CEA Comité Español de Automática
XLVII JORNADAS DE AUTOMÁTICA CÓRDOBA 2026
UNIVERSIDADE DA CORUÑA

Jornadas de Automática

Cribado de CSAM basado en visión artificial

Pablo Morais[a], Eduardo Fidalgo[a,*], Víctor González-Castro[a], Waqar Tanveer[a], Milad Mirjalili[a], Enrique Alegre[a]

[a]*Group for Vision and Intelligent Systems, I4 Institute, Universidad de León, 24071, León, Spain.*

To cite this article: Morais, P., Fidalgo E., González-Castro, V., Tanveer, W., Mirjalili, M. 2026. Computer vision-based CSAM screening. Jornadas de Automática, 47. https://doi.org/10.17979/ja-cea.2026.47.13766

Resumen

La detección de material de abuso sexual infantil (CSAM) es un desafío relevante en informática forense y protección de menores, cuya investigación está limitada por restricciones legales, éticas y de privacidad. Este trabajo presenta un pipeline modular de visión artificial para el cribado de riesgo de CSAM basado en tres señales visuales proxy: detección facial, estimación de edad aparente y clasificación de contenido adulto. El sistema integra YOLOv12n-face, SwinFace y un clasificador de pornografía basado en ViT-B/16, aplicando una regla de decisión que marca una imagen únicamente cuando coinciden evidencia de contenido adulto y presencia de un posible menor. Además, se proponen dos datasets proxy de tres clases, APD-2M-3C y NPDI-2K-3C, orientados a diferenciar imágenes pornográficas, no pornográficas sencillas y no pornográficas difíciles. Al no utilizar CSAM real, la evaluación se centra en el rendimiento temporal y los falsos positivos sobre datasets sin CSAM. Los resultados preliminares muestran una capacidad eficiente para procesar grandes volúmenes de imágenes, aunque se requiere validación adicional antes de un uso operativo.

Palabras clave: Aprendizaje Automático, Procesamiento de Imágenes, Técnicas de Inteligencia Artificial, Seguridad, Privacidad, Criminalidad

Computer vision-based CSAM screening

Abstract

The detection of Child Sexual Abuse Material (CSAM) is a major challenge in digital forensics and child protection, while research in this area is constrained by legal, ethical, and privacy restrictions. This paper presents a modular computer vision pipeline for CSAM risk screening based on three proxy visual signals: face detection, apparent age estimation, and adult-content classification. The system integrates YOLOv12n-face, SwinFace, and a ViT-B/16-based pornography classifier, applying a decision rule that flags an image only when evidence of adult content coexists with facial evidence suggesting a possible minor. In addition, two three-class proxy datasets, APD-2M-3C and NPDI-2K-3C, are introduced to distinguish pornographic images, simple non-pornographic images, and challenging non-pornographic images. Since no real CSAM was used, the evaluation focuses on processing performance and false-positive behavior on CSAM-free datasets. Preliminary results show that the pipeline can efficiently process large-scale image collections, although further validation with controlled law-enforcement datasets is required before operational deployment.

Keywords: Machine Learning, Image Processing, Artificial Intelligence techniques, Security, Privacy, Criminality

1. Introduction

The increase of child sexual abuse material (CSAM) on the Internet represents one of the most severe threats to child safety nowadays. Furthermore, driven by AI, this content has grown dramatically in recent years, making manual detection and moderation increasingly infeasible (UNICEF, 2026). Au-

*Autor para correspondencia: eduardo.fidalgo@unileon.es

tomated detection systems are becoming essential tools for law enforcement agencies, online platforms and child protection organisations in their effort to identify and eliminate this kind of material (Westlake et al., 2022).

Despite the critical need for robust detection methods, progress in this area remains slow due to significant challenges. The scarcity of properly labelled datasets, combined with the highly sensitive nature of the data, limits the development and evaluation of machine learning models (Lee et al., 2020). Furthermore, most datasets in this field are designed for binary classification, i.e., porn vs. non-porn, which may not capture the full complexity of real-world scenarios (Gangwar et al., 2021).

