2024 International Conference of the Biometrics Special Interest Group (BIOSIG) | 979-8-3503-7371-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIOSIG61931.2024.10786730

An End-to-End Framework to Classify and Generate Privacy-Preserving Explanations in Pornography Detection

Margarida Vieira
Faculty of Engineering, University of Porto
INESC TEC
Porto, Portugal
margarida.j.vieira@inesctec.pt

Tiago Gonçalves
Faculty of Engineering, University of Porto
INESC TEC
Porto, Portugal
tiago.f.goncalves@inesctec.pt

Wilson Silva
AI Technology for Life,
Department of Information and Computing Sciences,
Department of Biology,
Utrecht University
Utrecht, The Netherlands
w.j.dossantossilva@uu.nl

Ana F. Sequeira
Faculty of Engineering, University of Porto
INESC TEC
Porto, Portugal
ana.f.sequeira@inesctec.pt

***Abstract*—The proliferation of explicit material online, particularly pornography, has emerged as a paramount concern in our society. While state-of-the-art pornography detection models already show some promising results, their decision-making processes are often opaque, raising ethical issues. This study focuses on uncovering the decision-making process of such models, specifically fine-tuned convolutional neural networks and transformer architectures. We compare various explainability techniques to illuminate the limitations, potential improvements, and ethical implications of using these algorithms. Results show that models trained on diverse and dynamic datasets tend to have more robustness and generalisability when compared to models trained on static datasets. Additionally, transformer models demonstrate superior performance and generalisation compared to convolutional ones. Furthermore, we implemented a privacy-preserving framework during explanation retrieval, which contributes to developing secure and ethically sound biometric applications.**

***Index Terms*—Biometrics, Computer Vision, Deep Learning, Explainable Artificial Intelligence, Pornography Detection, Privacy Preservation**

I. Introduction

The Internet's growing influence has escalated concerns regarding the acquisition and dissemination of explicit material, particularly pornography. While this issue transcends global legislative frameworks and law enforcement efforts, there is also a concern related to early exposure to such content, which can lead to numerous adverse effects (e.g., poor mental health, sexism, objectification, and increased risk of sexual violence). Moreover, professionals who analyse this material experience significant psychological distress [1]. Despite pornography's widespread consumption, research on its implications is limited and often overlooked in mainstream psychology, but with studies already indicating detrimental effects across cultural and age groups [2].

Current methods for detecting pornographic content lack clarity in their inner-functioning processes. Aiming to address this significant gap, this research pioneers the exploration of explainability within pornography detection. We investigate the potential of several deep learning (DL) architectures in 2D pornography detection while giving the first steps toward the degree of explainability of these methods.

Another critical aspect to consider is the privacy of the depicted individuals. Advances in generative artificial intelligence and the emergence of deepfakes, among other factors, have taken this preoccupation to new levels, like in 2017 when a Reddit user created pornographic content featuring non-consenting female celebrities' faces[1]. Given the inherently sensitive nature of this content, especially when we enter its illegal realm where individuals most certainly did not consent, this work also touches on the topic of privacy preservation through face anonymisation by proposing an end-to-end framework that integrates detection, explainability, and anonymisation efforts towards an effective, transparent, and

This work is co-financed by Component 5 - Capitalisation and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021 - 2026, within project NewSpacePortugal, with reference 11. It was also financed by National Funds through the Portuguese funding agency, Portuguese Foundation for Science and Technology (FCT) through the Ph.D. Grant "2020.06434.BD".

Note: This paper contains sexually explicit materials.

[1]https://inhope.org/articles/what-is-a-deepfake

ethical system (see Fig. 1).

We consider that this work makes the following contributions (see Section III):

1) A 2D pornography detection approach by fine-tuning existing pre-trained convolutional neural networks (CNNs) and transformer models, achieving performances on par with other more complex state-of-the-art strategies;
2) The exploration of various explainability techniques and discussion on the crucial need for more research into explainable artificial intelligence (XAI) to ensure safer and more reliable pornography detection, particularly given the sensitive nature of this content;
3) An end-to-end approach for detecting pornography and explaining the decision-making process while safeguarding the privacy of the individuals depicted through face anonymisation.

