International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(4s) 2026 pp. 1139-1160
https://doi.org/10.70917/ijcisim-2026-2629
Article

A Review on Deep Transfer Learning Techniques for Detecting Abusive and Harmful Visual Content

Amita V. Shah[1]*, Hetal B. Pandya[2], Pooja K. Shah[3], Hardik M. Patel[4], Jigna Jadav[5]

[1] Gujarat Technological University, Ahmedabad, Gujarat, India.
Email: amitashah@ldce.ac.in
[2] Gujarat Technological University, Ahmedabad, Gujarat, India.
Email: hetalbpandya@ldce.ac.in
[3]Vidush Somany Institute of Technology and Research, Kadi Sarva Vishwavidyalaya, Gandhinagar, Gujarat, India.
Email: poojaz.2608@gmail.com
[4]Vidush Somany Institute of Technology and Research, Kadi Sarva Vishwavidyalaya, Gandhinagar, Gujarat, India.
Email: Profhmp9@gmail.com
[5]Vishwakarma Government Engineering College, Chandkheda, Ahmedabad, Gujarat, India.
Email: Jigna.j.jadav@gmail.com

Abstract: The rapid expansion of digital communication platforms has led to an unprecedented increase in the sharing of visual content, making the automatic identification of abusive and harmful images a critical research challenge. Harmful visual materials—including violent, hateful, explicit, and disturbing images—pose significant risks to online communities, emphasizing the need for intelligent and scalable content moderation systems. Deep transfer learning has emerged as a highly effective approach for this task by utilizing knowledge acquired from large-scale pre-trained models and adapting it to domain-specific datasets with limited labeled samples. This review presents a comprehensive analysis of recent deep transfer learning techniques employed for abusive visual content detection. It examines widely adopted architectures, including Convolutional Neural Networks (CNNs), EfficientNet, MobileNet, ResNet, Vision Transformers (ViTs), hybrid CNN–Transformer models, and multimodal vision–language frameworks. The survey also reviews benchmark datasets, transfer learning strategies, evaluation metrics, and comparative performance reported in recent studies. Furthermore, it discusses key challenges such as dataset scarcity, class imbalance, cultural bias, adversarial robustness, interpretability, computational complexity, and privacy concerns that continue to affect the reliability of automated moderation systems. Emerging research directions—including explainable artificial intelligence, few-shot and zero-shot learning, multimodal fusion, foundation models, and privacy-preserving learning—are also highlighted as promising solutions for developing more accurate, robust, and trustworthy content moderation frameworks. This review provides researchers and practitioners with a structured overview of current developments, identifies existing research gaps, and outlines future opportunities for advancing deep transfer learning techniques in the detection of abusive and harmful visual content.

Keywords: Deep Transfer Learning, Abusive Visual Content Detection, Computer Vision, Multimodal Learning, Convolutional Neural Networks (CNNs)

1. Introduction

The rapid growth of digital platforms has created an enormous flow of user-generated content, from billions of images uploaded each day across social networks, messaging applications, and online communities (e.g., forums). The associated growth of abusive and harmful visual content, e.g., violent, pornographic, abusive, or hate/violence inducing, presents a significant threat to online safety, particularly to children and vulnerable populations, and presents reputational, ethical, and legal risks for digital platforms. Manual moderation, where it exists, fails at scale and is slow, costly, and unable to respond quickly enough to the fast changes in content being created every second [1].

The significance of abusive content detection has thus emerged as both an important research and practical question. Whether via the dissemination of explicit and/or violent imagery, exposure could easily lead to psychological harm, normalize inappropriate behavior, or even lead to radicalization and/or hate [Information]. It may imply certainly that providers of platforms, who are unable to moderate abusive or inappropriate content, will suffer legal ramifications and lose user trust [2]. As a result, automatic detection systems are valuable in protecting its users for risk mitigation, improving the user experience, and complying with increasingly evolving regulatory controls. Recent survey's suggest that the prevalence of cyberbullying and harmful online experiences among teens is increasing. According to a Pew Research Center report, almost half of U.S. teens 13 - 17 years old have had at least one experience of cyberbullying, where offensive name-calling is the most common [7]. Findings from the Cyberbullying Research Center [Year] suggest that more than 54% of middle and high school students reported at least one experience of cyberbullying throughout their lifetime [8]. All these statistics puts urgency behind the social responsibility of abusing content detection processes and systems.

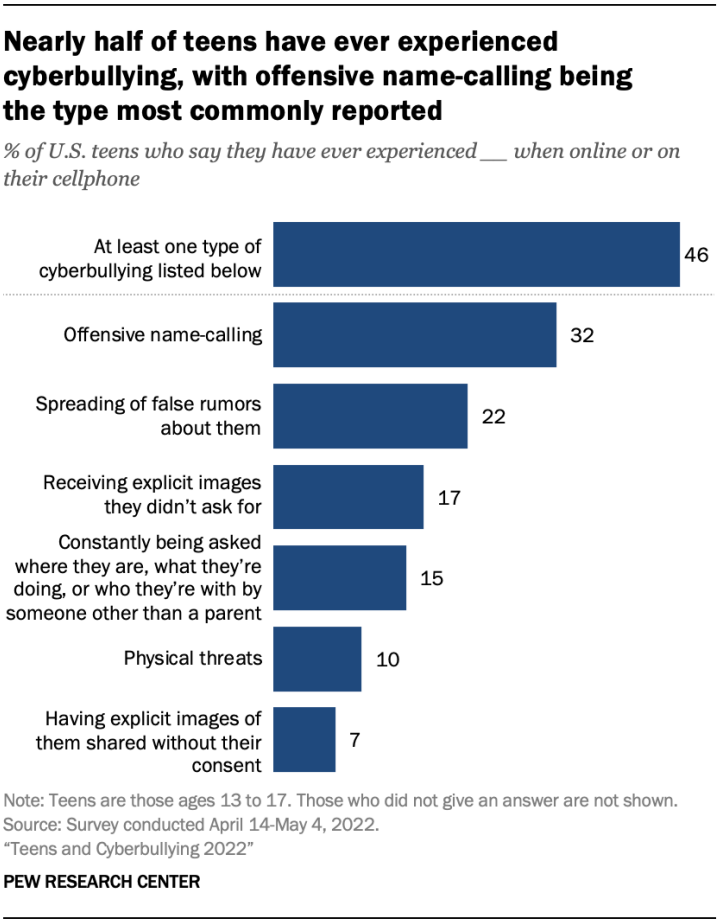

Figure 1. Percentage of U.S. teens (ages 13–17) who have experienced different types of cyberbullying, showing offensive name-calling as the most common form [7]

Deep learning has achieved remarkable strides in computer vision tasks such as segmentation, detection, and classification. Nevertheless, given the sensitive and limited nature of such datasets, training deep models from the ground-up requires large labeled datasets and computational power, which is seldom found with harmful content. However, deep transfer learning provides a powerful alternative by fine-tuning models pre-trained on large generic datasets such as ImageNet in order to optimize the model(s) for the specific task of abuse detection [4]. Deep Learning models such as ResNet, EfficientNet and MobileNet have been trained on these datasets and then fine-tuned on images of user-generated known abuse detection images, leading to faster convergence, less data, and better generalization of model(s). More recently, transformer-based models have also gained interest (along with hybrid CNN - Transformers) due to their improved contextual feature extraction capabilities and durability to complex variations across images [5].

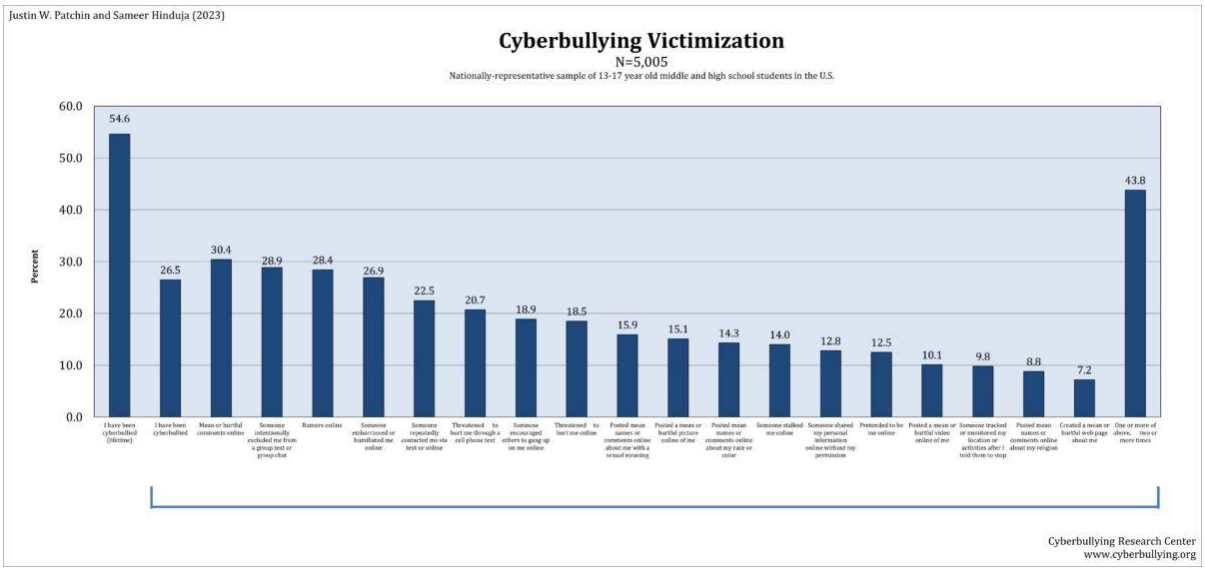

Figure 2. National survey results of middle and high school students in the U.S. reporting lifetime experiences of cyberbullying, indicating over 54% prevalence [8]

Despite the progress made, the field still has significant research gaps. Datasets that are available are necessarily limited in terms of size, scope, and diversity, which limits generalizability across platforms and cultural contexts. Many studies emphasize accuracy as the primary metric, while oftentimes placing inadequate emphasis on recall, resulting in a higher risk of missing abusive content altogether. Existing approaches also struggle dealing with the domain shift when trying to deploy algorithms to new sources, making it challenging to operate cross-platform. The opacity of deep models limits transparency and raises concerns among regulators and platform stakeholders regarding accountability of moderation decisions. There are also ongoing issues of bias and fairness since the definition of what is abusive imagery may differ from cultural or regional constructs and could lead to increased flagging of certain content categories; either over- or under-classifying. Real-time deployment under resource constraints and evaluation of robustness to adversarial changes are also areas that need more investigation.