Traditional approaches to automated CSAM detection rely on hash-based techniques, such as the popular PhotoDNA from Microsoft (Steinebach, 2023). Even though these techniques are powerful, they are reactive, meaning they can only identify material already catalogued. Furthermore, any manipulation of the image can evade hash-based detection. Recent approaches have shifted towards machine learning and deep learning models capable of identifying CSAM based on its visual content. They typically combine multiple signals, such as nudity or skin detection, with age estimation from facial features (Gangwar et al., 2021) (Pereira et al., 2020). However, these approaches still face the aforementioned limitations.

This article addresses both challenges already mentioned. First, we present a modular computer vision pipeline for CSAM risk screening based on face detection, apparent age estimation and adult-content classification. Second, we introduce two proxy three-class datasets, APD-2M-3C (Gangwar et al., 2021) and NPDI-2K-3C (Moreira et al., 2016), designed to distinguish pornographic, easy non-pornographic and hard non-pornographic images. Finally, we perform a preliminary evaluation of processing time and false-positive behaviour on non-CSAM datasets, highlighting current limitations and future validation needs.

In the following sections, we present the state of the art for each module, the methodology we followed, the experiments conducted, the results, and, finally, some conclusions with possible lines of future work.

2. State of the art methods

2.1. Face recognition

(Amirgaliyev et al., 2025) performed a systematic review of the state of the art in a variety of problems, one of them being face recognition. After reviewing 142 relevant articles, they concluded that the best classical computer vision-based method was EigenFaces, achieving 96 % accuracy on SCD 2500, and the best deep learning-based method was ArcFace, achieving 98 % accuracy.

(Yang, 2025) improved AdaBoost with particle swarm optimisation and dual threshold classification, then integrated a facial sample generation learning algorithm based on the improved AdaBoost. This PSO-AdaBoost reduced the average training time to 53 seconds and achieved a maximum sample recognition rate of 96.41 % and a minimum misclassification rate of 3.24 %.

Object detection models have also been used for face detection tasks. For example, (Zhang and Hu, 2025) created a pipeline which combined a YOLO with EfficientNet for Face detection and a feature extraction model based on FaceNet. When compared against the YOLOv3 series, the proposed model achieved an accuracy increase of approximately 26.30 % and a computation time of 3.78 s (YOLOv3-tiny achieved 3.12 s). However, this model has some security vulnerabilities because it lacks live body detection capabilities.

A model to combat the problem of RIFD (Rotation-Invariant Face Detection), called the Robust Polar Transformation Network (RP-Net), was introduced in (Kaewkorn et al., 2025). They used the CenterNet model architecture, which employs MobileNetV2 and a Feature Pyramid Network to manage diverse scales and complexities of facial features. This model was evaluated using the FDDB dataset, in which it achieved an average accuracy of 92.4 % with an inference speed of 36 FPS on CPU and 110 on GPU, being the fastest model while having a medium size (MTCNN-Aug was smaller).

2.2. Age estimation

A model for age estimation, whose backbone combined CNN with Vision Transformers (ViT) and modifications to ResNet was proposed by (Lin et al., 2024). This novel model was evaluated using AgeDB, ChaLearn16 and UTKFace datasets against other CNN models such as EfficientNet-B2 or Inception-v4, in which it obtained mixed results, having the lowest Mean Absolute Error (MAE) and highest Confidence Score (CS) in AgeDB and ChaLearn16 (7.08, 45.9 % and 6.16, 55.8 % respectively) while in UTKFace, EfficientNet achieved better metrics.

Liu et al. (2025) proposed a novel framework named Global and Local Aware age estimation (GLA-Age) composed of three components: an age-related region attention module, an age Vision Transformer, and an adaptive attention dropping strategy. This framework was evaluated using both inter-dataset and cross-dataset evaluation protocols in FG-NET, MORPH2, KANFACE and CACD in which it always obtained the best MAE and CS, with the MAE consistently below 7, CS above 48 %, and top accuracy above 55 %.