The code related to our methods' implementation is publicly available in a GitHub repository[2].

II. Related Work

A. Pornography Detection

Early research on pornography detection relied on traditional uni-modal computer vision and machine learning (ML) methods, such as support vector machines (SVMs), color histograms, or bag-of-visual-words (BoVW) [3], [4]. Later, the field moved towards leveraging the potential of multi-modal approaches, with visual, motion, and audio features being used to improve the predictive performance of models available until then [5], [6]. With the advent of DL, there was a paradigm shift towards automated feature extraction and classification, starting with CNNs [7]. Further advancements integrated CNNs with long short-term memory (LSTM) networks and explored fusion techniques to combine static and motion features [8]. More recent works employed advanced multi-modal and attention-based models to push the detection boundaries even further, leveraging the strengths of both CNNs and self-attention mechanisms in capturing complex patterns in visual and auditory data [9], [10].

B. Explainable Artificial Intelligence

XAI is a crucial research area that addresses complex models' often opaque nature. While DL models excel in various tasks, their decision-making processes are frequently inscrutable, which is problematic in applications requiring high trust and accountability. Explainability techniques aim to clarify these processes, enhancing trust, facilitating debugging, and ensuring ethical compliance.

One popular practice in the domain of XAI involves defining taxonomies to organise and distinguish the various explainability methods. However, their intricate and extensive nature hinders the creation of a universally applicable classification system; thus, multiple distinct taxonomies coexist. One common taxonomy relates to a method's stage as either *ante-hoc* - inherently interpretable models, designed with transparency in mind - or *post-hoc* - the focus of this work, applied to already trained models [11].

There already exist several *post-hoc* explainability methods, such as **Integrated Gradients (IG)** [12], which calculates feature contributions by integrating gradients along a path from a baseline (typically an all-zero image) to the actual input; **DeepLIFT** [13], which decomposes the output difference between an input and a baseline into contributions from each feature; **Layer-wise Relevance Propagation (LRP)** [14], which assigns relevance scores by backpropagating the model's output through its layers, using specific rules to redistribute the relevance; and **Occlusion** [15], a perturbation-based approach that evaluates feature importance by systematically masking parts of the input and measuring the effect on the model's output.

A major drawback of most XAI methods is their objective evaluation. Aiming to contribute to the solution of this open question Hedström et al. [16] proposed four different metric categories: **Complexity** (checks the conciseness of explanations), **Faithfulness** (checks if features deemed relevant affect model predictions more strongly), **Localisation** (checks if the explanatory evidence is located around a given region of interest), and **Robustness** (checks if explanations are stable when subject to input perturbations). Despite the sensitive nature of pornographic content, to our knowledge, this is the first work to explore XAI in the context of pornography detection.

III. Methodology

A. Data

We used three widely-used and publicly available [3] datasets in the experimental part of this work: the **Pornography-800** [17], composed of 800 videos (400 non-pornographic and 400 pornographic); the **Pornography-2k** [6], an extension of the latter, composed of 2000 videos (1000 non-pornographic and 1000 pornographic); and the **Adult Pornography Detection (APD-2M)**[4] [18], composed of approximately 2 million images (1,070,035 pornographic and 1,150,295 non-pornographic). As a pre-processing step for the video datasets, we started by extracting frames from videos (20 frames per video), following two different strategies: the **middle extraction strategy**, which focuses on the central portion of the video, under the assumption that the most relevant content of a video is likely to be located near its middle; and the **evenly-spaced extraction strategy**, which extracts frames at regular intervals throughout the video. We split the data into train (80%) and test (20%) partitions, ensuring that frames from the same video were kept in the same partition. We reserve 10% of the train as the validation partition.

[3]Access to these datasets, which contain sexually explicit content, is restricted and requires a formal request. Researchers wishing to use these data must submit a request to corresponding authors and agree to the legal terms and conditions governing their use. This ensures that the data will be used exclusively for research purposes and that researchers fully understand and comply with all legal obligations.