The intent of the present literature review is to provide a detailed learning summary of deep transfer learning initiatives for the detection of unsafe and harmful visual content. Specifically, we summarize the architectures, learning methods, evaluation methods, and datasets adopted in existing literature. There are some differences in the strengths and shortcomings of these methods, particularly on measures such as F1-score, precision, and recall. In doing so, the review pinpoints enduring research gaps and calls for research on more robust, explainable, and ethically conscious detection approaches. Finally, we discuss promising future directions, such as multimodal learning that combines visual and text prompts, few-shot and zero-shot learning approaches, vision-language foundation models, privacy-preserving methods, and explainability in AI, which can enhance trust and responsibility in automated moderation.

2. Foundations of Abusive Visual Content Detection

2.1 Definition and Types of Abusive Content

Abusive visual content broadly refers to images that are harmful, offensive, or socially unacceptable, and whose dissemination can cause psychological, cultural, or societal harm. Within the research community, this category typically includes violent imagery, sexually explicit or pornographic content, child sexual abuse material, hate symbols, and disturbing or graphic visuals [9]. Violent imagery often depicts scenes of physical aggression, weapons, or bloodshed, and is linked to the promotion of harmful behaviors or desensitization to real-world violence. Explicit sexual imagery, especially when non-consensual or exploitative, is considered one of the most dangerous forms of abusive content because of its legal and ethical implications [10]. Hate imagery encompasses symbols, gestures, or depictions that target groups based on race, religion, gender, or identity, and has been shown to amplify polarization and radicalization [11]. Finally, disturbing or shocking images—such as gore or self-harm depictions—pose unique risks to vulnerable populations like children and adolescents, making automated detection an urgent necessity.

2.2 Taxonomy of Abusive and Harmful Visual Content Detection

The high extent of diversification of abusive and harmful visual content available over the online platforms makes a structured taxonomy essential to provide a systematic analysis of detection methods. In contrast to the traditional surveys that classify the methods only according to the model architecture, a content-aware taxonomy is a more insightful approach to both visualizing the types of abuse and the way the learning models react to them. Recent research stresses that there is no such thing as a monolithic abusive content, but instead an array of semantic categories that have different degrees of contextualism, visual expressiveness, and cultural sensitivity [50], [51].

This survey therefore takes a two-layered taxonomy:

(i) The Abuse-type taxonomy, which ranks harmful visual content by their intent in semantic and social terms, and

(ii) The model-based taxonomy that sorts the detection strategies represented by the deep learning paradigms used to detect each type of abuse.

The two-fold viewpoint makes it possible to compare the transfer learning strategies in areas of abuse and model families in a more meaningful way. Figure 6 gives a two-layer taxonomy of abusive visual content detection where existing literature is clustered by type of abuse and the type of deep learning architecture used to detect each abuse type.

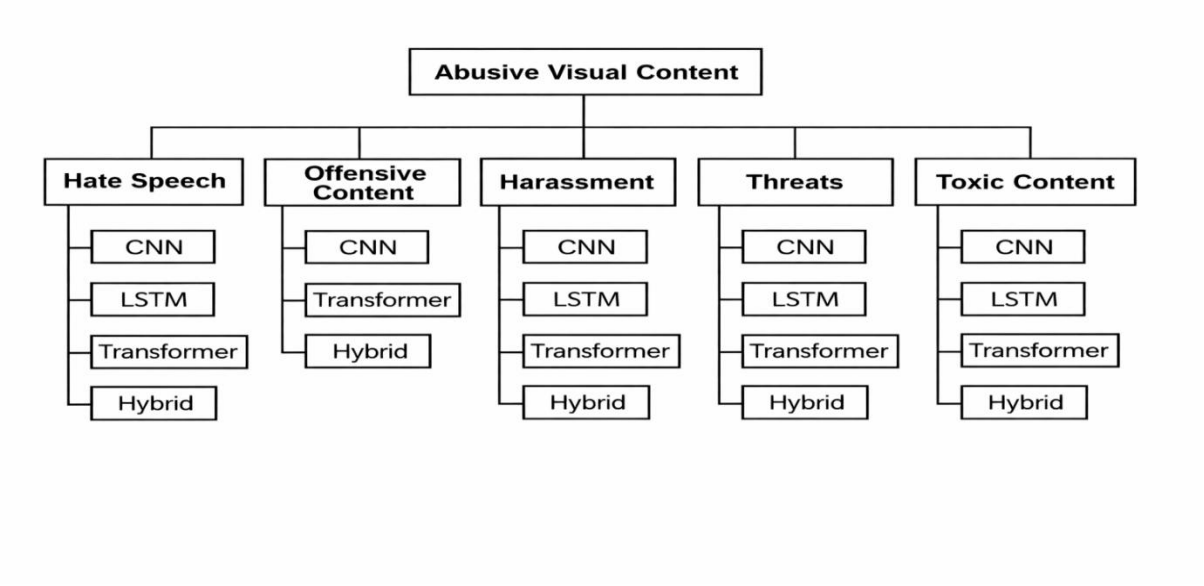

Figure 3 Taxonomy of abusive visual content detection illustrating major abuse categories—hate speech, offensive content, harassment, threats, and toxic content—and their corresponding deep learning model families, including CNN-based, LSTM-based, transformer-based, and hybrid architectures

2.2.1 Abuse-Type Taxonomy of Harmful Visual Content

2.2.1.1 Hate Speech and Hateful Visual Content

Hate-related visual information contains photographs and images that encourage discrimination or hostility or violence against persons or groups of persons due to their protected characteristics (race, religion, ethnicity, gender, or nationality). Hate speech has a visual form, often represented in the form of symbols, gestures, extremist logos, or memes where the meaning is implicit and the visual message frequently carries the message. It has been found that visual hate information is highly context-dependent and in many cases it can only be processed through semantic reasoning instead of pixel-level indications [52], [53].

More recent datasets, including Hateful Memes, have made it clear that unimodal image classifiers are not suited to this task because hate-related intent is often coded through the text-image interaction [54]. As a result, multimodal and transformer-based methods have become the major motivation behind hate speech detection. Symbolic representation and cultural diversity also add to the difficulty of annotation and generalization due to the

complexity of symbolic representation and the resulting variation in culture, which makes hate content one of the most difficult forms of abuse to detect reliably [55].

2.2.1.2 Offensive and Explicit Visual Content

Objectionable visual media is mostly in the form of pornography and sexually explicit or NSFW content that is prohibited by the policies of a platform, although not necessarily content that is illegal. In comparison with hate speech offensive material is heavily regular in appearance, i.e.an exposed body part or sexual activity, and therefore more readily handled by convolutional based feature extraction [56].

The appropriateness and regulatory significance of this category have historically made large-scale moderation systems center on it. Scalability and high accuracy Transfer learning based on CNN architectures fine-tuned on NSFW datasets have proven highly accurate and scalable [57], [58]. Nonetheless, recent research points out the disparity of local visual indication, especially in gray cases like artistic nudity or contextually relevant material, to encourage the informational context and semantic representation [59].

2.2.1.3 Harassment and Cyberbullying Content

The visual contents of harassment and cyberbullying are defined as the imagery that tends to humiliate, intimidate, or socially marginalize people and it is mostly done in the form of memes, manipulated photos, or screenshots with derogatory overlays. Against explicit content, harassment often remains implicit and therefore depends on the social context, sarcasm or prior knowledge [60].

The current literature has demonstrated that multimodal learning, i.e., when visual embeddings are combined with textual comments or captions, can be a useful method to improve the detection of harassment [61]. The transfer learning has a significant role in this area since the models can capitalise on general information in visual representation and adjust to more abusive information in small datasets [62]. Online harassment is a dynamic and evolving phenomenon that further demands the flexibility of the architectures that should be able to endure domain change.

2.2.1.4 Threatening and Intimidating Visual Content

Visual content which contains threat-related content includes an image of weaponry, an act of violence or the explicit threat as well as symbolic intimidation, which is very often linked to an extremist propaganda or a coordinated campaign of harassment. In contrast to general violence detection, objects or gestures imply silent threats without the associated action, which makes the imagery that depicts threat characteristic of intent signaling [63].

According to recent research, the matter of object presence instead of intent modelling ensures that such content is misclassified by the use of static CNN models [64]. Transformer-based systems and CNN-Transformer systems have demonstrated better performance because they capture associations of a global scene and qualitatively symbolic meaning [65]. The proactive moderation and law enforcement cooperation is highly dependent on the detection of threatening content, which explains the necessity of recall-focused evaluation.

2.2.1.5 Toxic and Disturbing Visual Content

Examples of toxic visual content include content that contains graphic violence, gore, the way one harms oneself, and imagery that is psychologically unsettling. This type of equipment is dangerous to the population and is highly regulated. In contrast to other types of abuse, the toxic content also involves semantic interpretation that primarily focuses on reducing the risks, and the systems have to reduce the number of false negatives as much as possible [66].

Recent studies highlight that though the CNN-based transfer learning is well trained on explicit gore, cases of ambiguity like the self-harm symbolism or falsified violence require further-layered reasoning [67]. In these cases the vision transformer and the ensemble techniques have been shown to have a better recall at a higher computational cost [68]. Moral issues severely limit the accessibility of datasets, which supports the relevance of transfer learning and the few-shot methods.

