 17–20 February 2008; Volume 2; pp. 1435–1438. [[CrossRef](#)]
86. Wang, J.Z.; Li, J.; Wiederhold, G.; Firschein, O. System for screening objectionable images. *Comput. Commun.* **1998**, *21*, 1355–1360. [[CrossRef](#)]
87. Endeshaw, T.; Garcia, J.; Jakobsson, A. Classification of indecent videos by low complexity repetitive motion detection. In Proceedings of the Applied Imagery Pattern Recognition Workshop, Washington, DC, USA, 15–17 October 2008. [[CrossRef](#)]
88. Qu, Z.; Liu, Y.; Liu, Y.; Jiu, K.; Chen, Y. A method for reciprocating motion detection in porn video based on motion features. In Proceedings of the 2nd IEEE International Conference on Broadband Network and Multimedia Technology, Beijing, China, 18–20 October 2009; pp. 183–187. [[CrossRef](#)]
89. Ulges, A.; Schulze, C.; Borth, D.; Stahl, A. Pornography detection in video benefits (a lot) from a multi-modal approach. In Proceedings of the 2012 ACM Workshop on Audio and Multimedia Methods for Large-Scale Video Analysis, Nara, Japan, 2 November 2012; pp. 21–26. [[CrossRef](#)]
90. Behrad, A.; Salehpour, M.; Ghaderian, M.; Saiedi, M.; Barati, M.N. Content-based obscene video recognition by combining 3D spatiotemporal and motion-based features. *Eurasip J. Image Video Process.* **2012**, *2012*, 1. [[CrossRef](#)]
91. Jung, S.; Yoon, J.; Sull, S. A real-time system for detecting indecent videos based on spatiotemporal patterns. *IEEE Trans. Consum. Electron.* **2014**, *60*, 696–701. [[CrossRef](#)]
92. Schulze, C.; Henter, D.; Borth, D.; Dengel, A. Automatic detection of CSA media by multi-modal feature fusion for law enforcement support. In Proceedings of the ACM International Conference on Multimedia Retrieval, Glasgow, UK, 1–4 April 2014; Association for Computing Machinery: New York, NY, USA, 2014; pp. 353–360. [[CrossRef](#)]
93. Liu, Y.; Gu, X.; Huang, L.; Ouyang, J.; Liao, M.; Wu, L. Analyzing periodicity and saliency for adult video detection. *Multimed. Tools Appl.* **2020**, *79*, 4729–4745. [[CrossRef](#)]
94. Mahadeokar, J.; Pesavento, G. Open Sourcing a deep learning Solution for Detecting NSFW Images. *Yahoo Eng.* **2016**, *24*, 2018.
95. Nian, F.; Li, T.; Wang, Y.; Xu, M.; Wu, J. Pornographic image detection utilizing deep convolutional neural networks. *Neurocomputing* **2016**, *210*, 283–293. [[CrossRef](#)]
96. Vitorino, P.; Avila, S.; Perez, M.; Rocha, A. Leveraging deep neural networks to fight child pornography in the age of social media. *J. Vis. Commun. Image Represent.* **2018**, *50*, 303–313. [[CrossRef](#)]
97. Wang, Y.; Jin, X.; Tan, X. Pornographic image recognition by strongly-supervised deep multiple instance learning. In Proceedings of the International Conference on Image Processing, Phoenix, AZ, USA, 25–28 September 2016; Volume 2016, pp. 4418–4422. [[CrossRef](#)]
98. Xu, W.; Parvin, H.; Izadparast, H. Deep learning Neural Network for Unconventional Images Classification. *Neural Process. Lett.* **2020**, *52*, 169–185. [[CrossRef](#)]
99. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–15 June 2015; pp. 1–9. [[CrossRef](#)]
100. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [[CrossRef](#)]
101. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [[CrossRef](#)]

102. Lecun, Y.; Bottou, L.; Bengio, Y.; Ha, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324.
103. Wehrmann, J.; Simões, G.S.; Barros, R.C.; Cavalcante, V.F. Adult content detection in videos with convolutional and recurrent neural networks. *Neurocomputing* **2018**, *272*, 432–438. [\[CrossRef\]](#)
104. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [\[CrossRef\]](#)
105. Song, K.H.; Kim, Y.S. Pornographic video detection scheme using multimodal features. *J. Eng. Appl. Sci.* **2018**, *13*, 1174–1182. [\[CrossRef\]](#)
106. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; Volume 2015, pp. 4489–4497. [\[CrossRef\]](#)
107. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; Lecun, Y.; Paluri, M. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6450–6459. [\[CrossRef\]](#)
108. Singh, S.; Buduru, A.B.; Kaushal, R.; Kumaraguru, P. KidsGUARD: Fine grained approach for child unsafe video representation and detection. In Proceedings of the ACM Symposium on Applied Computing, Limassol, Cyprus, 8–12 April 2019; Volume F1477, pp. 2104–2111. [\[CrossRef\]](#)
109. Papadamou, K.; Papasavva, A.; Zannettou, S.; Blackburn, J.; Kourtellis, N.; Leontiadis, I.; Stringhini, G.; Sirivianos, M. Disturbed Youtube for kids: Characterizing and detecting inappropriate videos targeting young children. In Proceedings of the 14th International AAAI Conference on Web and Social Media, Atlanta, GA, USA, 8–11 June 2020; pp. 522–533. [\[CrossRef\]](#)
110. Chaves, D.; Fidalgo, E.; Alegre, E.; Alaiz-Rodríguez, R.; Jáñez-Martino, F.; Azzopardi, G. Assessment and Estimation of Face Detection Performance Based on Deep Learning for Forensic Applications. *Sensors* **2020**, *20*, 4491. [\[CrossRef\]](#)
111. Yang, S.; Luo, P.; Loy, C.C.; Tang, X. WIDER FACE: A Face Detection Benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
112. Nada, H.; Sindagi, V.A.; Zhang, H.; Patel, V.M. Pushing the Limits of Unconstrained Face Detection: A Challenge Dataset and Baseline Results. *arXiv* **2018**, arXiv:1804.10275.