[2]https://github.com/margaridav27/end-to-end-framework-pornography-detection [4]https://gvis.unileon.es/datasets-apd-2m/

B. 2D Pornography Detection

In a binary classification setting (i.e., pornography vs. non-pornography), we fine-tuned ten CNNs — ResNet (50, 101, 152) [19], DenseNet (121, 169, 201) [20], AlexNet [21], VGG (16, 19) [22], and MobileNetV2 [23] — and two transformers — Vision Transformer (ViT) [24] and Data-efficient image Transformer (DeiT) [25] — pre-trained on ImageNet [26]. We trained every model using the stochastic gradient descent (SGD) as the optimisation algorithm and the binary cross entropy as the loss function. Depending on the dataset, the number of training epochs varied between 5 (APD-2M), 50 (Pornography-2k) or 100 (Pornography-800), and the learning rate and its scheduler varied between 1×10^{-4} with *Lambda* (Pornography-800), and 1×10^{-3} with *Step* (Pornography-2k, APD-2M). We performed experiments using images from both frame extraction strategies. All images were resized to 224×224 and normalised with a z-score normalisation using ImageNet's mean and standard deviation. Regarding data augmentation, we applied random cropping, random rotations, random brightness, contrast, and saturation changes, and random gamma correction. For the video datasets (i.e., Pornography-800 and Pornography-2k), we evaluated every model on its test set using the two frame extraction (FE) strategies, middle extraction (M) and evenly-spaced extraction (E), whereas each transformer was evaluated using only the latter, both with (Y) and without (N) data augmentation (DA). We applied the same evaluation strategy to the APD-2M dataset, except for the FE strategy. Regarding performance metrics, we computed the Accuracy (Acc.), Precision (Prec.), Recall (Rec.), and F1-Score (F1) [27]. Additionally, we also performed cross-dataset testing, to assess the robustness and generalisability of the models to different data distributions.

C. Explaining Pornography Detection

In this work, we generated visual explanations only for the correctly predicted cases and tested several methods to ensure that our interpretation of results did not depend solely on a single method. For the CNN-based architectures, we performed experiments with Integrated Gradients (IG) [12], DeepLIFT [13], Layer-wise Relevance Propagation (LRP) [14] with different propagation rules, and Occlusion [15]. In addition, we report results obtained for two different Python libraries, `Captum` [28] and `Zennit` [29]. For the transformer-based architectures, we used the LRP framework designed by Chefer et al. [30]. Moreover, we leveraged the `Quantus` library [16] to objectively assess the quality of these explanations against three criteria - Faithfulness, Robustness, and Complexity. Regarding Faithfulness, we generated results for Faithfulness Correlation (Correl.), Selectivity (Select.), and Region Perturbation (Reg. Pert.). Regarding Robustness, we computed Max Sensitivity (Max Sens.), Relative Input Stability (RIS), and Relative Output Stability (ROS). Finally, for Complexity, we calculated Sparseness (Sparse.) and Complexity (Complex.).

D. End-to-End Approach to Anonymous Visual Explanations

Fig. 1 presents our proposed framework to generate privacy-preserving explanations. Given an input image, it is simultaneously fed into both the pornography detection and face detection models. For the pornography detection model, the image is resized and normalised as required (see Section III). The model then classifies the image, and an explanation for this decision is generated using an explainability method. In parallel, the original version of the image is fed into the face detection model, which outputs bounding boxes with coordinates of detected faces. Having mapped these coordinates onto the resized image, the framework blurs the detected faces in the image and overlays the explanation, preserving privacy without compromising the model's explainability. We argue that this end-to-end process ensures that explanations are informative and privacy-preserving, addressing the sensitive nature of pornography detection.

