Comparative Analysis of YOLO Architectures for Human Body Part Detection: Towards Symbiotic AI in Human-AI Interaction

Vita Santa Barletta[1,*,†], Danilo Caivano[1,†], Giovanni Dimauro[1,†], Massimiliano Morga[2,†], Alberto Maria Ricchiuti[1,†], Beatrice Scavo[1,†] and Federico Valentino[1,†]

[1]Università degli studi di Bari Aldo Moro, Piazza Umberto I, 70121 Bari, Apulia, Italy
[2]SER&Practices, Spin-off of the University of Bari Aldo Moro

Abstract

Cyber Social Security requires effective tools for the identification and automated moderation of harmful visual content, such as non-consensual nudity, sextortion, and online pornography. Addressing this issue requires not only accurate AI-based moderation tools but also systems that align with ethical, trustworthy, and human-centered design principles. In this study, we present a comparative analysis of two versions of the YOLO framework (YOLOv5 and YOLO11), evaluated across their respective model sizes (n, s, m, l, x, xl) and tested with both pretrained and randomly initialized weights. The goal is to determine the most effective configuration for the task of nudity detection. To this end, we constructed a dedicated dataset of over 5,000 annotated images across ten sensitive classes, with a focus on semantic balance and annotation quality. The models were tested under various configurations, revealing that YOLO11m with pretrained weights offers the best trade-off between accuracy and computational efficiency. The results confirm the potential of YOLO-based models for real-time automated moderation applications, while also highlighting the need for further improvements in localization accuracy.

Keywords

Cyber Social Security, Trustworthy AI System, Nudity Detection, YOLO Framework, Real-Time Object Detection

1. Introduction

The development of the internet and digital technologies has both enhanced social interaction and the availability of damaging material. Concerned Cyber Sociologists need mechanisms capable of automated non-consensual image-sharing detection and online sexual exploitation deterrence, capable of nudity detection as it relates to image analysis [1]. Discriminative models utilize photographic skin color or shape and use multi-stage pipelines associated with explicit content classification. Such approaches suffer from high false positive rates, lack of generalization, increased computing resources or time expenditures, and disjointed execution.

In addition, recent research explores the potential of human-AI symbiosis and human body part detection for advanced human-machine interaction. Willcox & Rosenberg [2] propose a Symbiont AI that learns to assist humans in real-time through Embodied Symbiotic Learning, fostering a partnership with shared expectations. In [3], the authors emphasizes augmented cognition to enhance human-machine symbiosis through mutual understanding and support. In the realm of human body part detection, Kuang et al. [4] introduce a method integrating human body part information to improve Human Object Interaction detection. Instead, Xu et al. [5] present AIP-Net, an anchor-free instance-level human part detection network that achieves state-of-the-art performance on the COCO Human Parts Dataset and demonstrates practical application in human-robot interaction. These advancements collectively

COL-SAI 2025: Workshop on COllaboration and Learning through Symbiotic Artificial Intelligence, in conjunction with the 16th Biannual Conference of the Italian SIGCHI Chapter (CHItaly 2025), October 6-10 2025, Salerno, Italy (2025)

*Corresponding author.
†These authors contributed equally.
vita.barletta@uniba.it (V. S. Barletta); danilo.caivano@uniba.it (D. Caivano); giovanni.dimauro@uniba.it (G. Dimauro); m.morga@serandp.com (M. Morga); a.ricchiuti20@studenti.uniba.it (A. M. Ricchiuti); b.scavo@studenti.uniba.it (B. Scavo); f.valentino7@studenti.uniba.it (F. Valentino)
0000-0002-0163-6786 (V. S. Barletta); 0000-0001-5719-7447 (D. Caivano); 0000-0002-4120-5876 (G. Dimauro)

contribute to the development of more effective and intuitive human-AI interactions, leveraging body part information and symbiotic learning approaches.

