Forensic Science International: Digital Investigation 42-43 (2022) 301436

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

Detecting and ranking pornographic content in videos

Mark Borg [a, *], André Tabone [a], Alexandra Bonnici [a], Stefania Cristina [a], Reuben A. Farrugia [b], Kenneth P. Camilleri [a]

[a] Department of Systems and Control Engineering, University of Malta, Malta
[b] Department of Communications and Computer Engineering, University of Malta, Malta

ARTICLE INFO

Article history:
Received 10 February 2022
Received in revised form
12 July 2022
Accepted 15 July 2022
Available online 4 August 2022

Keywords:
Pornography detection
Sexual object detection
Deep learning

ABSTRACT

The detection and ranking of pornographic material is a challenging task, especially when it comes to videos, due to factors such as the definition of what is pornographic and its severity level, the volumes of data that need to be processed, as well as temporal ambiguities between the benign and pornographic portions of a video.

In this paper we propose a video-based pornographic detection system consisting of a convolutional neural network (CNN) for automatic feature extraction, followed by a recurrent neural network (RNN) in order to exploit the temporal information present in videos. We describe how our system can be used for both video-level labelling as well as for localising pornographic content within videos. Given pornographic video segments, we describe an efficient method for finding sexual objects within the segments, and how the types of detected sexual objects can be used to generate an estimate of the severity ('harmfulness') of the pornographic content. This estimate is then utilised for ranking videos based on their severity, a common requirement of law enforcement agencies (LEAs) when it comes to categorising pornographic content.

We evaluate our proposed system against a benchmark dataset, achieving results on par with the state of the art, while providing additional benefits such as ranking videos according to their severity level, something which to the best of our knowledge has not been attempted before. We perform further investigations into model generalisability by performing an out-of-distribution (o.o.d.) test, investigate whether our model is making use of shortcut learning, and address the issue of explainability. The results obtained indicate that our model is using strong learning, thus further validating our proposed approach and the results obtained.

1. Introduction

The field of pornography detection encompasses a wide and varied set of applications and use cases, ranging from image and video content filtering (Kelly et al., 2008; Behrad et al., 2012; Moreira et al., 2016), automatic censoring of pornographic material (via removing, obfuscating, or blurring the offending material) (More et al., 2018; de Freitas et al., 2019; Mallmann et al., 2020), to pornography type classification (Oronowicz-Jaskowiak, 2018).

With regards to content filtering, this can involve restricting access to pornographic material which though legal, is not suitable for certain demographics, like minors, or for certain environments, such as schools or workplaces (de Freitas et al., 2019; Zhelonkin and Karpov, 2020). Another use case is the automatic detection of pornographic material which is illegal, such as tools to assist LEAs in detecting child sex abuse material (CSAM) (Gangwar et al., 2017; Macedo et al., 2018; Lee et al., 2020; Laranjeira da Silva et al., 2022).

Although many advancements have been done in the field of pornography detection, numerous challenges still remain. This is made more difficult by the fact that there is no universal agreement of what differentiates pornographic from benign material, and this normally depends on various factors such as culture, context and the application domain (Short et al., 2012; McKee et al., 2020). Much of the research work in pornography detection deals with image content rather than video material (Oronowicz-Jaskowiak, 2018; Shen et al., 2018; Vitorino et al., 2018; Zhelonkin and

* Corresponding author.
E-mail addresses: mark.j.borg@um.edu.mt (M. Borg), andre.tabone@um.edu.mt (A. Tabone), alexandra.bonnici@um.edu.mt (A. Bonnici), stefania.cristina@um.edu.mt (S. Cristina), reuben.farrugia@um.edu.mt (R.A. Farrugia), kenneth.camilleri@um.edu.mt (K.P. Camilleri).

https://doi.org/10.1016/j.fsidi.2022.301436
2666-2817/

Karpov, 2020; Al Dahoul et al., 2021; Gangwar et al., 2021). Videos present a number of additional challenges to pornography detection, due to the larger amounts of data involved, real-time operating constraints, and the typically more ambiguous nature of what constitutes pornographic actions on a temporal scale. A commonly-adopted approach is to reduce the task of video pornography detection to an image-based one, by running pornography detection on a frame-by-frame basis, or on a selected subset of video frames, typically the keyframes (Caetano et al., 2014; Moustafa, 2015; Jung et al., 2017; da Silva and Marana, 2019; Mallmann et al., 2020). One disadvantage of this approach is that temporal information present in videos is not exploited to the full.