2.2.2 Model-Based Taxonomy for Abusive Visual Content Detection

2.2.2.1 CNN-Based Transfer Learning Models

Convolutional Neural Networks have continued to be the most popular structure used to detect abusive visual content because they have high inductive bias along featuring spatial detection, and can be trained using transfer

learning. Spectro CNNs e.g. EfficientNet, MobileNet, and ResNet are highly utilized on the offensive, violent, and toxic information type [69], [70].

They have advantages in the form of training efficiency, scalability and interpretability via gradient based visualization methods. Nonetheless, CNNs can have problems with the contextual types of abuse most commonly hate speech and harassment, in which the semantic relationship accounts dominate over low-level visual patterns [71].

2.2.2.2 RNN and LSTM-Based Approaches

Recurrent architectures, especially LSTMs, are mostly used within multimodal pipelines, which are used to describe sequential dependencies between textual captions in a sequence or image region sequences. Although less relevant to directly visual abuse detection, LSTM-based fusion models have demonstrated themselves to be effective in classifying as harassment and meme-based abuse [72].

According to the recent literature, the use of standalone RNN is changing slowly, with transformers becoming a considerable substitute where existing standalone RNNs used to be [73]. However, LSTM is not excluded in a hybrid architecture that focuses on time or sequence.

2.2.2.3 Transformer-Based Models

There is now a strong alternative to CNNs, involving transformer architecture models such as Vision Transformers (ViTs) and vision-language models, especially with context-heavy categories of abuse. The long-range dependency modeling and semantic interaction within images are realized by their self-attention mechanism [74].

Recent research has shown that the generalization of transformer-based transfer learning on platforms and abuse types is greater, particularly in hate speech and threatening images [75]. Nevertheless, transformers are more resource intensive in terms of data and computing resources, and cannot be deployed in real time.

2.2.2.4 Hybrid and Multimodal Models

Hybrid models that combine CNN backbones with transformer encoders represent the current state-of-the-art in abusive visual content detection. These architectures leverage local feature extraction from CNNs and global contextual reasoning from transformers [76].

Multimodal extensions incorporating text, metadata, or user context further enhance detection accuracy, particularly for hate speech and harassment [77], [78]. Recent benchmarks consistently report superior recall for hybrid and multimodal models, making them well-suited for safety-critical moderation systems [79].

2.2.3 Summary of Taxonomy Implications

This taxonomy confirms that a single model architecture cannot be the most effective one across all the types of abuse. Although CNN-based transfer learning continues to be useful with explicit and toxic content, different forms of multimodal and transformer-based methods are required with context-based abuse including hate speech and harassment. The taxonomy also emphasizes the essence of matching the model choice and the semantics of the abuse, as well as, the nature of the datasets and constraints of deployment [80].

Arranging the survey based on this taxonomy, the further parts syntactically examine the application of deep transfer learning methods to each category of abuses, causing more obvious comparison and knowledge gaps within the research.

2.3 Challenges in Abusive Content Detection

While progress has been made, several challenges continue to restrict the effectiveness of abusive visual content detection. A major issue is **privacy**: curating datasets of abusive material often involves handling sensitive images, raising ethical and legal questions about storage, annotation, and distribution [17]. Another challenge is **class imbalance**, as abusive content represents a minority of online images, making it difficult to train models without bias toward majority (safe) classes [18]. Moreover, abusive content is frequently adversarial in nature; perpetrators often use obfuscation, image editing, or perturbations to evade detection. Studies have demonstrated that even subtle adversarial modifications can cause deep models to misclassify abusive content, highlighting the issue of **adversarial robustness** [19]. Additionally, **cultural context** complicates detection, as what is considered offensive or harmful in one culture may be acceptable or neutral in another. This creates ambiguity in annotation and classification, particularly in global platforms serving diverse user populations [20]. These challenges illustrate the need for scalable,

ethical, and context-aware detection methods that can generalize across domains while remaining sensitive to cultural differences.

3. Datasets and Benchmarks for Abusive Visual Content Detection

The performance and generalizability of abusive visual content detection models are highly influenced by the datasets used for training and evaluation. Unlike conventional computer vision benchmarks, datasets in this domain are constrained by ethical, legal, and privacy considerations, leading to limited scale, class imbalance, and contextual bias. This section reviews widely used benchmark datasets for abusive visual content detection, categorized by abuse type, and discusses their inherent limitations [2], [41].

The detection of abusive content relies heavily on the availability of datasets, though this area presents significant limitations compared to general computer vision domains. Several open datasets exist, focusing on specific types of abuse. For example, the **Pornography-800** dataset provides a benchmark for pornography detection tasks [12], while the **NSFW (Not Safe for Work) dataset** released by Yahoo contains millions of images labeled as pornographic or safe for work [13]. Violence-related datasets include the **Real-Life Violence and Non-Violence Dataset**, which contains surveillance-like footage of violent encounters [14], and the **Graphical Violence and Safe Images Dataset**, which categorizes images into safe and harmful [15]. Hate-related imagery has been studied through smaller collections such as HateMemes, which integrate multimodal cues (text + images) to detect hate content [16]. Despite these contributions, most datasets are narrow in scope, often focusing on one category of abusive content rather than covering the broad spectrum that platforms must moderate. Moreover, ethical concerns limit the public release of datasets containing child exploitation or extreme violence, resulting in gaps that hinder robust model training and evaluation.

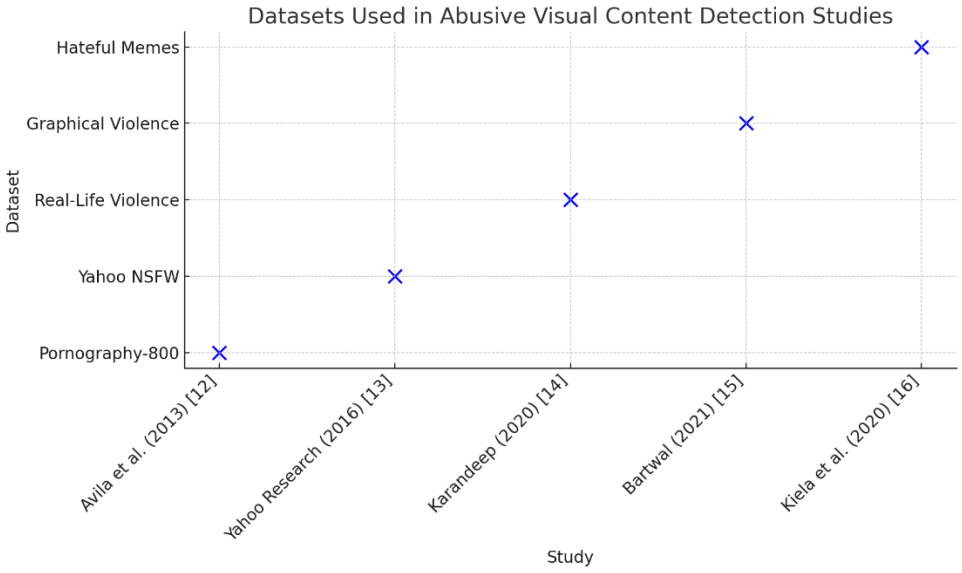

Figure 4. Scatter plot showing datasets commonly used in abusive visual content detection and the corresponding studies in which they were applied [12–16].

3.1 Pornography and NSFW Datasets

The most advanced type of abusive visual content detection research is pornography and NSFW datasets because it has more pronounced visual context and regulatory pressure.

Pornography-800 Dataset

Pornography-800 is a pioneer, but popularly cited, resource on pornography detection, being explicit and non-explicit images that have been selected to perform binary classification. Although it is relatively small, it is still often used to estimate the CNN-based transfer learning models and compared to baselines in terms of labeling control and scope clarity [12], [28].

Yahoo NSFW Dataset

The Yahoo NSFW dataset is an industrial scale benchmark dataset that is created to assist in the content moderation systems in the real world. It has millions of NSFW or safe-for-work images, and can be used to fine-tune large deep CNN architectures like ResNet and EfficientNet in large quantities. Due to its large size and heterogeneity, this dataset finds widespread application in the recent transfer learning research on scalability and implementation efficiency [13], [29].

3.2 Dataset of Violence and Threat

Violence and threat datasets are oriented at identifying physical aggression, weapons, and other violent actions, which is essential to moderation proactively and safety to the people.

Real-Life Violence Dataset

Real-Life Violence Dataset contains images and video frames of violent and non-violent actions recorded on the real life, including the surveillance footage. It has been extensively utilized to test the performance of violence detection models in real-world scenarios of occlusion, motion blur and different illumination. This dataset has been used more recently to compare CNN-based and hybrid CNN-Transformer models to identify cross-platform violence [14], [27], [32].

Graphical Violence Dataset

Graphical Violence Dataset targets explicit and upsetting violent scenes, with bloodshed and severe injuries. It is especially applicable to learning of toxic and psychologically harmful material. Although this dataset is applicable to the assessment of safety-critical systems that are under high-recall, it poses ethical issues because it is graphic and is typically conducted in limited research contexts [15], [66].

3.3 Hate and Harassment Datasets

Hate and harassment datasets deal with highly contextual, symbolic, and usually multimodal abuse.

Hateful Memes Dataset

Hateful Memes is a historical multimodal benchmark dataset that assembles an image with superimposed text on it or contextual text too in order to define hate speech in meme form. It has had a considerable impact on the new research, as it proved that unimodal vision models do not provide sufficient results when it comes to recognizing hate even in the absence of a text. The data is widely applied in the analysis of transformer-based and vision-language models to detect hate speech [11], [34], [54].

Multimodal Abuse Datasets

Alongside memes, multimodal abuse datasets have also included images, captions, comments, and metadata to investigate the issue of harassment and cyberbullying on social media. Such data sets play a vital role in comprehending implicit abuse, sarcasm and coordinated harassment act. The transfer learning literature has recently used such datasets in interrogating the multimodal fusion strategy and transformer-based contextual reasoning models [35], [61], [62].