IV. Results and Discussion

The detection results for the video datasets (Pornography-800 and Pornography-2k) in Table I, the APD-2M in Table II, as well as the cross-dataset results in Table III, highlight the varying performance of CNN and transformer models depending on the datasets they are fine-tuned on. The best results by architecture are highlighted in bold, and the best results overall are underlined. Only the models that achieved the highest value for the most metrics on at least one of the datasets are presented. Models trained on diverse and dynamic datasets like Pornography-2k generalise better across different datasets, whereas those trained on static datasets like APD-2M tend to overfit and perform poorly on video datasets. Notably, all models tested on APD-2M show increased precision despite performance drops in other metrics, suggesting that APD-2M's pornographic instances are more explicit and easier to detect. This raises questions about the effectiveness of state-of-the-art models tested only on curated datasets, emphasising the importance of cross-dataset testing. Furthermore, transformer models demonstrate superior performance and generalisation compared to CNN ones. The evenly-spaced frame extraction strategy is more effective, offering diverse dataset representation. However, data augmentation's impact is mixed, sometimes improving specific metrics but also introducing noise. Overall, the findings underscore the importance of dataset diversity and careful consideration of training data characteristics for developing robust and generalisable pornography detection models.

This study also explores the decision-making processes of the best-performing CNN and transformer models. Fig. 2 illustrates one example of the generated explanations. Our observational analysis revealed that these models rely on a broader range of cues beyond explicit content, underscoring once more the nuanced and complex nature of pornographic content. Models fine-tuned on Pornography-2k frequently focused on logos, whereas those fine-tuned on Pornography-800 did not, suggesting areas for further exploration and

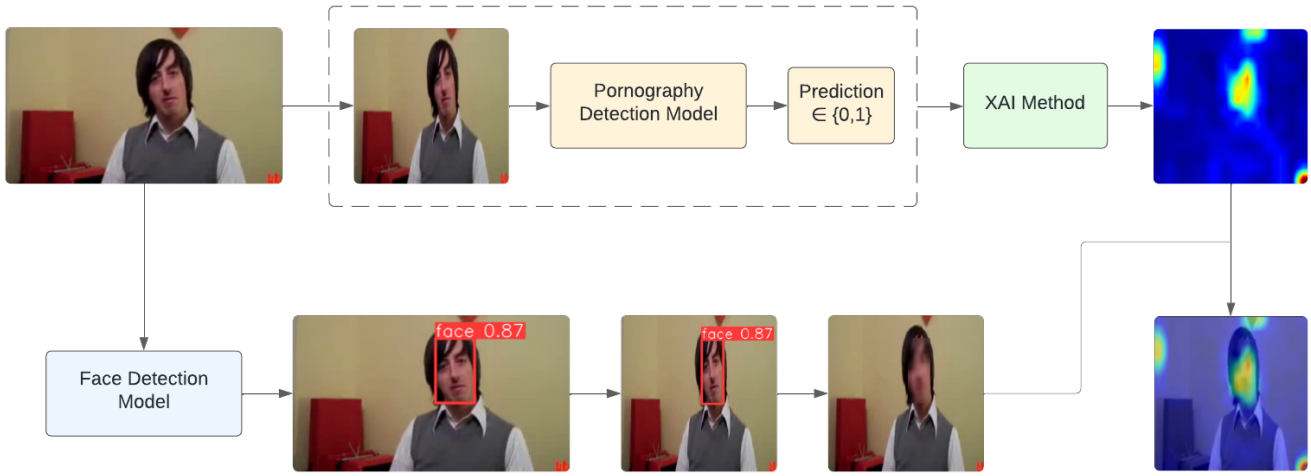


Fig. 1: End-to-end framework for generating privacy-preserving explanations in the context of pornography detection.

TABLE I: Performance metrics for the best CNN and transformer models fine-tuned and tested on Pornography-800 and Pornography-2k datasets.

Model	FE	DA	Pornography-800 Acc.	Pornography-800 Prec.	Pornography-800 Rec.	Pornography-800 F1	Pornography-2k Acc.	Pornography-2k Prec.	Pornography-2k Rec.	Pornography-2k F1
VGG19	E	N	**92.19**	**90.47**	**94.31**	**92.35**	92.66	91.92	93.63	92.77
		Y	89.34	86.62	93.06	89.73	**93.50**	93.42	93.66	**93.54**
ViT	E	N	**96.22**	**96.19**	96.25	**96.22**	96.03	96.48	95.57	96.03
		Y	95.94	95.48	**96.44**	95.96	**96.27**	**96.64**	**95.92**	**96.28**

TABLE II: Performance metrics for the best CNN and transformer models fine-tuned and tested on the APD-2M dataset.