In this paper, we propose a solution to tackle video-based pornography detection that utilises the full video data available. Our solution employs an efficient model that is able to run on low-end hardware, operating at near real-time frame rate. We also describe how our solution can go beyond a simple pornography/benign classification, by including other functionality such as pornographic segment localisation and video severity ranking in order to support a wider range of applications. In the next section we provide a brief literature review, and describe how our proposed solution compares with the state of the art.

2. Literature review

Early works in pornography detection tackled mostly nudity detection, using either skin colour or texture based techniques (Zheng and Daoudi, 2004; Karavarsamis et al., 2013), or local features combined with bag of visual words (BoVW) approaches (Caetano et al., 2014; Valle et al., 2011; Avila et al., 2013). More recently, deep learning (DL)-based approaches demonstrated a marked improvement in accuracy, as well as the ability to work with more challenging datasets (Moustafa, 2015; Wehrmann et al., 2017; Perez et al., 2017; da Silva and Marana, 2019).

While image-based pornography detection is by far more prevalent, a number of DL-based works have attempted pornography detection on videos. A common way of doing this is to perform pornography detection on each separate video frame (or selected keyframes), and then combining the frame-wise results into a video-level label via majority voting (Moustafa, 2015; Jung et al., 2017; Qamar Bhatti et al., 2018).

In order not to lose important temporal/motion information, other DL-based works make use of the full video data (or video clips with multiple frames) (Wehrmann et al., 2017; Perez et al., 2017; da Silva and Marana, 2019; Song and Kim, 2020). For example, Perez et al. (2017) propose a two-stream CNN system, with one stream for image-based features, while the second utilising motion features (optical flow and MPEG motion vectors) across frames. da Silva and Marana (2019) employ a 3D CNN to operate on the temporal data, while Wehrmann et al. (2017),Song and Kim (2020), and Yousaf and Nawaz (2022), use a CNN for feature extraction, followed by an RNN for temporal reasoning.

For our solution, we adopt a similar approach to that of the last three works, that is, a combination of CNN and RNN. Like them, we make use of a majority voting scheme to arrive at the video-level label. However, we then also localise the pornographic content within videos and extract these segments for further processing, something which to the best of our knowledge is not attempted by any other pornography detection work.

On the extracted video segments we perform sexual object detection (SOD), and then rank the video segments based on their estimated severity of pornographic content. Several other DL-based works have attempted SOD on image content (Shen et al., 2018; Mallmann et al., 2020; Tabone et al., 2020, 2021). Both Shen et al. (2018) and Mallmann et al. (2020) utilise a CNN-based object detector to detect four and six sexual objects respectively, with the latter work additionally employing a Bayesian network to exploit the contextual information between the sexual objects. Tabone et al. (2020) adopt a two-step approach: first performing detection to quickly distinguish pornographic images from benign ones, then running an object detector to detect nine sexual objects. In our work, we utilise the method of Tabone et al. (2020) for SOD.

As regards to the ranking of video segments based on severity ('harmfulness') of their pornographic content, limited work has been attempted so far. Oronowicz-Jaskowiak (2018) classifies pornographic images into seven different categories (BDSM; sexual activities: individual, group, and animated; fetishisms: feet, knee high socks, and paraphilic infantilism), but this is not used for image ranking purposes or for estimating the severity of the pornographic content. Tabone et al. (2021) also adopt a classification approach, classifying images into 19 types and relying on the class label description for a rudimentary ranking of the images. Laranjeira da Silva et al. (2022) list nudity levels and age as examples of classification criteria. Vitorino et al. (2018), mention the need for further research into automatic severity estimation of CSAM based on the COPINE scale (Taylor et al., 2001), but little follow-up research appears to have been done so far.