Occlusion leads to a loss of critical facial features; therefore, Yu and Zhao (2025) presented a method for reconstructing occluded facial features to enhance age estimation accuracy in occluded conditions. Three age estimation models are used as teachers: CORAL (ordinal regression), DLDL (label distribution learning), and MWR (predicts both the mean and variance of age for confidence intervals). The best model was KD-CORAL, with an average MAE of 5.23 across AFAD, CACD, and IMDB-WIKI, though it had the third-highest CS.

Shou et al. (2025) introduced a new model for accurate age estimation on face images named Masked Contrastive Graph Representation Learning (MCGRL) that uses graph neural networks (GNN) instead of CNN and also introduces a self-supervised masked graph autoencoder (SMGAE). The method was evaluated in FG-NET and MORPH. FACES, SC-FACE-ROT and SC-FACE-SUR, with a top average confidence score of 67.28 % and MAE of 4.87. It also achieved lower training times of 344, 17 and 662 seconds in MORPH, FG-NET and CACD.

Figure 1: Graphic abstract of the proposed work in this article

2.3. *Adult pornography detection*

Recent works have also dealt with pornography detection. For instance, (Rautela et al., 2024) developed a custom model, DVRGNet, which combines a hierarchical structure of convolutional neural networks, including DenseNet, VGGNet, ResNet, and GoogLeNet, with a video obscenity classifier based on a 5-layer Bi-LSTM. The proposed model achieved an accuracy of 99.42 % on the NPDI-2K dataset and 99.04 % on the NPDI-800 dataset.

(Gangwar et al., 2024) proposed a deep learning-based architecture called DeepHSAR that combines two multi-label classification streams: a global, fully supervised stream and a local, semi-supervised stream. Both streams are based on attention-driven CNN networks and are fused to improve accuracy. The authors created their own multi-label dataset called SexualActs-150k. Several versions of the DeepHSAR model were evaluated against other models, including RDAR, SSGRL, and KSSNet; the DeepHSAR variant that combines both scoring streams achieved the best results, with accuracies of 78.94 % and 84.42 % on the positive and negative classes, respectively. The model achieved the highest accuracy among the models compared, including DVRGNet (Rautela et al., 2024), on the NPDI-2K dataset (99.85 %).

A hybrid model combining Few-shot learning with Transformers was introduced by (Coelho et al., 2024). The model employs an episode-based learning approach, in which conventional architectures (CNN) and transformers were compared; transformers achieved the highest accuracy (73.50 %) in the 5-shot setting. The final model was pre-trained on ImageNet-1k and fine-tuned on Places600, achieving 63.38 % accuracy on scenes provided by the Brazilian Federal Police.

Other authors have explored pornography detection in video. For example, (Zhu et al., 2024) developed a model called MFOVD (Multi-Frame Obscene Video Detection) based on multiple Vision Transformers (ViT) to extract information from different frames, along with an interaction module that incorporates temporal encoding to model relationships between consecutive frames. The model was evaluated on the NPDI dataset, achieving an accuracy of 96.2 %; it was also tested on a new dataset called OPD (which includes content that is difficult to detect), yielding an accuracy of 88.4 %. When the interaction module was removed, accuracy dropped sharply to 80.3 % (on the OPD dataset).

3. Methodology

3.1. *Pipeline*

The pipeline consists of three independent modules working to process each image sequentially. These modules are named as follows: Face Detection (FD), Age Estimation (AE), and Adult Pornography Detection (APD). Each module has its own pre-trained model, balancing precision and computational cost to ensure the workflow is as fast and precise as possible. For the FD module, the YOLOv12n-face model (Tian et al., 2025) was chosen for its high performance at low cost. The AE module uses the age estimation functionality of SwinFace (Qin et al., 2023), and the APD module uses a ViT-B/16 model (Dosovitskiy et al., 2020).