Model	DA	Acc.	Prec.	Rec.	F1
ResNet152	Y	**99.97**	99.96	**99.98**	**99.97**
DenseNet201	Y	**99.97**	**99.97**	99.97	**99.97**
ViT	N	**99.97**	99.97	**99.97**	**99.97**
	Y	**99.97**	99.97	**99.97**	**99.97**

improvement. The XAI quantitative evaluation results, presented in Table IV, show negative faithfulness correlation values for some methods, indicating that current XAI methods may not accurately be capturing the models' behaviour. The differential selectivity between pornographic and non-pornographic content, as revealed by density plots, indicates that current methods effectively identify critical features in pornographic instances, with a peak on the negative side. However, robustness metrics highlight challenges in achieving stable explanations. Overall, the study emphasises the need for improved XAI methods to accurately capture complex model behaviours and address the loosely defined concepts inherent in this type of content.

V. Conclusions and Future Work

A. Conclusions

This study investigates the efficacy of pre-trained CNN and transformer models in 2D pornography detection, emphasising the importance of thoughtful dataset selection and fine-tuning. Transformers, particularly the ViT model, outperformed CNNs due to their ability to capture long-range dependencies and contextual information. The study also highlights the limitations of current explainability methods, with the Occlusion method providing the most human-interpretable explanations for CNNs and transformer explanations generally being the most aligned with our human intuition and expectations. The results underscore the need for improved XAI methods to better capture the nuanced nature of pornographic content.

B. Future Work

Future research should focus on evaluating models on more complex video datasets to better assess generalisation capabilities and performance and on developing methods to automatically detect and remove confounding elements, such as logos. Additionally, techniques to minimise the number of irrelevant or confounding frames in video datasets should be explored. Further investigation into more explainability methods is encouraged to improve explanations' faithfulness, stability, and human interpretability. Enhancing face detection tools with well-labelled facial datasets specific to pornography content and exploring more advanced facial anonymisation techniques, including the use synthetically generated data, are also crucial for privacy preservation in this sensitive domain.

References

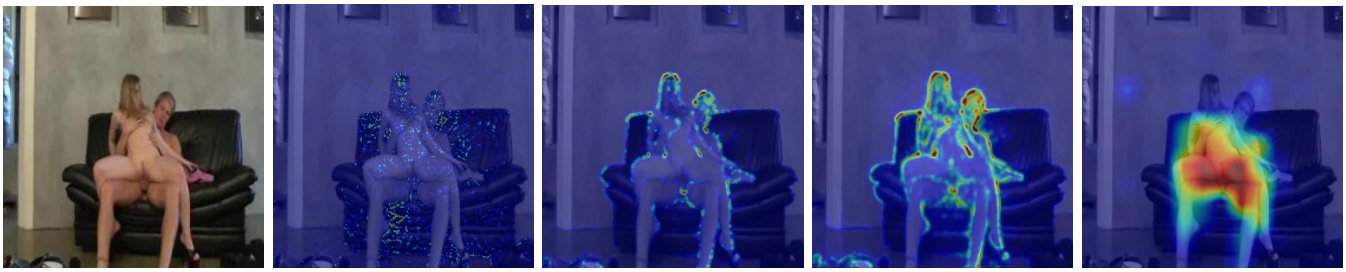
[1] K. C. Seigfried-Spellar, "Assessing the Psychological Well-being and Coping Mechanisms of Law Enforcement Investigators vs. Digital Forensic Examiners of Child Pornography Investigations," *Journal of Police and Criminal Psychology*, pp. 215–226, 2018. [Online]. Available: https://doi.org/10.1007/s11896-017-9248-7

[2] J. B. Grubbs and S. W. Kraus, "Pornography Use and Psychological Science: A Call for Consideration," *Current Directions in Psychological Science*, pp. 68–75, 2021, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.1177/0963721420979594

[3] H. Lee, S. Lee, and T. Nam, "Implementation of high performance objectionable video classification system," in *2006 8th International Conference Advanced Communication Technology*, 2006, pp. 4 pp.–962.