The rest of this paper is structured as follows: we first introduce the datasets used in our experiments; then describe our proposed approach, followed by the experimental analysis, and evaluation; we finish off the paper with investigations into model generalisation and explainability, and finally the concluding remarks.

3. Datasets

Publicly-available video-based datasets for pornography detection are limited both in number and in size.

3.1. The NPDI dataset

The NPDI dataset (Avila et al., 2013), considered a benchmark dataset, consists of nearly 80 h of 400 pornographic and 400 benign videos, obtained from public websites. These videos are post-processed to identify video shots, and then a key frame (the middle video frame) is extracted for each video shot. One limitation of this dataset is that the key frames are sampled at a very low and irregular rate. Furthermore, for many of the videos, only few key frames are available: 126 out of the 800 videos (15.7%) consist of only a single key frame.

3.2. The APD-VIDEO dataset

To address and overcome the limitations of existing datasets, we created our own dataset, which we call APD-VIDEO (University of Malta, 2021). We make this dataset publicly available.

3.2.1. Dataset creation

Videos taken from the *Kaggle Thumbzilla* pornography list[1] constitute the positive videos. This is a publicly-available list of 191,532 videos, plus metadata. The Thumbzilla website is actually a video aggregator, meaning that it collects and organises videos from several other websites, thus ensuring diversity in content and format. Since the number of videos available in this list is very large, we only download a randomly sampled subset of these videos.

For the negative (benign) class, we take videos from the *MPII Human Pose dataset* (Andriluka et al., 2014). The MPII dataset[2] has

[1] Available at: https://www.kaggle.com/ljlr34449/porn-data.
[2] Available at: http://human-pose.mpi-inf.mpg.de/.

Table 1
APD-VIDEO dataset distribution.

	Positive videos	Negative videos	
	Kaggle Thumbzilla	MPII dataset	Youtube confusable
Total videos in original dataset	191,352	2821 (18,080 labelled activities)	115
Downloaded videos	4215	2435	115
Video segments	8201	7289	420
Total video frames	7,237,218	3,383,823	983,414

410 different activities, some of which are of special interest to our classification problem, such as: pilates, aerobic, calisthenics, breastfeeding, gymnastics, massage, therapy, yoga, volleyball, reclining, workouts, and boxing, among others.

We augment the negative set with a list of 115 publicly-available YouTube videos (which we call 'YouTube-confusable'), containing videos that are easily confused with the positive class. For example, we have film trailers of a sexually suggestive nature, fashion shows, swimwear fashion shows, boxing, wrestling, yoga, bodybuilding competitions, and sumo wrestling. Table 1 gives a breakdown of the number of videos downloaded for the positive and negative sets. We attempt to keep the training portions of the two sets approximately equal (in terms of video frames), in order to get a balanced dataset.

3.2.2. Video segments

For training purposes, we extract video segments of a predefined length (30 s) from each downloaded video, sampled at 12 frames per second (FPS). For the Kaggle Thumbzilla videos and the Youtube-confusable videos, we select random positions from where we extract the video segments. In the case of the MPII dataset, since this dataset comes with 18,080 labelled positions of human activity segments, we randomly choose a subset of these existing and labelled video segments.

3.2.3. Groundtruthing

For training and evaluation purposes, we perform framewise labelling of the video segments and the test videos, in addition to just using a single label for the whole video. For determining the labels, we follow the definition of pornography as given by Short et al. (2012).

At this stage, only one annotator labelled the video frames. Other challenges encountered during framewise labelling arise from temporal ambiguity (determining the exact boundary between a pornographic scene from a non-pornographic one), how to label sexual activity interspersed with other shots (like close-up shots of the face, background, or persons speaking), as well as the general ambiguity associated with what constitutes pornography. We thus acknowledge that the labels at frame level are noisy and are prone to have annotation errors; later, in §4.2, we describe an approach to mitigate this issue.

4. Proposed approach

Our proposed approach, illustrated in Fig. 1, consists of a pipeline made up of four stages: the first stage performs pornography detection on videos, classifying the videos into pornographic or benign. In the second stage, the framewise results are used to locate segments within the videos that contain pornographic content from other segments that do not. In the third stage, sexual objects (private body parts) are detected and classified within the positive video segments. While in the fourth stage, the detected sexual objects are used to estimate the severity of the pornographic content and to rank the videos according to this severity measure.