Therefore, considering the need to identify new tools for social security and the literature on nudity detection in social contexts, this paper describes a system for detecting nudity that uses exclusively the YOLO architecture which trains to find and mark nude boundaries in static images. While two-stage approaches must be less accurate, focus on inference speed, and greater ease-of-use, single-stage models derive better results. In this study, we attempt to determine which variant of YOLOv5 and YOLOv11 enables real-time moderation of explicit content based on accuracy, efficiency, and resource consumption.

2. Related Work

The automated detection of pornographic and sexually explicit content is a central challenge within the broader field of Cyber Social Security [6, 7], where it supports the mitigation of digital harms such as online grooming, sextortion, and unwanted exposure—particularly in vulnerable populations [8, 1]. Effective content moderation systems are critical for law enforcement, platform compliance, and the maintenance of healthy digital ecosystems.

Early approaches to visual explicit content detection primarily relied on color-based models to identify skin-toned regions under various lighting and pose conditions [9]. Although computationally efficient, these methods exhibited high false positive rates, often misclassifying sports scenes or skin-colored backgrounds. To address this, shape-based techniques introduced spatial constraints to better delineate potentially explicit regions [10], yet these approaches still lacked semantic understanding and generalization.

To improve robustness, mid-level representations such as the Bag of Visual Words (BoVW) were introduced, combining local feature descriptors with classifiers like SVMs for enhanced discrimination [11]. In video settings, the inclusion of motion-based features—such as MPEG-4 motion vectors, histograms of motion (MHIST), and periodicity detection (PER)—further enhanced detection accuracy, as shown by Jansohn et al. [12].

A major leap occurred with the advent of deep learning, particularly Convolutional Neural Networks (CNNs). AGNet, an ensemble of AlexNet and GoogLeNet, achieved 89.2% accuracy on the NPDI dataset by aggregating predictions across frames [13]. However, its lack of temporal modeling limited its effectiveness in video contexts. To address this, Perez et al. [14] extended GoogLeNet to incorporate sequential motion features, improving F1-score by 4–5% over AGNet. Subsequent work emphasized multi-task learning to enhance semantic richness. AttM-CNN, for instance, combined pornography detection with age estimation using a dual-branch CNN based on ResNet and Inception architectures [15, 16, 17]. Trained on over two million images, the model reached 92.7% accuracy, outperforming forensic tools like NuDetective by more than 20%.

More recently, the focus has shifted toward computational efficiency and real-time deployment. Mallmann et al. [18] introduced PPCensor, a CNN-based pipeline that reframes nudity detection as an object detection task. By applying localized obfuscation to private body regions, the system allows for granular moderation without discarding entire frames, while maintaining near real-time performance on edge hardware.

In parallel, transformer-based architectures have gained attention for their ability to capture global context. He et al. [19] demonstrated that Vision Transformers (ViTs) significantly outperform traditional CNNs such as ResNet in classifying sensitive content, thanks to their self-attention mechanisms.

YOLO-based methods have also emerged as promising alternatives for adult content detection. Typically, these systems follow a two-stage architecture: first detecting people or sensitive body parts using YOLO, followed by a secondary classification network [20, 21]. While effective, this separation introduces architectural complexity and additional inference latency.

Our work departs from this paradigm by employing YOLO in a fully end-to-end manner. We train the network directly to detect explicit regions without auxiliary classifiers, resulting in a single-stage

architecture that reduces latency and simplifies deployment—particularly in real-time applications. Unlike prior video-based methods that apply naive frame-by-frame processing [14, 18], our system focuses on static image analysis, leveraging YOLO's speed and spatial precision to isolate nudity with high fidelity. This provides a solid foundation for future extensions to multimodal, temporally aware moderation systems in large-scale platforms.

3. YOLO

YOLO (You Only Look Once) is a unified, real-time approach to object detection proposed by Redmon et al. (2016) [22], which reformulates the detection problem as a single regression task that directly maps from image pixels to bounding box coordinates and class probabilities.