TABLE III: Performance metrics for the cross-dataset experiments. The values in parentheses represent the percentage change in performance when the best-performing models are tested on a different dataset compared to their original fine-tuning dataset.

Fine-tuning Dataset	Testing Dataset	Model	FE	DA	Acc.	Prec.	Rec.	F1
Pornography-800	Pornography-2k	VGG19	E	N	87.12 (↓ 5.50%)	91.14 (↑ 0.74%)	82.39 (↓ 12.64%)	86.54 (↓ 6.29%)
		ViT			93.31 (↓ 3.02%)	95.80 (↓ 0.40%)	90.67 (↓ 5.79%)	93.16 (↓ 3.18%)
	APD-2M	VGG19			89.55 (↓ 2.86%)	82.51 (↓ 8.80%)	99.42 (↑ 5.42%)	90.18 (↓ 2.35%)
		ViT			95.50 (↓ 0.75%)	91.52 (↓ 4.85%)	99.93 (↑ 3.82%)	95.54 (↓ 0.71%)
Pornography-2k	Pornography-800	VGG19	E	Y	96.97 (↑ 3.71%)	95.74 (↑ 2.48%)	98.31 (↑ 4.96%)	97.01 (↑ 3.71%)
		ViT			98.62 (↑ 2.44%)	98.38 (↑ 1.80%)	98.88 (↑ 3.09%)	98.63 (↑ 2.44%)
	APD-2M	VGG19			96.27 (↑ 2.96%)	95.82 (↑ 2.57%)	96.47 (↑ 3.00%)	96.14 (↑ 2.78%)
		ViT			98.64 (↑ 2.46%)	97.45 (↑ 0.84%)	99.80 (↑ 4.04%)	98.61 (↑ 2.42%)
APD-2M	Pornography-800	ResNet152	-	Y	82.06 (↓ 17.92%)	87.39 (↓ 12.58%)	74.94 (↓ 20.04%)	80.69 (↓ 19.29%)
		DenseNet201			78.12 (↓ 21.86%)	85.38 (↓ 14.59%)	67.87 (↓ 32.11%)	75.63 (↓ 24.35%)
		ViT		N	87.38 (↓ 12.59%)	90.57 (↓ 9.40%)	83.44 (↓ 16.53%)	86.86 (↓ 13.11%)
		ViT		Y	87.91 (↓ 12.06%)	92.38 (↓ 7.59%)	82.63 (↓ 17.35%)	87.23 (↓ 12.74%)
	Pornography-2k	ResNet152		Y	84.27 (↓ 15.70%)	93.18 (↓ 6.78%)	74.13 (↓ 25.86%)	82.57 (↓ 17.41%)
		DenseNet201			83.17 (↓ 16.81%)	93.27 (↓ 6.70%)	71.69 (↓ 28.29%)	81.07 (↓ 18.91%)
		ViT		N	87.12 (↓ 12.85%)	95.00 (↓ 4.97%)	78.51 (↓ 21.47%)	85.97 (↓ 14.00%)
		ViT		Y	87.22 (↓ 12.75%)	96.32 (↓ 3.65%)	77.54 (↓ 22.44%)	85.92 (↓ 14.05%)



(a) Orig. Image (b) IG — Zennit (c) LRP EGB — Zennit (d) LRP EPF — Zennit (e) Chefer et al. [30]

(f) IG — Captum (g) Captum — DeepLIFT (h) LRP — Captum (i) LRP Comp. — Captum (j) Occlusion — Captum

Fig. 2: Explanations generated for a Pornography-800's example using all the experimented methods for the best-performing CNN and transformer models fine-tuned on the said dataset. The colour map (jet) ranges from blue (indicating low attribution), through green and yellow to red (indicating high attribution).