The pipeline works as follows: first, the input enters the FD module, where YOLOv12n-face detects possible faces in the image, for each face, we get the parameters of a *bounding box*, meaning, the location, height and width of the face inside the image; these *bounding boxes* are then sent to the AE module, which processes them as if they were the whole image and from the output of the SwinFace model, we take only the age estimation parameter; lastly, the full image enters the APD module, where we get the probability of an image being pornographic. All the information is then sent to the logic part of the pipeline, if an image has faces estimated to be under 20 years old (The mean absolute error (MAE) of SwinFace is 2.5 years (Qin et al., 2023), this threshold was chosen as to minimize the false positives, as those are more dangerous in this

Table 1: Summary of previous work with the APD module (best models for each category studied).

Model	Dataset	Accuracy	Precision	Recall	F1-Score
CustomCNN [1]	APD-Subset	0.5002	0.5556	0.0025	0.0050
ConvNeXtTiny	APD-Subset	0.9960	0.9960	0.9960	0.9960
Vit-B/16	APD-Subset	0.9997	1.000	0.9994	0.9997
Vit-B/32 [2]	APD-Subset	0.9993	0.9995	0.9990	0.9992
CustomCNN	NPDI-800	0.7212	0.7399	0.6781	0.6781
ConvNeXtTiny	NPDI-800	0.9499	0.9017	0.9815	0.9399
Vit-B/16	NPDI-800	0.8687	0.8684	0.8687	0.8685
Vit-B/32	NPDI-800	0.8794	0.8692	0.8338	0.8456

specific case than the false negatives) and the APD module has established that the image is pornographic (the probability of having pornographic content is higher than the threshold selected when creating the pipeline, for the experiments, this threshold was trivially chosen as 0.6), the image is flagged as *CSAM*.

Some shortcuts have been introduced in the pipeline workflow to reduce unnecessary processing time: if an image does not contain faces, it will skip the AE module and will be flagged as *Not CSAM* regardless of the APD module. This shortcut may create a potential risk of real CSAM images altered to remove the faces to be wrongfully flagged as *Not CSAM*, in production environments this risk should not be allowed. When creating the pipeline, the shortcut was introduced because of the high amount of images that did not contain any faces at all, but future iterations of the pipeline could try to mitigate this by moving this kind of images to a specific folder to be manually inspected by a human agent, the current version, however, does not address this problem.

The schematic design of the pipeline is shown in Figure 1.

In this work, we have focused on refining the APD module, a continuation of previous work (Morais, 2025) (unpublished), which compared different models for image classification tasks in adult pornography detection. The best-performing model in this work was the pre-trained ViT-B/16 trained on the ImageNet dataset (Dosovitskiy et al., 2020), then fine-tuned on a subset of 20,000 images from the APD-2M dataset (Gangwar et al., 2021) (named APD-Subset). A summary of the comparison from the previous work is shown in Table 1.

Part of this optimization process involved investigating the optimal configuration of the APD module, which meant trying different ViT-B/16 models trained on the adult pornography datasets (APD-Subset, NPDI-800, and NPDI-2K) and different threshold configurations. To evaluate the pipeline's performance, four datasets were selected for processing: SEA-Faces-30k (Gangwar et al., 2023), Juvenile-80K (Gangwar et al., 2021), SexualActs-150K (Gangwar et al., 2024), and the full APD-2M dataset (Gangwar et al., 2021). The pipeline was also improved to increase inference speed (i.e., the time taken to process an image), reducing it from an average of 0.275 seconds to 0.04 seconds.

3.2. Datasets

For this work, we have also created two variants of existing datasets. This arose from the difficulty of finding high-quality adult pornography datasets. For example, one of the problems with the NPDI-800 (Avila et al., 2013) dataset is that the images are low-resolution, which introduces noise and makes training less reliable.

The design of these datasets was inspired by NPDI-800 (Avila et al., 2013), which is divided into three categories: porn, non-porn-easy, and non-porn-hard. The main advantage of this structure is the inclusion of the non-porn-hard category, which enables the model to learn from cases where an image may seem pornographic but is not. Examples include images taken at beaches or swimming pools, where there is a high degree of skin exposure, but there is no pornography involved in it. The source datasets selected for this process were NPDI-2K (Moreira et al., 2016) and APD-2M (Gangwar et al., 2021). These datasets were chosen because prior work had already used them, facilitating their integration into the proposed methodology.