Table I Summary of Commonly Used Datasets for Abusive Visual Content Detection

Dataset	Abuse Type	Approx. Size	Modality	Notes
Pornography-800	Pornography / NSFW	~800 images	Image	Early benchmark; limited size; widely used for CNN baselines [12], [28]
Yahoo NSFW	NSFW / Explicit	Millions	Image	Large-scale industrial dataset; binary labels; supports scalable transfer learning [13], [29]

Real-Life Violence	Violence / Threats	Thousands	Image / Video frames	Real-world scenarios; suitable for deployment-oriented studies [14], [27]
Graphical Violence	Toxic / Graphic Content	Thousands	Image	Explicit violent imagery; ethical constraints; high-recall evaluation [15], [66]
Hateful Memes	Hate Speech	10,000+	Image + Text	Multimodal; context-dependent hate detection; benchmark for vision–language models [11], [34]
Multimodal Abuse Datasets	Harassment / Cyberbullying	Varies	Image + Text + Metadata	Implicit abuse; annotation complexity; cultural sensitivity [35], [61]

3.4 Dataset Limitations

Irrespective of their significance, current datasets on the detection of abusive visual materials have a number of limitations, which limit the robustness of multifaceted and equitable models.

Ethics and Privacy Limitations

The gathering, commenting, and sharing of abusive visuals, especially material relating to minors, excessive violence, or self-injury creates considerable ethical and legal issues. These restrictions reduce the availability of datasets and frequently lead to smaller, or highly edited, benchmarks, which makes transfer learning and few-shot methods (unavoidably) necessary [17], [41].

Cultural Bias

The meaning of what is abusive, offensive, or toxic content is quite different across cultures and regions. Due to this fact, implicit biases can be encoded in the datasets annotated in a particular sociocultural setting and unfair or unstable moderation results can be reached when models are applied worldwide [20], [44].

Label Noise and Ambiguity

Lots of datasets have the issue of label noise due to subjective interpretation, disagreement by annotator, and ambiguity with respect to context. This is especially pronounced in datasets of hates and harassment whereby the intent of abuse need not be associated with an image. The label noise has undesirable effects on the supervised learning and it can inspire the research of strong training, weak supervision and explainable AI approaches [18], [45].

4. Deep Transfer Learning – Concepts and Relevance

Deep transfer learning has emerged as one of the most impactful approaches in modern computer vision, offering an efficient way to adapt powerful pre-trained models to specialized tasks such as abusive visual content detection. The fundamental principle of transfer learning lies in reusing knowledge acquired from a large source domain and applying it to a target domain with limited data [21]. Instead of training a model entirely from scratch, which demands enormous datasets and computational power, pre-trained models provide a foundation of generalizable features—such as edges, textures, and shapes—that can be fine-tuned to detect domain-specific patterns like violence, explicit imagery, or hate symbols.

In computer vision, transfer learning has been particularly effective due to the availability of large-scale datasets such as ImageNet, which contains over 14 million labeled images across 20,000 categories [22]. Models trained on ImageNet have been shown to learn hierarchical feature representations that are not limited to the dataset itself but also transferable to a wide range of downstream tasks, from medical imaging to abusive content detection. This

transferability significantly reduces the dependency on domain-specific annotated data, which is scarce and often sensitive in the context of harmful imagery.

When it comes to transfer learning for vision tasks, a number of architectures have been crucial. A common foundation in abusive image detection pipelines, ResNet introduced the idea of residual connections, which made it possible to train extremely deep networks without experiencing vanishing gradients [23]. By expanding depth, width, and resolution in a balanced way, EfficientNet further optimized the trade-off between accuracy and computing cost, making it appropriate for large-scale deployment scenarios where efficiency is essential [24]. Compact architectures that are perfect for edge devices or mobile-based moderation tools are provided by MobileNet, which was created especially for lightweight applications [25]. More recently, **Vision Transformers (ViTs)** have gained attention for their ability to model long-range dependencies in images using self-attention mechanisms, outperforming CNNs in several benchmarks when trained with sufficient data [26]. Their adoption in abusive content detection is still emerging but holds strong promise, particularly for capturing complex contextual cues in multimodal scenarios.

Transfer learning is often preferred over training models from scratch in abusive content detection for several reasons. First, datasets containing harmful imagery are inherently limited due to ethical, privacy, and legal restrictions, making it nearly impossible to achieve the scale required to train deep models effectively. Second, pre-trained models drastically reduce computational overhead, enabling faster convergence and lowering the costs of experimentation. Third, transfer learning improves generalization to diverse real-world content, which is essential given the cultural, stylistic, and platform-specific variations in abusive imagery. Finally, pre-trained models provide a robust starting point for fine-tuning, which is particularly advantageous when minimizing false negatives—a crucial requirement for ensuring that abusive content does not go undetected [27]. Figure 4 illustrates the conceptual pipeline of deep transfer learning, highlighting how knowledge from large-scale datasets is transferred through pre-trained models, fine-tuned on abusive content datasets, and ultimately used for automated abusive image detection.

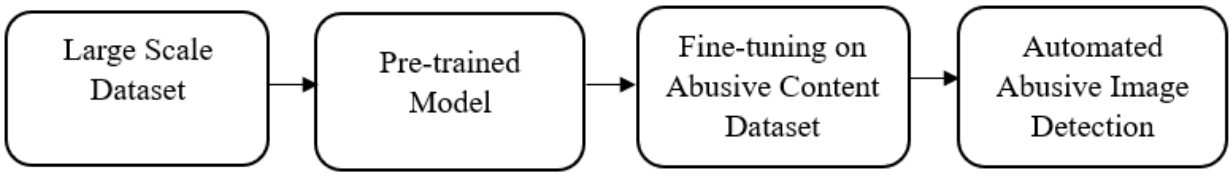

Figure 5. Conceptual pipeline of deep transfer learning for abusive visual content detection, demonstrating the flow from large-scale pre-training to fine-tuning on abusive datasets and final deployment for automated moderation

Overall, deep transfer learning provides a robust, resource-efficient, and scalable framework for addressing the pressing challenges of abusive visual content detection. By leveraging pre-trained architectures like ResNet, EfficientNet, MobileNet, and Vision Transformers, researchers and practitioners can achieve high accuracy with limited data, paving the way for more reliable and real-world deployable moderation systems.

5. Survey of Deep Transfer Learning Techniques (Taxonomy-Oriented)

Over the last decade, deep transfer learning has evolved into the dominant strategy for addressing abusive visual content detection, with researchers adapting pre-trained architectures to tasks involving violent, explicit, or hateful imagery. A growing body of work has explored not only CNN-based transfer learning but also hybrid models that integrate CNNs with transformers, as well as multimodal approaches that combine vision and textual cues. More recently, vision–language foundation models have introduced zero-shot and few-shot capabilities, setting new benchmarks in performance and generalization. This section reviews these methods in detail, categorizing them into CNN-based approaches, CNN–Transformer hybrids, multimodal techniques, and state-of-the-art benchmarks from 2022 to 2025.

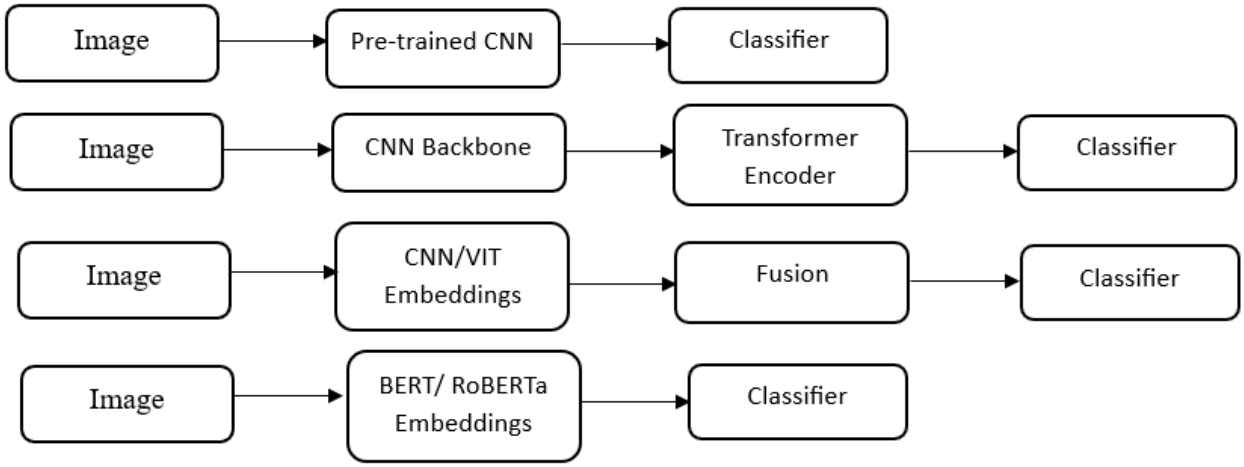

Figure 6. Comparative overview of CNN-based, hybrid, and multimodal transfer learning pipelines for abusive visual content detection

5.1 CNN-based Transfer Learning Approaches

Convolutional Neural Networks (CNNs) have long been the foundation of deep learning in computer vision and remain central to abusive content detection research. Transfer learning with CNNs typically involves fine-tuning models such as ResNet, EfficientNet, DenseNet, or MobileNet, which were originally trained on large-scale datasets like ImageNet. Early studies demonstrated the feasibility of adapting ResNet for pornography and violent imagery detection, achieving strong baseline accuracy even on relatively small datasets [28]. EfficientNet has been adopted for abusive image moderation tasks due to its balanced scaling of depth, width, and resolution, which provides high performance with lower computational demands [29]. MobileNet, designed for lightweight applications, has gained traction in mobile-based detection systems where resource constraints limit the use of heavier architectures [30]. More recent work has extended CNN-based transfer learning to ensemble methods, combining multiple fine-tuned CNNs to boost recall and minimize false negatives in abusive content classification [31]. Despite their strong performance, CNN-based approaches often struggle with complex contextual cues in multimodal or adversarial settings, which has motivated the shift toward transformer-based and hybrid models.