4.1. Pornography detection

The first stage of our proposed solution consists of a CNN for automatic feature extraction, working on a frame-by-frame basis, followed by an RNN for temporal reasoning across video frames. This is illustrated in Fig. 1 (top part). We make use of the convolutional base of the binary classifier of Tabone et al. (2020) for the extraction of the CNN-based features. This CNN is composed of a pre-trained MobileNetV2, with the weights fine-tuned by training on a subset of the APD-2M pornographic image corpus (University of Leon), with the addition of benign images taken from the VOC2012 (Everingham et al., 2012), COCO (Lin et al., 2014), and MPII datasets (Andriluka et al., 2014).

The MobileNetV2 architecture (Sandler et al., 2019) consists of a relatively small CNN that makes use of depth-wise separable convolutions in order to achieve a lightweight and efficient model. Thus, it is less reliant on needing a powerful GPU to run. The MobileNetV2 architecture includes linear bottlenecks in between layers to help preserve information as it traverses the network. The output of this network consists of a vector $\mathbf{x}$ of 1024 CNN features per video frame. Given a video with L frames and the corresponding set $\{\mathbf{x}_i\}_1^L$ of CNN features, we employ a sliding window approach to feed the features to the RNN. In our method, we use a sliding window size of 60 video frames, and a window step size based on the ratio of the original FPS of the video to the required processing rate of the system (set to 12 FPS in our experiments).

The adopted RNN architecture is illustrated in Fig. 2. This consists of two layers of bi-directional gated recurrent units (GRUs) (Cho et al., 2014), followed by three fully-connected layers serving as the final binary classifier. The RNN architecture described here was determined after running a number of model selection experiments (described in §5.2). Dropout is used both within the RNN layers, as well as for the first two fully-connected layers of the classifier. Batch normalisation is applied in between the RNN layers and the classifier. The first two fully-connected layers of the classifier use ReLU as the activation function, while the final classification layer uses softmax. Binary cross-entropy is used as the loss function for training the network.

The output of the RNN and the binary classifier consists of a set of frame-wise class predictions $\{y_i\}_1^L$ and their associated confidence scores $\{s_i\}_1^L$, indicating whether the frames up to frame i are considered to be pornographic or not.

4.2. Label smoothing

Working with noisy labels can have a detrimental effect on deep learning systems, as evidenced by several works (Chen et al., 2019; Song et al., 2020). And although other works demonstrated marked robustness to strong label noise (Rolnick et al., 2018), recent studies suggest that deep learning systems achieve this via increased