The process of creating the new versions of these datasets, NPDI-2K-3C and APD-2M-3C, was similar for both, but with some caveats, mainly because NPDI-2K-3C comprises videos rather than images. The first step was to create a *teacher model* (a ViT-B/16 trained on NPDI-800) optimised for three-class classification. Then, pseudo-labels were generated for both datasets using this initial teacher. After applying a confidence filter to remove low-confidence images, self-training began, allowing the *student model* to learn from these pseudo-labels and subsequently replace the teacher. This loop was performed up to 10 times to obtain the *final student model*. This model was then used in a procedure called *offline hard negative mining*, where it performed inference on the dataset (the processed APD-2M and NPDI-2K), then the images wrongfully categorised were manually added to the new dataset, and the process began again. This entire process had to be performed multiple times.

4. Experiments and results

4.1. Pipeline

First, it is important to note that all experiments were conducted on datasets that did not contain any CSAM due to legal restrictions. As a result, the experiments could only evaluate the false positive rate (FPR), and all adjustments to the pipeline were aimed at reducing it. The final set of experiments is presented in Table 2. These datasets consisted of multiple folders, each containing a different class. To speed up processing and identify which classes had higher FPR, the datasets were processed in batches based on the folders they contained. The

results shown in the Table represent the average of the original output across all processed folders.

When looking at the images classified as CSAM, which as previously mentioned are always false positives, we can appreciate the following: First, there are multiple images wrongly classified without a simple explanation simply due to model accuracy; secondly, as previously mentioned, the threshold of the AE module was selected to prioritize the false negative rate over the false positive rate, which means that many images are flagged as underage when they are not; lastly, it was observed that when the image had low resolution or had a high content of skin exposure, the APD module was prone to flag them as pornographic.

4.2. Datasets

NPDI-2K-3C is made with the entirety of the original NPDI-2K dataset separated into frames, which resulted in a sample size of 20,000, but after the whole workflow, it was reduced to around 17,000 images, due to low confidence scores on multiple images, which were removed from the final dataset. A similar approach was taken for APD-2M-3C. Instead of using the full dataset of 2 million images, a balanced subset of 100,000 images, half of each class, was used, resulting in a final size of around 80,000 images, due to the same reasons as NPDI-2K-3C. Specific details about the final datasets are shown in Table 3, and some samples from the final versions are depicted in Figure 2 and Figure 3.

Due to time constraints, only preliminary experiments could be done in both datasets, which can be seen in Table 4, this table shows the baseline metrics of InceptionNext (Yu et al., 2025) on the APD-Subset and NPDI-2K datasets and both the new datasets, it is important to state that NPDI-2K was already created with 3 categories in mind, which contrasts APD-2M, henceforth the results obtained. Furthermore, the impact of the unbalanced state of the datasets was not evaluated and could be a potential future line of work, however, it can be theorized that the models trained with these datasets will be biased towards classifying images as porn, especially in the APD-2M-3C case.

Table 4: Preliminary experiments in the new datasets with InceptionNext

Dataset	Accuracy	Precision	Recall	F1-Score
APD-Subset	0.9951	0.9953	0.9950	0.9951
NPDI-2K	0.9411	0.9459	0.9347	0.9407
APD-2M-3C	0.9271	0.8087	0.7778	0.7885
NPDI-2K-3C	0.9080	0.9029	0.8936	0.8975

5. Conclusions and Future Work

This work presented a modular computer vision pipeline for CSAM risk screening based on proxy visual signals: face detection, apparent age estimation and adult-content classification. We also introduced two three-class proxy datasets, APD-2M-3C and NPDI-2K-3C, designed to improve adult-content classification in challenging non-pornographic scenarios. Since no real CSAM was used, the current validation is limited to processing efficiency and false-positive behaviour on non-CSAM datasets. Future work will focus on threshold calibration, ablation studies, stronger validation protocols and controlled evaluation with law-enforcement-supervised data where legally and ethically feasible. Other lines of work may include different purpose modules, for example, a skin-attention module which focuses on the importance of the amount of skin shown in an image and its correlation to pornographic content.