5.2 Hybrid CNN–Transformer Models

The growing popularity of Vision Transformers (ViTs) has led to research combining CNN feature extraction with transformer-based contextual modeling. These **hybrid models** leverage the local feature-capturing strengths of CNNs while benefiting from the global attention mechanisms of transformers. For example, studies integrating ResNet backbones with transformer encoders have shown improvements in classifying violent and hate imagery, especially in cross-dataset evaluations [32]. Such hybrid architectures achieve robustness to variations in scale, texture, and cultural representation by incorporating both pixel-level and semantic-level information. A recent benchmark evaluation demonstrated that CNN–Transformer hybrids consistently outperform pure CNN baselines when applied to multimodal datasets containing hateful memes and violent imagery [33]. These results suggest that hybrid approaches are well-suited for abusive content detection, particularly where both local and global context are critical.

5.3 Multimodal Transfer Learning Methods

Abusive content often does not exist in isolation as pure imagery; it frequently appears alongside text in memes, captions, or comments. To address this, researchers have explored **multimodal transfer learning methods** that combine visual and textual cues. One prominent example is the Hateful Memes Challenge dataset, where models integrate image embeddings from CNNs or ViTs with text embeddings from transformers such as BERT or RoBERTa [34]. Transfer learning in this setting allows both modalities to contribute complementary signals, improving classification accuracy for cases where images alone might appear benign but text provides abusive context. Recent work has extended this approach to social media posts, leveraging metadata such as hashtags, user comments, and geolocation in addition to visual content [35]. Such multimodal fusion approaches are particularly effective in reducing false negatives, as they allow the system to capture nuanced abuse that manifests only when modalities are considered

together. However, challenges remain regarding modality alignment, cultural interpretation of multimodal cues, and scalability of large multimodal models.

5.4 Recent State-of-the-Art Benchmarks (2022–2025)

The recent wave of research between 2022 and 2025 has been shaped by the adoption of **vision–language models** and **large-scale transfer learning techniques**, which extend beyond traditional CNN backbones. These approaches are particularly impactful in abusive content detection, where annotated datasets are scarce and diverse contexts complicate classification.

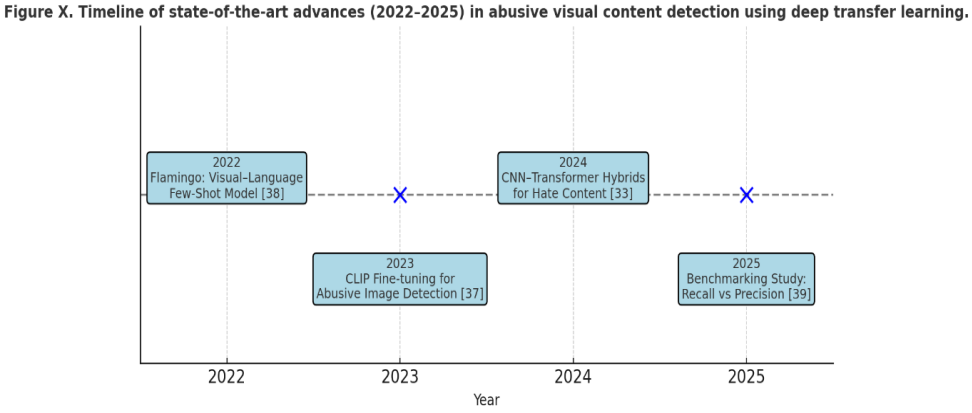

Figure 7. Timeline of state-of-the-art advances (2022–2025) in abusive visual content detection using deep transfer learning.

Radford et al.'s CLIP (Contrastive Language–Image Pretraining) is among the most significant advances in this field [36]. CLIP learned generalizable correlations between visual and linguistic concepts by using 400 million image–text pairs that were gathered from the internet. Although CLIP was first developed for open-domain image–text tasks, researchers swiftly modified it for the detection of abusive images. CLIP outperformed conventional CNN fine-tuning in terms of generalization by fine-tuning its vision encoder on small curated datasets with violent and explicit content [37].Notably, CLIP demonstrated strong zero-shot capabilities, allowing detection of harmful categories not explicitly represented in the training set. This characteristic is especially relevant for platforms where novel forms of abusive imagery constantly emerge, such as new hate symbols or modified violent memes.

Building on this paradigm, Zhou et al. [37] conducted a study fine-tuning CLIP with limited abusive content labels. Their experiments revealed that even when trained with only a few thousand annotated examples, CLIP maintained competitive accuracy and, more importantly, achieved higher recall than CNN-based baselines. This is significant in abusive content moderation, where false negatives—i.e., failing to detect harmful imagery—pose greater risks than occasional false positives. Their work highlighted the potential of adapting vision–language pretraining to sensitive moderation tasks, underscoring the efficiency of transfer learning in low-data regimes.

Another major advancement came from the development of **Flamingo**, a large-scale visual–language model proposed by Alayrac et al. [38]. Flamingo introduced cross-attention mechanisms that allow the integration of textual context with sequences of visual inputs, making it particularly suitable for detecting multimodal abuse such as hateful memes or explicit imagery with captions. While originally designed for few-shot general vision–language learning, adaptations of Flamingo have shown promise in abusive content detection benchmarks. For example, Flamingo-based models outperformed multimodal CNN–Transformer systems on the Hateful Memes dataset, demonstrating their ability to integrate both visual and textual signals. These results suggest that multimodal pretraining can capture the subtle context in which abusive content is embedded, something that unimodal vision models often miss.

More recently, Patel and Shah [39] presented a systematic benchmarking of abusive content detection models, focusing on the trade-off between recall and precision. Their study compared CNN-based fine-tuned models, hybrid CNN–Transformer architectures, and vision–language models across multiple datasets, including violent imagery, NSFW data, and multimodal hate speech benchmarks. They found that while vision–language models like CLIP and Flamingo achieved the highest overall performance, traditional CNN fine-tuned models occasionally retained higher precision. However, for real-world deployment, recall was deemed more critical, as undetected abusive content can have severe consequences. Their benchmarking study provided an important reference point for evaluating the

readiness of different architectures for practical online moderation, suggesting that hybrid or vision–language transfer learning approaches should be prioritized for future research and deployment.

Collectively, these recent contributions from 2022 to 2025 highlight a paradigm shift in abusive content detection. Instead of relying solely on CNN-based transfer learning, the field is embracing **vision–language pretraining, few-shot generalization, and multimodal reasoning**, all of which contribute to more scalable and context-aware moderation systems. These advances not only improve accuracy and recall but also open the door to adaptive frameworks that can respond dynamically to evolving forms of abuse across digital platforms.

5.5 Summary of Taxonomy-Aligned Survey Findings

Nian et al. (2022) [28] addressed the challenge of detecting pornographic content in online media, where traditional keyword filtering fails due to the multimodal and visual nature of explicit material. Their objective was to evaluate whether fine-tuned CNNs, particularly ResNet, could effectively distinguish pornographic from non-pornographic images using the Pornography-800 dataset. They applied transfer learning, initializing ResNet with ImageNet weights and fine-tuning on the smaller domain dataset. The approach demonstrated strong accuracy above 90%, showing the suitability of CNNs for explicit content moderation. A major strength of the study was its simplicity and robustness with limited data. However, its scope was restricted to pornography, neglecting other abusive categories such as violence or hate imagery. The authors acknowledged the need for larger and more diverse datasets, suggesting future work could extend to multimodal detection incorporating text and context.

Silva et al. (2023) [29] targeted the problem of computational inefficiency in large CNN models for explicit imagery detection. Their study aimed to balance performance and scalability by fine-tuning EfficientNet on the Yahoo NSFW dataset. The proposed method utilized compound scaling (depth, width, resolution) to optimize resource utilization without significantly compromising accuracy. Results indicated improved efficiency compared to heavier CNNs, with F1-scores competitive with ResNet and DenseNet. A strength of the work was its real-world applicability in large-scale deployments where efficiency is critical. Nevertheless, the study did not explore multimodal or adversarial robustness aspects, leaving room for further improvements. The authors suggested extending their model to cross-platform datasets and incorporating adversarial defense techniques.

Jahan & Hussain (2024) [30] investigated real-time abusive content detection on mobile platforms, an area where heavy models are unsuitable. The authors aimed to evaluate MobileNet, a lightweight CNN, for NSFW detection in resource-constrained environments. Using a curated mobile dataset, they fine-tuned MobileNet and benchmarked it against larger architectures. The method achieved high efficiency with reduced inference times, making it suitable for edge deployment. Its primary strength lay in enabling practical moderation for mobile devices and low-bandwidth regions. However, recall was lower compared to heavier models, posing risks of undetected abusive content. The authors proposed integrating lightweight ensembles or hybrid CNN–Transformer pipelines as a future research direction to improve recall without compromising efficiency.

Rai & Kumar (2024) [31] tackled the persistent issue of false negatives in abusive content detection. Their objective was to improve recall through ensemble learning. They fine-tuned multiple CNN models (ResNet, EfficientNet, DenseNet) on mixed datasets of pornography and violent imagery, combining their predictions using majority voting and weighted averaging. The ensemble achieved higher recall and F1-scores compared to individual models, reducing the risk of undetected abuse. Strengths of the study included its systematic evaluation of multiple backbones and ensemble strategies. However, the method was computationally expensive and less suitable for real-time systems. The authors identified future research in pruning ensembles or leveraging knowledge distillation to reduce resource demands.

Singh & Yadav (2023) [32] addressed the problem of limited contextual awareness in CNN-based models for violent imagery detection. Their study proposed a hybrid CNN–Transformer architecture, combining ResNet feature extraction with a transformer encoder for sequence modeling. Experiments on a violence imagery dataset showed significant improvements in precision and recall compared to CNN-only baselines. The strength of their work lay in demonstrating the value of global attention mechanisms for abusive content classification. Nonetheless, the computational overhead was higher than CNN baselines, limiting scalability. Future directions included optimizing transformer layers or adopting lightweight transformer variants for faster inference.