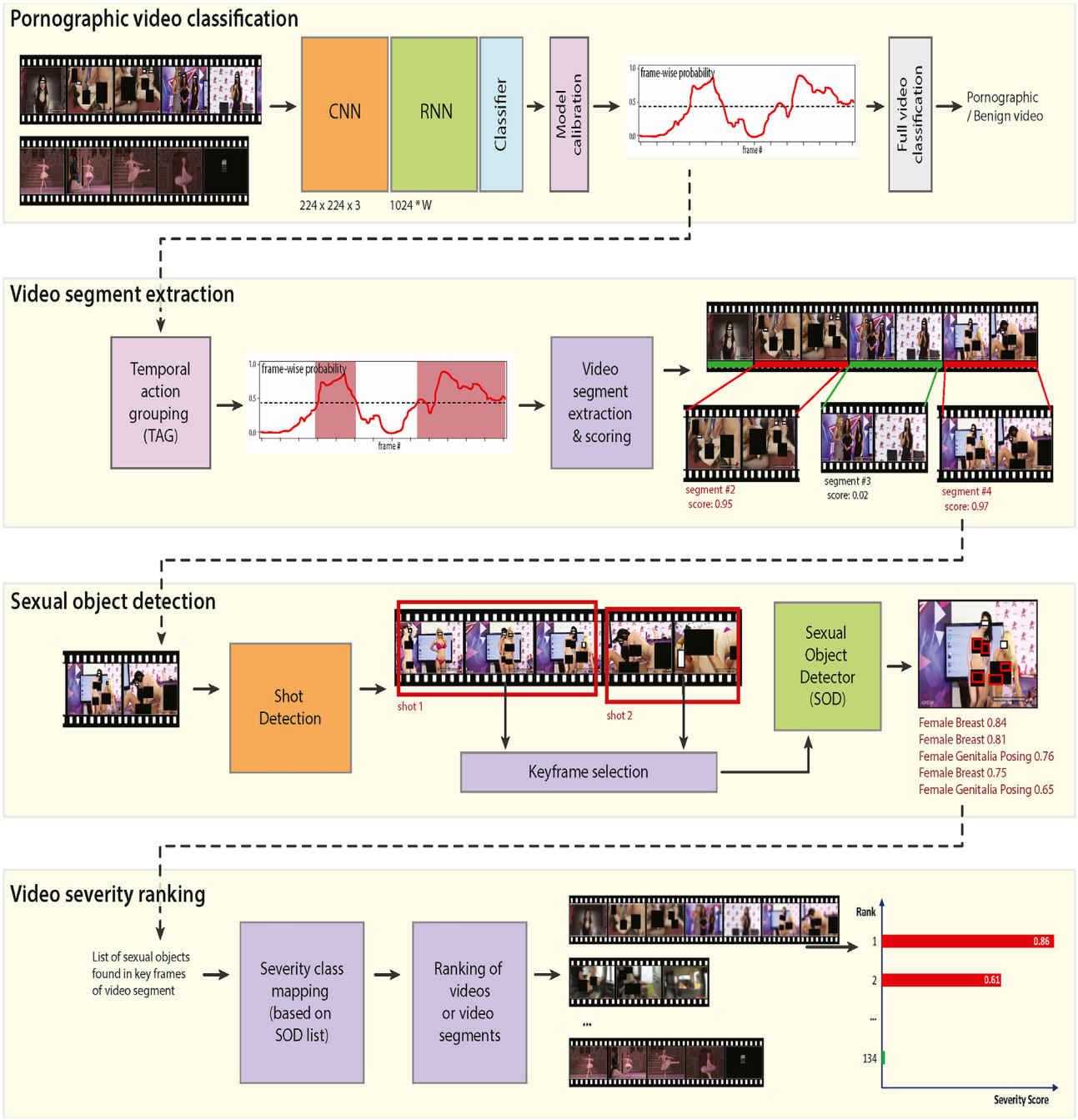

Fig. 1. An overview of the pipeline of our proposed pornographic video detection system. Our system can support different video-based applications, ranging from classifying whole videos into pornographic or benign, locating the pornographic content within a specific video, identifying sexual objects in video frames, and ranking videos in terms of a simple severity measure based on the types of identified sexual objects within the video.

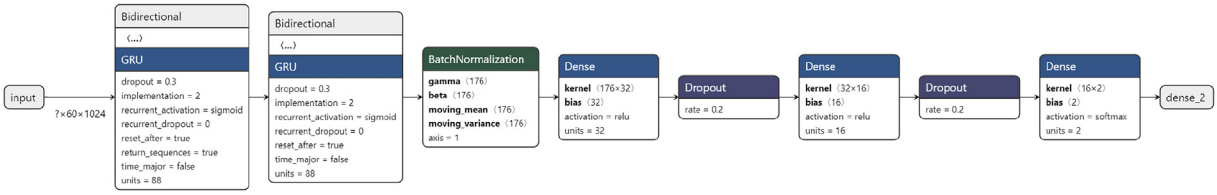

Fig. 2. The proposed RNN model consisting of two bi-directional GRUs, followed by the classifier consisting of three fully-connected layers.

memorisation at the expense of strong learning (Arpit et al., 2017; Karimi et al., 2020). One has to keep in mind that deep learning models are able to memorise large datasets even when these have completely randomised labels (Zhang et al., 2017).

As discussed in §3.2.3, the labelling of the datasets used for this work may be considered to have a high level of noise, mainly due to issues like subjectivity, temporally-related ambiguities, and also compounded by the lack of multiple annotators. Thus we investigate ways of dealing with such noisy labels.