The architecture of YOLO is based on a unified convolutional neural network that processes the entire image in a single pass. The image is divided into a grid of size $S \times S$, where each cell is responsible for detecting objects whose center falls within it. Each cell predicts B bounding boxes, each with a confidence score that reflects both the probability of the presence of an object and the spatial accuracy of the prediction, calculated using the Intersection Over Union (IoU) metric. In parallel, each cell provides a single conditional probability distribution over the C classes, which is computed only if the cell contains an object.

YOLO was chosen for its:

- **Speed** – since it treats the problem as a regression task, it does not involve a complex pipeline;
- **Contextualization ability** – as it has a global view of the image during both training and testing;
- **Generalization capability** – as it learns generalized representations of objects.

3.1. YOLOv5

YOLOv5 incorporates the Cross Stage Partial Network (CSPNet) [23] into its backbone (Darknet). CSPNet reduces redundant gradient information during training, thereby improving the model's efficiency. It splits the feature map into two flows: one is processed through a series of convolutional blocks, while the other remains unchanged. In the end, the two flows are concatenated, reducing the overall number of parameters and computational cost (in terms of FLOPs), without compromising performance.

In the neck, the model adopts the Path Aggregation Network (PANet) [24], which enhances information transmission between different levels of the network by adding a bottom-up path to the traditional top-down structure of the Feature Pyramid Network (FPN). This enables better propagation of both low- and high-resolution features, contributing to more accurate object localization.

Finally, the head of the network consists of three convolutional layers. The activation functions used are SiLU and Sigmoid: the former is applied in the hidden layers, while the latter is used in the output layer. The model outputs three types of predictions: the classes of the detected objects, their bounding boxes, and their objectness scores. The CIoU (Complete Intersection over Union) is used to compute the location loss.

3.2. YOLO11

YOLO11 [25] represents a significant advancement of the YOLO framework. The main innovations introduced in YOLO11 include:

- **C3k2 Block**: a more efficient variant of the classic CSP Bottleneck module. It uses two convolutions with smaller kernels instead of a single larger one, reducing computational cost while maintaining good performance. Its behavior can vary based on the `c3k` parameter, allowing for deeper structures when needed.
- **C2PSA Block**: introduces a spatial attention mechanism that helps the model focus more effectively on the most relevant areas of the image, improving detection accuracy, especially in complex scenes or with small or partially occluded objects.

- **CBS Blocks (Convolution-BatchNorm-SiLU)**: combine convolution, batch normalization, and SiLU activation to enhance the quality of the extracted features, making the learning process more stable and effective, and contributing to greater accuracy.

With respect to traditional YOLO architecture, the innovations introduced are arranged as follows:

- **Backbone**: replacement of the C2f block with the more efficient C3k2, retention of the SPPF block, and introduction of the new C2PSA to enhance spatial attention.
- **Neck**: use of the C3k2 block to improve speed and reduce computational complexity, along with integration of the C2PSA block to increase the relevance of features, especially for difficult-to-detect objects.
- **Head**: combined use of C3k2 and CBS blocks to process feature maps and increase detection accuracy. This section ends with 2D convolutional layers and the Detect module, which produces the final output (bounding boxes, confidence scores, and classes). The behavior of the C3k2 block is governed by the `c3k` parameter, which adjusts its internal structure.

For both YOLO versions, the Ultralytics platform offers model implementations in two configurations:

- one pre-trained on the COCO dataset;
- another initialized with randomly assigned weights.

4. Experimental Setting

The primary objective of our experimental evaluation is to identify the most suitable YOLO model variant for end-to-end nudity detection in static images. To this end, we systematically investigate how architectural variations within the YOLO family affect both detection accuracy and computational efficiency.