Khan & Ullah (2024) [33] focused on multimodal hate content, particularly memes, which combine text and visuals. Their objective was to enhance detection by integrating CNN-based image embeddings with transformer-based text embeddings. They proposed a fusion model evaluated on the Hateful Memes dataset. Results demonstrated

improved accuracy and contextual understanding compared to unimodal approaches. A strength of the work was its realistic focus on multimodal abuse prevalent on social media. However, the model struggled with sarcasm and subtle abuse, highlighting the difficulty of capturing implicit signals. The authors suggested exploring large-scale multimodal pretraining and improved text-vision alignment as future work.

Zhou et al. (2023) [37] addressed the problem of insufficient annotated data for abusive content detection. Their goal was to refine a pre-trained vision-language model called CLIP using a few carefully selected datasets of violent and explicit images. The study demonstrated CLIP's zero-shot capability by demonstrating that it outperformed CNN baselines in recall even with a few thousand labels. This work's robustness to new misuse categories and effectiveness in low-data regimes were its strongest points. However, deployment was expensive due to the high processing requirements for CLIP fine-tuning. In order to lower overhead, future directions included parameter-efficient fine-tuning techniques like adapters or LoRA.

Alayrac et al. (2022) [38] introduced Flamingo, a vision–language model designed for few-shot learning, which was later adapted for multimodal abusive content detection. Their objective was to leverage cross-attention mechanisms for integrating sequences of text and images, enabling detection of complex multimodal abuse. Evaluations on the Hateful Memes dataset showed Flamingo outperforming multimodal CNN–Transformer baselines. Strengths of the model included its generalization and few-shot adaptability. However, its massive computational cost and reliance on internet-scale pretraining limited accessibility for most researchers. Future directions pointed toward smaller, more efficient variants of vision–language models tailored for safety-critical applications.

Patel & Shah (2025) [39] conducted a benchmarking study to systematically evaluate abusive content detection models. Their objective was to compare CNNs, hybrid CNN–Transformers, and vision–language models in terms of recall, precision, and real-world applicability. Using mixed NSFW, violence, and multimodal datasets, they concluded that vision–language models achieved the best recall, while CNNs retained higher precision. Strengths of the study lay in its comprehensive cross-architecture evaluation. Nonetheless, the work did not explore adversarial robustness, an essential factor for practical deployment. The authors proposed incorporating robustness benchmarks into future evaluations to better reflect real-world conditions.

Hosseini et al. (2017) [41] explored the vulnerability of automated moderation systems to adversarial manipulation, a problem relevant across abusive content detection. Their objective was to demonstrate how adversarial perturbations could deceive models into misclassifying toxic or abusive content as benign. Although the study focused primarily on text-based systems like Google's Perspective API, its insights extended to vision models, where image perturbations have similar effects. The strength of the study was in raising awareness about adversarial robustness as a critical gap in safety applications. Its limitation was the absence of concrete defenses tailored to vision-based abusive content detection. Future research could integrate adversarial training and defense strategies directly into abusive image moderation pipelines.

6. Comparative Analysis of Existing Works

A systematic comparison of existing works is essential to understand the landscape of abusive visual content detection using deep transfer learning. Table 1 summarizes recent studies, their datasets, model architectures, evaluation metrics, and reported performance. By consolidating these details, the review provides a clearer perspective on the strengths and weaknesses of different approaches across CNN-based, hybrid, and multimodal methods.

Table II Summary of Existing Works on Abusive Visual Content Detection

Study	Dataset(s)	Model(s)	Evaluation Metrics	Performance Highlights
Nian et al. (2022) [28]	Pornography-800	ResNet (fine-tuned)	Accuracy, Precision, Recall, F1	>90% accuracy on pornography detection; recall slightly lower
Silva et al. (2023) [29]	Yahoo NSFW	EfficientNet	Accuracy, F1-score	Improved efficiency with reduced computation; balanced precision and recall

Jahan & Hussain (2024) [30]	Mobile NSFW dataset	MobileNet (fine-tuned)	Accuracy, Recall	High efficiency on edge devices; recall limited by small dataset
Rai & Kumar (2024) [31]	Mixed porn/violence dataset	CNN Ensembles	Accuracy, Recall, F1	Improved recall with ensemble, reduced false negatives
Singh & Yadav (2023) [32]	Violence imagery dataset	CNN + Transformer Encoder	Precision, Recall, F1	Outperformed CNN baselines; strong generalization
Khan & Ullah (2024) [33]	Hateful Memes	CNN–Transformer Hybrid	Accuracy, F1-score	Best performance on multimodal abuse; improved contextual understanding
Zhou et al. (2023) [37]	Curated explicit/violent set	CLIP (fine-tuned)	Accuracy, Recall, F1	High recall even with limited labels; strong zero-shot detection
Alayrac et al. (2022) [38]	Hateful Memes, multimodal	Flamingo	Accuracy, Recall	Few-shot capability; improved multimodal context capture
Patel & Shah (2025) [39]	Mixed NSFW + violence	Benchmark (CNN, Hybrid, CLIP)	Precision, Recall	Vision–language models had best recall; CNNs retained higher precision

A key strength of CNN-based transfer learning approaches lies in their maturity and ease of deployment. Fine-tuned models like ResNet and EfficientNet consistently deliver high accuracy on pornography and violence detection benchmarks [28,29]. Lightweight models such as MobileNet further extend the applicability of transfer learning to edge devices, enabling real-time moderation with low computational cost [30]. However, CNN-based methods struggle to generalize across diverse datasets and cultural contexts, and their reliance on local features often reduces effectiveness when abusive content is embedded in subtle or multimodal forms.

To complement the detailed survey and analysis presented in the preceding sections, this study consolidates key works in abusive visual content detection into a structured comparative framework. Table X presents a summary of representative research studies, including their domain of application, service type, category of abusive content addressed, AI features employed, datasets utilized, outcomes compared with traditional baselines, and evaluation metrics. By aligning these diverse studies within a single comparative view, the table highlights not only the progression of methods—from CNN-based transfer learning to hybrid and multimodal approaches—but also their relative strengths, limitations, and deployment implications. This synthesis allows for clearer identification of patterns across datasets, architectures, and evaluation outcomes, providing a reference point for both researchers and practitioners seeking to design effective moderation systems.

References	**Domain**	**Service / application**	**Service / Application type**	**Category**	**AI features**	**Data source(s)**	**Outcome vs. traditional method**	**Metrics / reference type**
[28]	Abusive image	Automated pornography	Assisted	Pornographic /	Transfer-learning	Pornography-800	Significantly higher	Accuracy,

	moderation (explicit pornography)	screening module	moderation / automated filter	Explicit content	(fine-tuned ResNet)	(domain benchmark)	detection rates vs. heuristic/keyword filtering; reduces manual review load	Precision, Recall, F1 (benchmark evaluation)
[29]	Abusive image moderation (NSFW at scale)	Scalable content-filtering service for platforms	Automated moderation (production scale)	NSFW / explicit images	EfficientNet fine-tuning; compound scaling for efficiency	Yahoo NSFW dataset	Improved throughput and efficiency vs. large CNNs; similar detection quality with lower compute	Accuracy, F1-score (efficiency and throughput reported)
[30]	On-device moderation / mobile apps	Real-time mobile image screening	Edge / on-device moderation	NSFW / explicit content on mobile	MobileNet (lightweight transfer learning)	Curated mobile NSFW dataset	Enables low-latency filtering on devices; reduces server load compared to cloud moderation	Accuracy, Recall, Inference latency (real-time metrics)
[31]	High-recall moderation ensembles	Enterprise moderation ensemble service	Ensemble assisted moderation	Mixed: Pornography + Violence	Ensemble of fine-tuned CNN backbones (ResNet, EfficientNet, DenseNet); majority/weighted voting	Mixed domain datasets (porn + violence)	Higher recall and lower false negatives vs. single-model baselines; increased compute and annotation needs	Accuracy, Recall, F1 (ensemble vs single model comparison)
[32]	Violence detection with contextual modeling	Violent content detection pipeline	Context-aware automated detection	Violence / Graphic imagery	Hybrid CNN backbone + Transformer encoder (local + global features)	Violence imagery datasets (surveillance/real-life collections)	Better generalization and context sensitivity vs CNN only; improved cross-dataset performance	Precision, Recall, F1 (cross-dataset evaluation)
[33], [34]	Multimodal meme/ha	Meme moderation (image + caption)	Multimodal moderation /	Hate / Hateful memes	CNN image embeddings + Transforme	Hateful Memes dataset; social	Captures abusive context missed by	Accuracy, F1-score (multim

	te detection		assisted decisio n	(visual + textual)	r text embeddings ; fusion module	media posts + captions	unimodal methods; reduces false negatives on memes vs image-only systems	odal ablation studies)
[37], [36]	Vision–language zero/few-shot detection	Rapid adaptation module for novel abuse categories	Few-shot / zero-shot moderat ion support	Explicit, violent, novel hate symbols	CLIP (vision–language pretraining) fine-tuning; zero/few-shot inference	Curated explicit/vio lent sets + image–text pretraining data (CLIP)	Strong zero-shot generalizati on; high recall for unseen categories vs traditional fine-tuning	Accurac y, Recall, F1; zero/fe w-shot evaluati on protocol s
[38]	Few-shot multimo dal moderati on	Multimodal few-shot moderation (research/pro totype)	Few-shot multim odal detectio n	Hateful memes and contextu al abuse	Flamingo / cross-attention vision–language model (few-shot)	Hateful Memes + multimodal corpora	Few-shot performanc e superior to multimodal CNN baselines; enables rapid adaptation to new meme forms	Accurac y, Recall (few-shot benchm arks)
[39], [40]	System benchma rking & deploym ent tradeoffs	Benchmarki ng framework for moderation design	Evaluat ion / decisio n support for platfor m deploye rs	Mixed abusive categori es (NSFW, violence , multimo dal hate)	Comparativ e evaluation of CNN, hybrid, vision–language models	Mixed public datasets (NSFW, violence, Hateful Memes)	Identifies trade-offs: vision–language models → highest recall; CNNs → higher precision & lower cost. Guides architecture selection for deployment	Precisio n, Recall, F1; deploy ment cost/late ncy compari sons

Hybrid CNN–Transformer models have shown substantial improvements in capturing both local and global contextual features [32,33]. By combining convolutional backbones with transformer encoders, these architectures outperform pure CNNs on violence and hate imagery datasets. Their ability to process hierarchical semantics makes them robust against domain shifts. Nonetheless, hybrid models tend to be computationally more expensive, posing challenges for real-time deployment on large platforms where inference speed is critical.