[4] A. P. B. Lopes, S. E. d. Avila, A. N. Peixoto, R. S. Oliveira, M. d. M. Coelho, and A. d. A. Araújo, "Nude Detection in Video Using Bag-of-Visual-Features," in *2009 XXII Brazilian Symposium on Computer Graphics and Image Processing*, 2009, pp. 224–231, iSSN: 2377-5416. [Online]. Available: https://ieeexplore.ieee.org/document/5395206

[5] Y. Liu, Y. Yang, H. Xie, and S. Tang, "Fusing audio vocabulary with visual features for pornographic video detection," *Future Generation Computer Systems*, pp. 69–76, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X12001689

[6] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "Pornography classification: The hidden clues in video space–time," *Forensic Science International*, pp. 46–61, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0379073816304169

[7] M. Moustafa, "Applying deep learning to classify pornographic images and videos," in *arXiv*. arXiv, 2015, arXiv:1511.08899 [cs]. [Online]. Available: http://arxiv.org/abs/1511.08899

[8] M. Perez, S. Avila, D. Moreira, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "Video pornography detection through deep learning techniques and motion information," *Neurocomputing*, pp. 279–293, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231216314928

[9] K. H. Song and Y.-S. Kim, "Pornographic Video Detection Scheme Using Multimodal Features," in *arXiv*, 2018, pp. 1174–1182. [Online]. Available: https://medwelljournals.com/abstract/?doi=jeasci.2018.1174.1182

[10] N. Gautam and D. K. Vishwakarma, "Obscenity Detection in Videos Through a Sequential ConvNet Pipeline Classifier," *IEEE Transactions on Cognitive and Developmental Systems*, pp. 310–318, 2023, conference Name: IEEE Transactions on Cognitive and Developmental Systems. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9733936?casa_token=_hRtlk8bcYMAAAAA:7hUADUtHg9jaIOx6WdwbAHzZM3c687zKKRPZxgJbYqMTpU_cuRPZil7d-ro8vt1f8kmo7yhAewo&signout=success

[11] T. Speith, "A review of taxonomies of explainable artificial intelligence (xai) methods," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 2239–2250.

TABLE IV: Results on XAI quality assessment obtained for the VGG19 and the ViT fine-tuned using the evenly-spaced frame extraction strategy without data augmentation for Pornography-800. For each metric ↑/↓ means higher/lower is better.

(a) Captum [31].

Method	Faithfulness			Robustness		Complexity	
	Correl. (↑)	Select. (↓)	Reg. Pert. (↓)	Max Sens. (↓)	RIS (↓)	ROS (↓)	Sparse. (↑)	Complex. (↓)
IG	-1.3920×10^{-2}	3.8677×10^{2}	3.7754×10^{2}	6.6403	2.1275×10^{6}	3.7115×10^{9}	5.9422×10^{-1}	1.0173×10^{1}
DeepLIFT	-8.6127×10^{-3}	3.5748×10^{2}	4.0073×10^{2}	3.0581	2.8396×10^{5}	1.9999×10^{8}	7.9968×10^{-1}	1.0373×10^{1}
LRP	-1.1422×10^{-2}	4.5839×10^{2}	3.3968×10^{2}	7.2195	5.9373×10^{5}	2.1399×10^{8}	6.0151×10^{-1}	1.0156×10^{1}
LRP Comp.	4.5448×10^{-3}	3.8480×10^{2}	4.1920×10^{2}	2.8186	3.7223×10^{6}	1.5766×10^{10}	7.3736×10^{-1}	9.6248
Occlusion	8.4555×10^{-2}	6.7961×10^{2}	4.1738×10^{2}	4.7684	7.6138×10^{3}	3.3976×10^{7}	4.8241×10^{-1}	1.0423×10^{1}

(b) Zennit [29].