4.1. Dataset

For the development of our automatic human body part detection system, a dedicated dataset was constructed to address the specific requirements of the task. The dataset was developed through an iterative pipeline comprising repeated cycles of web-based image collection, manual annotation, and empirical evaluation of model performance. Particular attention was given to enhancing dataset quality and coverage through successive refinement steps, which included targeted augmentation of under-represented classes and exclusion of low-quality or ambiguous samples. This process allowed for the progressive improvement of the dataset in terms of both class balance and semantic diversity.

The final version of the dataset, employed for training the **YOLOv5** and **YOLOv11** object-detection models, consists of **5 090** images annotated with **8 247** bounding boxes. While the total number of samples remains relatively limited, the dataset reflects a considerable investment of time and manual effort, and its composition was carefully curated to optimize the training process for the intended detection task.

The dataset includes annotations for the following ten classes, encompassing both anatomical features and sexually explicit content: *anus, breast, buttocks, penis, vagina, oral-sex, penetration, penetration position, masturbation, porn.*

Annotations were performed using bounding boxes in accordance with a consistent labeling protocol designed to ensure inter-annotator agreement and reduce noise in the training data. Class frequencies were regularly monitored throughout the dataset-construction process, and specific measures were taken to mitigate class imbalance and prevent model bias. The resulting dataset thus provides a task-specific and well-structured foundation for the supervised training of explicit-content detection systems.

4.2. Model Variants

The evaluation focuses on a comparative analysis of multiple YOLO architectures, emphasizing both the well-established YOLOv5 family and the more recent YOLO11 series. The aim is to determine the optimal model configuration that balances detection performance with computational efficiency for the specific task of nudity detection.

The following model configurations were evaluated:

- **YOLOv5:** YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x
- **YOLO11:** YOLO11n, YOLO11s, YOLO11m, YOLO11l, YOLO11x

Each model was trained under two initialization strategies:

- Pre-trained weights from the COCO dataset;
- Random weight initialization.

Training was conducted for 100 epochs using the default hyperparameters provided by the respective implementations. All experiments employed consistent data augmentation strategies and loss functions. Input resolution and batch size were adapted per model to optimize GPU utilization while maintaining experimental comparability.

This setup allows for:

- Comparative analysis across lightweight, mid-sized, and high-capacity models.
- Identification of the YOLO variant offering the best trade-off between detection performance and computational efficiency.

4.3. Evaluation Metrics

Performance was assessed using the following metrics:

- **Classification of explicit content:** Precision and Recall.
- **Region-level detection:** mAP@0.5 and mAP@0.5:0.95

To ensure systematic monitoring and reproducibility, we utilized Weights & Biases and Comet throughout the training and evaluation phases. These tools enabled comprehensive tracking of:

- Precision, recall, and mAP over training steps;
- Learning curves and loss values;
- All relevant training hyperparameters.

This experimental protocol supports a fair, reproducible, and well-documented comparison of YOLO models of varying complexity under realistic deployment conditions.

5. Results

This section presents the comparative evaluation of YOLOv11 (Table 1) and YOLOv5 (Table 2) architectures on the task of nudity detection, performed on a curated dataset of over 5,000 annotated images across ten sensitive semantic classes. The goal was to assess the detection capabilities of each model variant, considering both pretrained and randomly initialized configurations, and to identify the optimal trade-off between accuracy and computational efficiency in view of real-time human-AI collaborative applications.

Among the YOLOv5 variants, YOLOv5x achieved the best results, with a mAP@0.5 of 0.367 and mAP@0.5:0.95 of 0.215. Precision and recall were 0.412 and 0.440 respectively, reflecting a reasonably balanced detection performance. Smaller configurations such as YOLOv5n and YOLOv5s showed a significant drop in recall, limiting their applicability in critical moderation tasks. The inclusion of pretrained weights yielded modest improvements across all sizes.