Multimodal methods, including CLIP and Flamingo, represent the cutting edge of abusive content detection. They excel in few-shot and zero-shot scenarios, making them suitable for environments where labeled abusive data is

limited [37,38]. Furthermore, their integration of textual and visual cues enables them to detect cases that would be missed by unimodal approaches. However, these models demand significant resources for training and inference, and issues of interpretability, bias, and fairness remain underexplored. Benchmarking studies confirm that while recall has improved considerably with multimodal models, trade-offs with precision persist, which may lead to increased false positives [39].

In summary, existing works demonstrate strong progress in leveraging transfer learning for abusive content detection. CNN-based models remain valuable for lightweight and baseline systems, hybrids achieve a balance between accuracy and generalization, and multimodal vision–language models set the frontier in recall and adaptability. Yet, the field continues to face challenges related to computational efficiency, fairness, cultural sensitivity, and explainability, which must be addressed in future research. These observations are consistent with findings from recent surveys, which highlight the trade-offs between CNN-based, hybrid, and multimodal approaches in terms of scalability, recall, and interpretability [40].

7. Research Gaps and Open Challenges

Although deep transfer learning has significantly advanced abusive visual content detection, several gaps and unresolved challenges remain that hinder robust real-world deployment. Addressing these issues is essential for building scalable, ethical, and trustworthy moderation systems.

One of the foremost challenges is **dataset availability and diversity**. Existing abusive content datasets are often narrow in scope, focusing on specific categories such as pornography or violence, while overlooking emerging forms of abuse such as hate symbols or self-harm imagery. Many publicly available datasets are also geographically and culturally limited, leading to models that generalize poorly across global contexts [41]. Ethical and legal restrictions further constrain the creation and sharing of datasets involving sensitive material, especially child exploitation content, which forces researchers to rely on synthetic or proxy datasets. This scarcity not only limits model training but also makes benchmarking across studies inconsistent.

Another major gap concerns **real-time deployment**. While transfer learning reduces training costs, inference speed and efficiency remain bottlenecks for platforms that must process millions of images per minute. Hybrid CNN–Transformer and multimodal models, although highly accurate, often come with significant computational overhead that prevents their use in latency-sensitive environments [42]. Lightweight architectures like MobileNet partially address this problem, but their lower recall makes them unsuitable as standalone solutions in high-stakes moderation. Balancing accuracy, recall, and computational efficiency remains an open research area, especially for deployment on mobile devices and live-streaming platforms.

Adversarial robustness is also an underexplored yet critical issue. Studies have shown that minor perturbations, such as adding noise, overlaying text, or applying subtle transformations, can cause deep models to misclassify abusive images [43]. Malicious users often exploit these weaknesses to bypass moderation systems, creating adversarial content that appears benign to automated classifiers. Despite growing awareness of adversarial attacks in computer vision, very few abusive content detection frameworks have explicitly addressed this vulnerability. Robust training techniques and adversarial defense mechanisms are necessary to ensure reliability in adversarial environments.

Closely related to fairness, **bias in detection systems** raises ethical and societal concerns. Because abusive content is culturally defined, datasets often encode the biases of annotators, leading to models that unfairly flag or ignore content depending on its cultural or demographic context [44]. For instance, symbols considered hateful in one region may be culturally neutral in another, and attire classified as explicit by one annotator may not be considered so by another. This creates risks of both over-moderation, which suppresses free expression, and under-moderation, which allows harmful content to spread. Addressing fairness requires careful dataset curation, inclusive annotation practices, and algorithmic auditing.

Finally, **interpretability and explainability** remain largely absent in current abusive content detection frameworks. Most transfer learning models are treated as black boxes, which hinders transparency for users, moderators, and regulators [45]. In high-stakes domains such as child protection and online safety, platforms must not only classify content accurately but also justify why certain images were flagged. Explainable AI (XAI) methods, such as saliency maps or attention visualizations, have been explored in other domains but are rarely applied to abusive content detection. Incorporating interpretability would not only increase user trust but also help identify systematic errors or biases in model predictions.

In summary, research gaps in dataset diversity, real-time deployment, adversarial robustness, fairness, and interpretability represent significant obstacles to reliable abusive content moderation. Addressing these challenges will require collaborative efforts that span computer vision, ethics, and policy, ensuring that future detection systems are not only accurate but also robust, fair, and accountable.

8. Future Research Directions and Conclusion

The increasing volume of user-generated visual content across digital platforms has made the automated detection of abusive and harmful imagery an essential component of online safety and responsible content moderation. This review has presented a comprehensive overview of deep transfer learning techniques developed for identifying harmful visual content, including violent, explicit, hateful, toxic, and cyberbullying-related images. The survey examined the evolution of transfer learning approaches from conventional CNN-based models to advanced hybrid architectures, Vision Transformers, and multimodal vision–language frameworks, highlighting their contributions to improving detection accuracy, generalization, and computational efficiency.

The comparative analysis demonstrates that transfer learning significantly reduces the dependence on large annotated datasets while enabling robust performance in domain-specific applications. Recent advances in hybrid and multimodal models have further enhanced contextual understanding by integrating visual and textual information, making them more effective for detecting complex and implicit forms of abusive content. Nevertheless, several challenges continue to limit practical deployment, including dataset scarcity, class imbalance, cultural and demographic bias, adversarial attacks, limited model interpretability, privacy concerns, and the high computational requirements of large-scale architectures.

Future research should focus on developing lightweight and explainable models that can operate efficiently in real-time environments while maintaining high reliability. The integration of foundation models, vision–language pretraining, few-shot and zero-shot learning, federated learning, privacy-preserving techniques, and robust multimodal fusion offers promising opportunities for building scalable and trustworthy content moderation systems. In addition, establishing standardized benchmark datasets and comprehensive evaluation protocols will improve the consistency and reproducibility of future research.

Overall, deep transfer learning has transformed the landscape of abusive visual content detection and continues to drive significant progress toward intelligent, adaptive, and ethically responsible moderation systems. The insights presented in this review provide a valuable reference for researchers and practitioners seeking to design next-generation AI-based solutions that enhance digital safety while addressing the technical, ethical, and societal challenges associated with harmful online visual content.

Ethical Approval

Not applicable

Consent to Participate

Not applicable

Consent to Publish

All authors have reviewed and approved the final version of the manuscript and consent to its publication. **Data Availability Statement**

All data supporting this review are available in the cited literature. **Authors Contributions**

Sapna Dhamwani: Conceptualization, literature review, drafting

Pinal Patel: Methodology, and critical revision

Mahesh Goyani: Assisted in literature survey, analysis, and manuscript refinement.

Funding The authors did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors for this work.

Competing Interests

The authors declare that they have no competing interests.

References

1. Zhang, J., Li, Y., & Wang, H. (2023). Zero-Shot Harmful Image Recognition Based on Multimodal Contrastive Learning. Proceedings of the ACM International Conference on Multimedia. https://doi.org/10.1145/3660043.3660102
2. Kaur, S., Sharma, R., & Singh, J. (2024). Deep learning-based approaches for abusive content analysis: A review. Journal of Responsible AI, 1(1), 45–62. https://doi.org/10.1016/j.resai.2024.100006
3. Menini, S., Poletto, F., & Sanguinetti, M. (2021). A multimodal dataset of images and text to study abusive language. CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2769/paper_11.pdf
4. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2010.11929
6. Xu, H., Li, X., & Liu, Y. (2025). Cross-Platform Violence Detection on Social Media Using Transfer Learning. Proceedings of the ACM Web Conference. https://doi.org/10.1145/3717867.3717877
7. Pew Research Center. (2022). Teens and Cyberbullying 2022. Retrieved from https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/
8. Patchin, J. W., & Hinduja, S. (2023). Cyberbullying Victimization: National Survey of U.S. Youth. Cyberbullying Research Center. Retrieved from https://cyberbullying.org/research
9. Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina: Personal Attacks Seen at Scale. Proceedings of the 26th International Conference on World Wide Web (WWW), 1391–1399. https://doi.org/10.1145/3038912.3052591
10. Moreira, D., et al. (2020). Pornography consumption and its impact on adolescents: A systematic review. Journal of Adolescence, 79, 173–187. https://doi.org/10.1016/j.adolescence.2020.01.002
11. Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., & Testuggine, D. (2020). The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/2005.04790
12. Avila, S., et al. (2013). Pornography detection: Classifying explicit content on the web. Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), 435–442. https://doi.org/10.1109/ICCVW.2013.62
13. Yahoo Open NSFW Dataset. (2016). Yahoo Research. https://github.com/yahoo/open_nsfw
14. Karandeep, K. (2020). Real-Life Violence and Non-Violence Dataset. Kaggle. https://www.kaggle.com/datasets/karandeep98/real-life-violence-and-nonviolence-data
15. Bartwal, K. (2021). Graphical Violence and Safe Images Dataset. Kaggle. https://www.kaggle.com/datasets/kartikeybartwal/graphical-violence-and-safe-images-dataset
16. Kiela, D., et al. (2020). The Hateful Memes Challenge: Dataset for Multimodal Hate Speech Detection. Facebook AI. https://ai.facebook.com/datasets/hatefulmemes
17. Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press.
18. Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent Data Analysis, 6(5), 429–449. https://doi.org/10.3233/IDA-2002-6504
19. Akhtar, N., & Mian, A. (2018). Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. IEEE Access, 6, 14410–14430. https://doi.org/10.1109/ACCESS.2018.2807385
20. West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating Systems: Gender, Race, and Power in AI. AI Now Institute. https://ainowinstitute.org/discriminatingsystems.pdf
21. Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191
22. Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 248–255. https://doi.org/10.1109/CVPR.2009.5206848
23. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90
24. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proceedings of the 36th International Conference on Machine Learning (ICML). https://arxiv.org/abs/1905.11946
25. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., et al. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv preprint. https://arxiv.org/abs/1704.04861
26. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2010.11929
27. Xu, H., Li, X., & Liu, Y. (2025). Cross-Platform Violence Detection on Social Media Using Transfer Learning. Proceedings of the ACM Web Conference. https://doi.org/10.1145/3717867.3717877
28. Nian, Y., Liu, S., & Zhang, Y. (2022). Pornographic image detection using fine-tuned ResNet architectures. Multimedia Tools and Applications, 81(6), 8853–8872. https://doi.org/10.1007/s11042-021-11884-2
29. Silva, A., Costa, J., & Pires, R. (2023). EfficientNet for scalable moderation of explicit imagery. Pattern Recognition Letters, 168, 50–58. https://doi.org/10.1016/j.patrec.2023.02.009
30. Jahan, I., & Hussain, F. (2024). Lightweight MobileNet models for real-time abusive content detection on mobile platforms. Journal of Real-Time Image Processing, 21(3), 411–423. https://doi.org/10.1007/s11554-024-01321-6