Method	Faithfulness			Robustness		Complexity	
	Correl. (↑)	Select. (↓)	Reg. Pert. (↓)	Max Sens. (↓)	RIS (↓)	ROS (↓)	Sparse. (↑)	Complex. (↓)
IG	-1.2449×10^{-2}	3.8649×10^{2}	3.7762×10^{2}	7.8293	6.2918×10^{5}	4.0624×10^{8}	5.9412×10^{-1}	1.0174×10^{1}
LRP (EGB)	1.9219×10^{-2}	2.9955×10^{2}	4.4062×10^{2}	8.8220×10^{2}	2.2823×10^{6}	1.3709×10^{10}	8.7476×10^{-1}	1.0372×10^{1}
LRP (EPF)	2.9687×10^{-2}	3.5828×10^{2}	4.2008×10^{2}	5.1061×10^{3}	3.7164×10^{5}	2.6417×10^{8}	5.3215×10^{-1}	1.0278×10^{1}

(c) Chefer et al. [30].

Faithfulness			Robustness		Complexity	
Correl. (↑)	Select. (↓)	Reg. Pert. (↓)	Max Sens. (↓)	RIS (↓)	ROS (↓)	Sparse. (↑)	Complex. (↓)
7.0540×10^{-2}	8.5463×10^{2}	1.0710×10^{2}	1.2413×10^{1}	3.4887×10^{1}	1.8171×10^{5}	4.6425×10^{-1}	1.0406×10^{1}

[Online]. Available: https://doi.org/10.1145/3531146.3534639

[12] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in *arXiv*. arXiv, 2017, arXiv:1703.01365. [Online]. Available: http://arxiv.org/abs/1703.01365

[13] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," in *arXiv*. arXiv, 2019, arXiv:1704.02685. [Online]. Available: http://arxiv.org/abs/1704.02685

[14] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLOS ONE*, p. e0130140, 2015, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140

[15] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *arXiv*. arXiv, 2013, arXiv:1311.2901. [Online]. Available: http://arxiv.org/abs/1311.2901

[16] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, and M. M. M. Höhne, "Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond," *Journal of Machine Learning Research*, pp. 1–11, 2023. [Online]. Available: http://jmlr.org/papers/v24/22-0142.html

[17] S. Avila, N. Thome, M. Cord, E. Valle, and A. de A. Araújo, "Pooling in image representation: The visual codeword point of view," *Computer Vision and Image Understanding*, pp. 453–465, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314212001737

[18] A. Gangwar, V. González-Castro, E. Alegre, and E. Fidalgo, "AttM-CNN: Attention and metric learning based CNN for pornography, age and Child Sexual Abuse (CSA) Detection in images," *Neurocomputing*, vol. 445, pp. 81–104, Jul. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092523122100312X

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *arXiv*. arXiv, 2015, arXiv:1512.03385 [cs]. [Online]. Available: http://arxiv.org/abs/1512.03385

[20] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *arXiv*. arXiv, 2018, arXiv:1608.06993 [cs]. [Online]. Available: http://arxiv.org/abs/1608.06993

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[22] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *arXiv*. arXiv, 2015, arXiv:1409.1556 [cs]. [Online]. Available: http://arxiv.org/abs/1409.1556

[23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *arXiv*. arXiv, 2019, arXiv:1801.04381 [cs]. [Online]. Available: http://arxiv.org/abs/1801.04381

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *arXiv*. arXiv, 2021, arXiv:2010.11929 [cs].

[25] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *arXiv*. arXiv, 2021.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[27] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[28] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for pytorch," in *arXiv*, 2020. [Online]. Available: https://arxiv.org/abs/2009.07896

[29] C. J. Anders, D. Neumann, W. Samek, K.-R. Müller, and S. Lapuschkin, "Software for dataset-wide xai: From local explanations to global insights with Zennit, CoRelAy, and ViRelAy," *CoRR*, 2021.

[30] H. Chefer, S. Gur, and L. Wolf, "Transformer Interpretability Beyond Attention Visualization," in *arXiv*. arXiv, 2021, arXiv:2012.09838. [Online]. Available: http://arxiv.org/abs/2012.09838

[31] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, and S. Lapuschkin, "Towards Best Practice in Explaining Neural Network Decisions with LRP," in *arXiv*. arXiv, 2020, arXiv:1910.09840. [Online]. Available: http://arxiv.org/abs/1910.09840