Table 1
Results YOLOv11 - 100 Epochs

	mAp50	mAP50-95	Precision	Recall	Train/box_loss	Train/cls_loss	Train/dfl_loss	Val/box_loss	Val/cls_loss	Val/dfl_loss
YOLO11n	0.4097	0.24192	0.54538	0.40525	0.96265	0.69773	1.14965	1.80369	2.06317	2.07752
YOLO11n_nw	0.26532	0.14282	0.33316	0.28161	1.42338	1.55267	1.72379	1.82086	2.39635	2.28682
YOLO11s	0.42492	0.24705	0.44267	0.40403	0.73974	0.47719	1.02935	1.85733	2.44981	2.24957
YOLO11s_nw	0.28574	0.16476	0.48582	0.28330	1.21506	1.20492	1.60142	1.88049	2.30250	2.46467
YOLO11m	0.43839	0.24324	0.40588	0.51567	0.65982	0.44588	1.04265	1.88873	2.35406	2.47389
YOLO11m_nw	0.29130	0.16420	0.48409	0.27488	1.12591	1.07718	1.55929	1.91592	2.28122	2.55150
YOLO11l	0.39429	0.22375	0.35385	0.38891	0.65880	0.44891	1.08961	1.86390	2.29913	2.61212
YOLO11l_nw	0.31476	0.17942	0.52056	0.31118	1.16698	1.17928	1.62193	1.79941	2.14943	2.44979
YOLO11x	0.42357	0.25149	0.47436	0.48488	0.64802	0.47013	1.10474	1.78357	2.21111	2.51209
YOLO11x_nw	0.31479	0.17559	0.34913	0.31639	1.08824	1.05677	1.57945	1.77708	2.22508	2.4833

YOLO11 – 100 Epochs

In contrast, YOLO11 models consistently outperformed YOLOv5 in both accuracy and recall. The YOLO11m configuration achieved the best results overall, with a mAP@0.5 of 0.438, mAP@0.5:0.95 of 0.243, and a recall of 0.516, outperforming all other configurations. Notably, models trained from scratch showed a marked decrease in performance—e.g., YOLO11m without pretrained weights reached only 0.291 in mAP@0.5—highlighting the importance of transfer learning, particularly in domain-specific visual tasks such as nudity detection.

A direct comparison between YOLOv5x and YOLO11m, summarized in Table 3, demonstrates the superior capability of YOLO11m, especially in detecting nuanced and sensitive content. These results underscore the potential of YOLO11-based architectures to support automated moderation systems that are both accurate and efficient.

From a human-AI symbiosis perspective, high recall and precision rates are essential to ensure user trust, system transparency, and ethical alignment. YOLO11m's performance enhances the reliability of AI-based moderators in identifying harmful visual content, minimizing both false positives and false negatives. Furthermore, the adaptability shown by pretrained configurations supports future personalization and domain transfer, critical for sensitive contexts such as healthcare, education, or platform moderation.

6. Conclusion

Extensive empirical evaluation revealed that the **Medium (M)** and **Large (L)** configurations of the YOLO architecture demonstrated the most favorable performance for human-body-part detection, particularly when initialized with pre-trained weights. These configurations offered an optimal compromise between detection accuracy and computational efficiency, rendering them appropriate for deployment in practical applications such as online-safety monitoring and content moderation.

Nonetheless, as illustrated by the visual results, the models' overall detection performance remained suboptimal. Although the networks exhibited a capacity to identify relevant anatomical features, they

Table 2
Results YOLOv5 - 100 Epochs

	mAp50	mAP50-95	Precision	Recall	Train/box_loss	Train/cls_loss	Train/dfl_loss	Val/box_loss	Val/cls_loss	Val/dfl_loss
YOLOv5n										
YOLOv5n_nw	0.27652	0.13867	0.28386	0.31522	0.044274	0.013946	-	0.04342	0.016226	-
YOLOv5s										
YOLOv5s_nw										
YOLOv5m										
YOLOv5m_nw										
YOLOv5l										
YOLOv5l_nw	0.34984	0.19117	0.37637	0.38883	0.028728	0.0063207	-	0.042647	0.015593	-
YOLOv5x	0.36713	0.21459	0.41207	0.44035	0.017037	0.0031023	-	0.040269	0.01844	-
YOLOv5x_nw	0.34423	0.18802	0.39816	0.34146	0.026648	0.0057727	-	0.039569	0.013265	-