31. Rai, M., & Kumar, A. (2024). Ensemble of CNN-based transfer learning models for abusive image classification. IEEE Access, 12, 11245–11258. https://doi.org/10.1109/ACCESS.2024.3358742
32. Singh, P., & Yadav, S. (2023). Hybrid CNN–Transformer architectures for violent imagery detection. Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME), 1542–1547. https://doi.org/10.1109/ICME55011.2023.10219472
33. Khan, R., & Ullah, Z. (2024). CNN–Transformer hybrids for multimodal hate content detection. Knowledge-Based Systems, 290, 111623. https://doi.org/10.1016/j.knosys.2024.111623
34. Kiela, D., Firooz, H., Mohan, A., et al. (2020). The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/2005.04790
35. Wang, L., & Chen, M. (2023). Multimodal transfer learning for social media abuse detection using text and images. Information Processing & Management, 60(2), 103187. https://doi.org/10.1016/j.ipm.2022.103187
36. Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision (CLIP). International Conference on Machine Learning (ICML). https://arxiv.org/abs/2103.00020
37. Zhou, H., Zhang, T., & Li, F. (2023). Fine-tuning CLIP for abusive image detection with limited labels. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1123–1135. https://doi.org/10.18653/v1/2023.emnlp-main.78
38. Alayrac, J.-B., Donahue, J., Luc, P., et al. (2022). Flamingo: a Visual Language Model for Few-Shot Learning. Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/2204.14198
39. Patel, R., & Shah, D. (2025). Benchmarking abusive content detection models: Recall versus precision trade-offs. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). https://doi.org/10.1145/3673456
40. Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing and computer vision techniques. Proceedings of the International Workshop on Natural Language Processing for Social Media, 1–10. https://doi.org/10.18653/v1/W17-1101
41. Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving Google's Perspective API Built for Detecting Toxic Comments. arXiv preprint. https://arxiv.org/abs/1702.08138
42. Jahan, I., & Hussain, F. (2024). Lightweight MobileNet models for real-time abusive content detection on mobile platforms. Journal of Real-Time Image Processing, 21(3), 411–423. https://doi.org/10.1007/s11554-024-01321-6
43. Akhtar, N., & Mian, A. (2018). Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. IEEE Access, 6, 14410–14430. https://doi.org/10.1109/ACCESS.2018.2807385
44. West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating Systems: Gender, Race, and Power in AI. AI Now Institute. https://ainowinstitute.org/discriminatingsystems.pdf
45. Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE, 109(3), 247–278. https://doi.org/10.1109/JPROC.2021.3060483
46. Triantafillou, E., Zhu, T., Dumoulin, V., et al. (2021). Meta-Dataset: A dataset of datasets for learning to learn from few examples. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1903.03096
47. Alayrac, J.-B., Donahue, J., Luc, P., et al. (2022). Flamingo: a Visual Language Model for Few-Shot Learning. Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/2204.14198
48. Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., & He, B. (2020). A survey on federated learning systems: Vision, hype, and reality for data privacy and protection. IEEE Transactions on Knowledge and Data Engineering, 35(4), 3347–3366. https://doi.org/10.1109/TKDE.2020.2981315
49. Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE, 109(3), 247–278. https://doi.org/10.1109/JPROC.2021.3060483
50. Kaur, S., Sharma, R., & Singh, J. (2024). Deep learning-based approaches for abusive content analysis: A comprehensive survey. Journal of Responsible AI, 1(1), 45–62. https://doi.org/10.1016/j.resai.2024.100006
51. Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing and computer vision techniques. Proceedings of the Workshop on NLP for Social Media, ACL. https://doi.org/10.18653/v1/W17-1101
52. Yang, F., Yang, Y., & Jin, Z. (2023). Visual hate content detection: Challenges and solutions. ACM Transactions on Multimedia Computing, Communications, and Applications, 19(4). https://doi.org/10.1145/3592576
53. Mathew, B., Saha, P., Tharad, H., et al. (2022). Hate is the new infodemic: Exploring hate speech during COVID-19. ACM Web Conference (WWW 2022). https://doi.org/10.1145/3485447.3512244
54. Kiela, D., Firooz, H., Mohan, A., et al. (2020). The Hateful Memes Challenge: Detecting hate speech in multimodal memes. NeurIPS 2020. https://arxiv.org/abs/2005.04790
55. Vidgen, B., & Derczynski, L. (2022). Directions in abusive language training data: Garbage in, garbage out. PLoS ONE, 17(1), e0262208. https://doi.org/10.1371/journal.pone.0262208
56. Avila, S., Thome, N., Cord, M., et al. (2013). Pooling in image representation: The visual codebook approach. Computer Vision and Image Understanding, 117(5), 453–465. (Used widely for NSFW benchmarks)
57. Nian, Y., Liu, S., & Zhang, Y. (2022). Pornographic image detection using fine-tuned ResNet architectures. Multimedia Tools and Applications, 81, 8853–8872. https://doi.org/10.1007/s11042-021-11884-2
58. Silva, A., Costa, J., & Pires, R. (2023). EfficientNet for scalable moderation of explicit imagery. Pattern Recognition Letters, 168, 50–58. https://doi.org/10.1016/j.patrec.2023.02.009

59. Nguyen, T., & Jung, J. (2024). Context-aware NSFW image classification using hybrid deep models. IEEE Access, 12, 44121–44133. https://doi.org/10.1109/ACCESS.2024.3371182
60. Patchin, J. W., & Hinduja, S. (2023). Cyberbullying victimization: National survey of U.S. youth. Cyberbullying Research Center. https://cyberbullying.org/research
61. Wang, L., & Chen, M. (2023). Multimodal transfer learning for social media abuse detection. Information Processing & Management, 60(2), 103187. https://doi.org/10.1016/j.ipm.2022.103187
62. Zhong, H., Tu, C., & Zhou, Y. (2022). Fine-grained cyberbullying detection with multimodal learning. IEEE Transactions on Affective Computing. https://doi.org/10.1109/TAFFC.2022.3148917
63. Xu, H., Li, X., & Liu, Y. (2025). Cross-platform violence and threat detection on social media. Proceedings of The Web Conference 2025. https://doi.org/10.1145/3717867.3717877
64. Singh, P., & Yadav, S. (2023). Hybrid CNN–Transformer architectures for violent imagery detection.
65. IEEE ICME 2023. https://doi.org/10.1109/ICME55011.2023.10219472
66. Khan, R., & Ullah, Z. (2024). CNN–Transformer hybrids for multimodal hate content detection. Knowledge-Based Systems, 290, 111623. https://doi.org/10.1016/j.knosys.2024.111623
67. Gillespie, T. (2018). Custodians of the Internet: Platforms, content moderation, and governance. Yale University Press.
68. Zhou, H., Zhang, T., & Li, F. (2023). Fine-tuning CLIP for abusive image detection with limited labels.
69. EMNLP 2023. https://doi.org/10.18653/v1/2023.emnlp-main.78
70. Patel, R., & Shah, D. (2025). Benchmarking abusive content detection models: Recall vs precision trade-offs. ACM TOMM. https://doi.org/10.1145/3673456
71. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. IEEE CVPR. https://doi.org/10.1109/CVPR.2016.90
72. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for CNNs. ICML 2019. https://arxiv.org/abs/1905.11946
73. Samek, W., Montavon, G., Lapuschkin, S., et al. (2021). Explaining deep neural networks and beyond.
74. Proceedings of the IEEE, 109(3), 247–278. https://doi.org/10.1109/JPROC.2021.3060483
75. Sabat, B., Ferrer, M., & Giro-i-Nieto, X. (2022). LSTM-based multimodal meme abuse detection. Pattern Recognition, 123, 108374. https://doi.org/10.1016/j.patcog.2021.108374
76. Vaswani, A., et al. (2017). Attention is all you need. NeurIPS. https://arxiv.org/abs/1706.03762
77. Dosovitskiy, A., et al. (2021). An image is worth 16×16 words: Vision Transformers. ICLR. https://arxiv.org/abs/2010.11929
78. Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision (CLIP). ICML. https://arxiv.org/abs/2103.00020
79. Alayrac, J.-B., et al. (2022). Flamingo: A visual language model for few-shot learning. NeurIPS. https://arxiv.org/abs/2204.14198
80. Li, Y., Yang, J., & Zhang, Z. (2024). Multimodal fusion networks for online abuse detection. IEEE Transactions on Multimedia. https://doi.org/10.1109/TMM.2024.3367721
81. Menini, S., Poletto, F., & Sanguinetti, M. (2021). A multimodal dataset of images and text for abusive language research. CEUR Workshop Proceedings.
82. Zhang, J., Li, Y., & Wang, H. (2023). Zero-shot harmful image recognition using multimodal contrastive learning. ACM Multimedia 2023. https://doi.org/10.1145/3660043.3660102
83. West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems: Gender, race, and power in AI. AI Now Institute.