Referencias

Amirgaliyev, B., Mussabek, M., Rakhimzhanova, T., Zhumadillayeva, A., 3 2025. A review of machine learning and deep learning methods for person detection, tracking and identification, and face recognition with applications. Sensors 25.
DOI: 10.3390/S25051410

Avila, S., Thome, N., Cord, M., Valle, E., AraúJo, A. D. A., 2013. Pooling in image representation: The visual codeword point of view. Computer Vision and Image Understanding 117 (5), 453–465.

Coelho, T., Ribeiro, L. S. F., Macedo, J., dos Santos, J. A., Avila, S., 9 2024. Transformers-based few-shot learning for scene classification in child sexual abuse imagery. In: Anais Estendidos da XXXVII Conference on Graphics, Patterns and Images (SIBGRAPI Estendido 2024). Sociedade Brasileira de Computação - SBC, pp. 8–14.
DOI: 10.5753/sibgrapi.est.2024.31638

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 10 2020. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021 - 9th International Conference on Learning Representations.
URL: https://arxiv.org/pdf/2010.11929

Gangwar, A., González-Castro, V., Alegre, E., Fidalgo, E., 7 2021. Attm-cnn: Attention and metric learning based cnn for pornography, age and child sexual abuse (csa) detection in images. Neurocomputing 445, 81–104.
DOI: 10.1016/j.neucom.2021.02.056

Gangwar, A., González-Castro, V., Alegre, E., Fidalgo, E., 2023. Triplebiggan: Semi-supervised generative adversarial networks for image synthesis and classification on sexual facial expression recognition. Neurocomputing 528, 200–216.
URL: https://www.sciencedirect.com/science/article/pii/S0925231223000346
DOI: https://doi.org/10.1016/j.neucom.2023.01.027

Gangwar, A., González-Castro, V., Alegre, E., Fidalgo, E., Martínez-Mendoza, A., 2024. Deephsar: Semi-supervised fine-grained learning for multi-label human sexual activity recognition. Information Processing and Management 61.
DOI: 10.1016/j.ipm.2024.103800

Kaewkorn, H., Zhou, L., Li, W., 2 2025. Rp-net: A robust polar transformation network for rotation-invariant face detection. Pattern Recognition 158.
DOI: 10.1016/J.PATCOG.2024.111044

Lee, H.-E., Ermakova, T., Ververis, V., Fabian, B., 2020. Detecting child sexual abuse material: A comprehensive survey. Forensic Science International: Digital Investigation 34, 301022.
URL: https://www.sciencedirect.com/science/article/pii/S2666281720301554
DOI: https://doi.org/10.1016/j.fsidi.2020.301022

Lin, C., Gou, J., Fan, Z., Liao, Y., 2024. A feature fusion-based resnet using the pooling pyramid for age estimation. Proceedings of the International Joint Conference on Neural Networks.
DOI: 10.1109/IJCNN60899.2024.10650545

Liu, X., Qiu, M., Zhang, Z., Shi, Y., Li, Z., Chen, X., Liu, Y., Chen, X., Yu, H., 3 2025. Enhancing facial age estimation with local and global multi-attention mechanisms. Pattern Recognition Letters 189, 71–77.
DOI: 10.1016/J.PATREC.2025.01.005

Morais, P., 2025. Aprendizaje profundo en la clasificación automatizada de imágenes en la detección de abusos sexuales a menores.

Moreira, D., Avila, S., Perez, M., Moraes, D., Testoni, V., Valle, E., Goldenstein, S., Rocha, A., 11 2016. Pornography classification: The hidden clues in video space–time. Forensic Science International 268, 46–61.
URL: https://www.researchgate.net/publication/308398120_Pornography_Classification_The_Hidden_Clues_in_Video_Space-Time
DOI: 10.1016/J.FORSCIINT.2016.09.010

Table 2: Processing results of the pipeline using ViT-B/16 trained on the APD-Subset.