YOLOv5 – 100 Epochs

Table 3
Performance comparison between YOLOv5x and YOLO11m on the nudity detection task

Model	mAP@0.5	mAP@0.5:0.95	Precision	Recall
YOLOv5x	0.367	0.215	0.412	0.440
YOLO11m	**0.438**	**0.243**	0.406	**0.516**

frequently encountered difficulties in achieving precise object localization and accurate delineation of bounding boxes. Among all evaluated variants, the **YOLOv11-M** model with pre-configured weights proved to be the most effective, yielding the highest precision scores. However, it still exhibited notable shortcomings in terms of boundary accuracy and spatial consistency.

These observations suggest that, while YOLO-based models hold promise for the task of body-part detection, further enhancements—such as more meticulous annotation, the incorporation of additional training samples, or architectural refinements—are required to improve localization precision and overall detection robustness.

7. Acknowledgments

This work was partially supported by the following projects: SERICS - ”Security and Rights In the CyberSpace - SERICS” (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU; Patto territoriale "Sistema universitario pugliese" – CUP F61B23000370006; Accordo Quadro CrASte - “Cyber Academy for Security and Intelligence”.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] V. S. Barletta, D. Caivano, G. Dimauro, F. Mantini, M. Morga, Exploring artificial intelligence challenges for monitoring cyber child abuse, volume 3978, 2025. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105008760266&partnerID=40&md5=fed46dfcbf71cc59bd9344aec2c4f01b.

[2] G. Willcox, L. B. Rosenberg, Symbiont ai and embodied symbiotic learning, Proceedings of the Future Technologies Conference (FTC) 2021, Volume 1 (2021). URL: https://api.semanticscholar.org/CorpusID:239802003.

[3] S. S. Grigsby, Artificial intelligence for advanced human-machine symbiosis, in: Interacción, 2018. URL: https://api.semanticscholar.org/CorpusID:51612552.

[4] H. Kuang, Z. Zheng, X. Liu, X. Ma, A human-object interaction detection method inspired by human body part information, in: 2020 12th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), 2020, pp. 342–346. doi:10.1109/ICMTMA50254.2020.00082.

[5] Y. Xu, Y. Zhang, Y. Leng, Q. Gao, Aip-net: An anchor-free instance-level human part detection network, Neurocomputing 573 (2024) 127254. URL: https://www.sciencedirect.com/science/article/pii/S0925231224000250. doi:https://doi.org/10.1016/j.neucom.2024.127254.

[6] V. Antoniol, F. Battista, P. Buono, D. Caivano, G. Calvano, G. Campesi, G. Cascione, A. Curci, M. de Gemmis, V. Gattulli, R. La Scala, R. Scardigno, A. L. Sciacovelli, A. Senaldi, P. Sorianello, V. Tamburrano, Cyber social security (css): A lens on methods for extraction of social sensor data, volume 3978, 2025. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105008758722&partnerID=40&md5=f6717d25ff68d5394e464db890b6ad62.

[7] V. S. Barletta, D. Caivano, M. Calvano, A. Curci, A. Piccinno, Craste: Human factors and perception in cybersecurity education, volume 3713, 2024, p. 75 – 81. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85198753881&partnerID=40&md5=35f9b858e583d214bb7a53c0a7dbf0da.

[8] M. T. Baldassarre, V. S. Barletta, V. Bavaro, D. Caivano, A. P. De Matteis, A. Lippolis, A. Piccinno, Llms to detect cyber child abuse in the in textual conversations, volume 3978, 2025. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105008757382&partnerID=40&md5=91bfb48c91b5043174e33d19f2ed45dd.