113. Lee, G.; Kim, M. Deepfake Detection Using the Rate of Change between Frames Based on Computer Vision. *Sensors* **2021**, *21*, 7367. [\[CrossRef\]](#) [\[PubMed\]](#)
114. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Niessner, M. FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; Volume 2019, p. 1. [\[CrossRef\]](#)
115. Kaggle. Deepfake Detection Challenge. Kaggle. 2020. Available online: <https://www.kaggle.com/c/deepfake-detection-challenge> (accessed on 1 December 2022).
116. Aldahoul, N.; Karim, H.A.; Abdullah, M.H.L.; Fauzi, M.F.A.; Ba Wazir, A.S.; Mansor, S.; See, J. Transfer detection of yolo to focus cnn's attention on nude regions for adult content detection. *Symmetry* **2021**, *13*, 26. [\[CrossRef\]](#)
117. Lovenia, H.; Lestari, D.P.; Frieske, R. What Did i Just Hear? Detecting Pornographic Sounds in Adult Videos Using Neural Networks. In Proceedings of the ACM International Conference Proceeding Series, St. Pölten, Austria, 6–9 September 2022; Association for Computing Machinery: New York, NY, USA, 2022; Volume 1, pp. 92–95. [\[CrossRef\]](#)
118. Gautam, N.; Vishwakarma, D.K. Obscenity Detection in Videos through a Sequential ConvNet Pipeline Classifier. *IEEE Trans. Cogn. Dev. Syst.* **2022**, *8920*, 1–10. [\[CrossRef\]](#)
119. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *201*, 5999–6009.
120. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Volume 12346, pp. 213–229. [\[CrossRef\]](#)
121. Wu, B.; Xu, C.; Dai, X.; Wan, A.; Zhang, P.; Yan, Z.; Tomizuka, M.; Gonzalez, J.; Keutzer, K.; Vajda, P. Visual Transformers: Token-based Image Representation and Processing for Computer Vision. *arXiv* **2020**, arXiv:2006.03677.
122. Simoes, G.S.; Wehrmann, J.; Barros, R.C. Attention-based Adversarial Training for Seamless Nudity Censorship. In Proceedings of the International Joint Conference on Neural Networks, Budapest, Hungary, 14–19 July 2019; Volume 2019, pp. 1–8. [\[CrossRef\]](#)
123. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
124. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-Token ViT: Training vision transformers from Scratch on ImageNet. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 538–547. [\[CrossRef\]](#)
125. Chollet, F. Xception: Deep learning with Depthwise Separable Convolutions. *arXiv* **2016**, arXiv:1610.02357.
126. Lin, X.; Qin, F.; Peng, Y.; Shao, Y. Fine-grained pornographic image recognition with multiple feature fusion transfer learning. *Int. J. Mach. Learn. Cybern.* **2021**, *12*, 73–86. [\[CrossRef\]](#)
127. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 2261–2269. [\[CrossRef\]](#)

128. Yousaf, K.; Nawaz, T. A deep learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos. *IEEE Access* **2022**, *10*, 16283–16298. [CrossRef]
129. Google. Google Open Dataset. 2022. Available online: <https://datasetsearch.research.google.com/> (accessed on 25 August 2022).
130. Athalye, A. Ribosome: Synthesize Photos from PhotoDNA Using Machine Learning. 2021. Available online: <https://github.com/anishathalye/ribosome> (accessed on 12 December 2022).
131. Kingma, D.P.; Welling, M. An Introduction to Variational Autoencoders. *Found. Trends Mach. Learn.* **2019**, *12*, 307–392. [CrossRef]
132. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic Routing between Capsules. In Proceedings of the 31st International Conference on Neural Information Processing Systems NIPS’17, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 3859–3869.
133. Yang, J.; Li, C.; Dai, X.; Yuan, L.; Gao, J. Focal Modulation Networks. *arXiv* **2022**, arXiv:2203.11926.
134. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv* **2022**, arXiv:2212.04356.
135. Zhang, L.; Agrawala, M. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv* **2023**, arXiv:2302.05543.
136. Laranjeira, C.; Macedo, J.; Avila, S.; dos Santos, J.A. Seeing without Looking: Analysis Pipeline for Child Sexual Abuse Datasets. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022; Association for Computing Machinery: New York, NY, USA, 2022; Volume 1.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.