Dataset	Number of folders	Images	Avg. time per image (s)	Total time (s)	False Positives
SEA-Faces-30K	3	30817	0.051	521	4
Juvenile Dataset	4	59539	0.062	861	4
SexualActs-150K	1	150646	0.038	5764	452
APD-2M	2	2220330	0.037	40875	2103

Table 3: Final configuration of the two new datasets.

Dataset	Total images	porn	non-porn-easy (s)	non-porn-hard (s)
APD-2M-3C	80640	50001	24194	6445
NPDI-2K-3C	17343	6482	7007	3854

(a) non-porn-easy sample

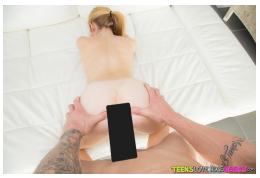
(b) non-porn-hard sample

(c) porn sample

Figure 2: Representative samples from the NPDI-2K-3C dataset.

(a) non-porn-easy sample

(b) non-porn-hard sample



(c) porn sample

Figure 3: Representative samples from the APD-2M-3C dataset.

Pereira, M., Dodhia, R., Anderson, H., Brown, R., 2020. Metadata-based detection of child sexual abuse material. arXiv preprint arXiv:2010.02387.

Qin, L., Wang, M., Deng, C., Wang, K., Chen, X., Hu, J., Deng, W., 8 2023. Swinface: A multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. IEEE Transactions on Circuits and Systems for Video Technology 34, 2223–2234.
URL: http://arxiv.org/abs/2308.11509http://dx.doi.org/10.1109/TCSVT.2023.3304724
DOI: 10.1109/TCSVT.2023.3304724

Rautela, K., Sharma, D., Kumar, V., Kumar, D., 2024. Obscenity detection transformer for detecting inappropriate contents from videos. Multimedia Tools and Applications 83, 10799–10814.
DOI: 10.1007/s11042-023-16078-2

Shou, Y., Cao, X., Liu, H., Meng, D., 2 2025. Masked contrastive graph representation learning for age estimation. Pattern Recognition 158.
DOI: 10.1016/J.PATCOG.2024.110974

Steinebach, M., 2023. An analysis of PhotoDNA. In: Proceedings of the 18th International Conference on Availability, Reliability and Security (ARES 2023). ACM.
DOI: 10.1145/3600160.3605048

Tian, Y., Ye, Q., Doermann, D., 2025. Yolov12: Attention-centric real-time object detectors.
URL: https://arxiv.org/abs/2502.12524

UNICEF, 2026. Artificial intelligence and child sexual abuse and exploitation. Accessed: 2026-05-27.
URL: https://www.unicef.org/reports/artificial-intelligence-and-child-sexual-abuse-and-exploitation

Westlake, B., Brewer, R., Swearingen, T., Ross, A., Patterson, S., Michalski, D., Hole, M., Logos, K., Frank, R., Bright, D., et al., 2022. Developing automated methods to detect and match face and voice biometrics in child sexual abuse videos. Trends and issues in crime and criminal justice (648), 1–15.

Yang, J., 12 2025. Facial recognition optimization based on adversarial sample generation in the field of artificial intelligence. Discover Artificial Intelligence 5.
DOI: 10.1007/S44163-025-00287-9

Yu, S., Zhao, Q., 6 2025. Improving age estimation in occluded facial images with knowledge distillation and layer-wise feature reconstruction. Applied Sciences (Switzerland) 15.
DOI: 10.3390/APP15115806

Yu, W., Zhou, P., Yan, S., Wang, X., 2025. Inceptionnext: When inception meets convnext.
URL: https://arxiv.org/abs/2303.16900

Zhang, J., Hu, N., 12 2025. Accuracy and robustness evaluation of deep learning algorithms in facial recognition systems. Systems and Soft Computing 7.
DOI: 10.1016/J.SASC.2025.200252

Zhu, D., Shan, X., Wu, C., Yung, K., Ip, A., 2024. Multi frame obscene video detection with vit: An effective for detecting inappropriate content. International Journal on Semantic Web and Information Systems 20.
DOI: 10.4018/IJSWIS.359768