[9] T. Gevers, A. W. Smeulders, Color-based object recognition, Pattern Recognition 32 (1999) 453–464. URL: https://www.sciencedirect.com/science/article/pii/S0031320398000363. doi:https://doi.org/10.1016/S0031-3203(98)00036-3.

[10] Q.-F. Zheng, W. Zeng, G. Wen, W.-Q. Wang, Shape-based adult images detection, in: Third International Conference on Image and Graphics (ICIG'04), 2004, pp. 150–153. doi:10.1109/ICIG.2004.128.

[11] T. Deselaers, L. Pimenidis, H. Ney, Bag-of-visual-words models for adult image classification and filtering, in: 2008 19th International Conference on Pattern Recognition, IEEE, 2008, pp. 1–4. doi:10.1109/ICPR.2008.4761366.

[12] C. Jansohn, A. Ulges, T. M. Breuel, Detecting pornographic video content by combining image features with motion information, in: Proceedings of the 17th ACM International Conference on Multimedia, MM '09, Association for Computing Machinery, New York, NY, USA, 2009, p. 601–604. URL: https://doi.org/10.1145/1631272.1631366. doi:10.1145/1631272.1631366.

[13] M. Moustafa, Applying deep learning to classify pornographic images and videos, 2015. URL: https://arxiv.org/abs/1511.08899. arXiv:1511.08899.

[14] M. Perez, S. Avila, D. Moreira, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, A. Rocha, Video pornography detection through deep learning techniques and motion information, Neurocomputing 230 (2017) 279–293. URL: https://www.sciencedirect.com/science/article/pii/S0925231216314928. doi:https://doi.org/10.1016/j.neucom.2016.12.017.

[15] A. Gangwar, V. González-Castro, E. Alegre, E. Fidalgo, Attm-cnn: Attention and metric learning based cnn for pornography, age and child sexual abuse (csa) detection in images, Neurocomputing 445 (2021) 81–104. URL: https://www.sciencedirect.com/science/article/pii/S092523122100312X. doi:https://doi.org/10.1016/j.neucom.2021.02.056.

[16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9. doi:10.1109/CVPR.2015.7298594.

[18] J. Mallmann, A. O. Santin, E. K. Viegas, R. R. dos Santos, J. Geremias, Ppcensor: Architecture for real-time pornography detection in video streaming, Future Generation Computer Systems 112 (2020) 945–955. URL: https://www.sciencedirect.com/science/article/pii/S0167739X19331073. doi:https://doi.org/10.1016/j.future.2020.06.017.

[19] H. He, C. Wilson, T. T. Nguyen, J. Dalins, Sensitive image classification by vision transformers, in: 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2024, pp. 1704–1711. doi:10.1109/SMC54092.2024.10831156.

[20] N. AlDahoul, H. Abdul Karim, M. H. Lye Abdullah, M. F. Ahmad Fauzi, A. S. Ba Wazir, S. Mansor, J. See, Transfer detection of yolo to focus cnn's attention on nude regions for adult content detection, Symmetry 13 (2021). URL: https://www.mdpi.com/2073-8994/13/1/26. doi:10.3390/sym13010026.

[21] D.-D. Phan, T.-T. Nguyen, Q.-H. Nguyen, H.-L. Tran, K.-N.-K. Nguyen, D.-L. Vu, A novel pornographic visual content classifier based on sensitive object detection, International Journal of Advanced Computer Science and Applications 12 (2021). URL: http://dx.doi.org/10.14569/IJACSA.2021.0120591. doi:10.14569/IJACSA.2021.0120591.

[22] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[23] OpenGenus Foundation, YOLO v5 model architecture [Explained], https://iq.opengenus.org/yolov5/, 2021.

[24] K. Wang, J. H. Liew, Y. Zou, D. Zhou, J. Feng, Panet: Few-shot image semantic segmentation with prototype alignment, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9197–9206.

[25] R. Khanam, M. Hussain, Yolov11: An overview of the key architectural enhancements, arXiv preprint arXiv:2410.17725 (2024).