Several different approaches have been proposed in the literature for training with noisy labels including amongst others: label denoising and cleaning techniques, loss function modification, data re-weighting, label smoothing, as well as curriculum learning (Song et al., 2020; Karimi et al., 2020).

For our work, we consider label smoothing (Szegedy et al., 2016) to be the most appropriate approach, due to its reported benefits (backed by empirical evidence) of better generalisation and model calibration (Muller et al., 2019; He et al., 2019), and the fact that we can apply it across all the training samples without the need to identify which specific labels are noisy and which are not. Label smoothing (Szegedy et al., 2016) replaces the hard labels with a softer version, consisting of a weighted combination of the labels themselves and a uniform distribution over the full label set. In essence, it introduces an element of uncertainty in the training labels that reflects the underlying noise in the labelling process. The hard labels are normally specified in terms of a one-hot encoded vector representation as follows:

$$p(y \mid \mathbf{x}_i) = \begin{cases} 1 & \text{if } y = y_i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $\mathbf{x}_i$ is the i^{th} training data sample and its corresponding label y_i. The smoothed version of (1) consists of:

$$p^{(a)}(y \mid \mathbf{x}_i) = \begin{cases} 1 - \varepsilon + \varepsilon\, U(y \mid \mathbf{x}_i) & \text{if } y = y_i \\ \varepsilon\, U(y \mid \mathbf{x}_i) & \text{otherwise} \end{cases} \tag{2}$$

where $\varepsilon \in [0, 1]$ is the weighting factor, and $U(y \mid \mathbf{x}_i) = \frac{1}{K}$ is a uniform distribution, with K being the number of class labels. In our solution, we apply label smoothing globally across all the training samples via (2), with weighting factor ε chosen to be 0.2. Thus, given the label $p(y = \text{pornographic}) \in [0, 1]$, its smoothed version becomes $p^{(a)} \in [0.1, 0.9]$.

During the labelling of the datasets, we experienced increased uncertainty when annotating frames at the boundary between the pornographic and non-pornographic portions of many of the videos. Thus we propose to add further smoothing at such points of transition to reflect this higher labelling uncertainty. We do this by first applying a smoothing factor prior to using (2), with this additional factor given below:

$$p^{(b)}(y \mid \mathbf{x}_i) = |y - \phi\, H(W)| \tag{3}$$

where $\phi \in [0, 1]$ controls the amount of additional smoothing done in transitional regions, W is the temporal window of video frames fed to the RNN at any given moment in time and centred on the frame at position i with label y which needs smoothing, while $H(\cdot)$ measures the entropy over window W according to: $H(W) = H(\text{porn}) + H(\text{benign})$.

In other words, if the video frames in window W all have the same label, then the entropy $H(W)$ factor is 0 and (3) performs no adjustment; if there is a mixture of labels (a transitional video segment), then the adjustment amount will depend on the number of pornographic to non-pornographic frames present in this temporal window, reflecting the potential increased confusion of the annotator in determining the boundary point.

Combining both smoothing factors (3) and (2) together, the final smoothed label for our binary classification problem now becomes:

$$p'(y \mid \mathbf{x}_i) = \begin{cases} 1 - \phi\, H(W) - \varepsilon/2 & \text{if } y = y_i \\ \phi\, H(W) + \varepsilon/2 & \text{otherwise} \end{cases} \tag{4}$$

4.3. Model calibration

Deep learning models in general tend to be highly over-confident in their predictions. This is mostly attributed to their large capacity, and further compounded by how their training strives to lower the cross-entropy loss (Guo et al., 2017; Lakshminarayanan et al., 2017; Karimi et al., 2020). As can be expected, our model also suffers from the problem of over-confidence: the frame-wise scores $\{s_i\}_1^L$ returned by our model do not reflect the true probabilities of whether the video at frame position i is pornographic or not. This miscalibration can be observed in Fig. 3.

A perfectly calibrated classification model would return scores s identical to the true probabilities p:

$$P\Big(y_{\text{pred}} = y_{\text{true}} \mid s = p\Big) = p \quad \forall p \in [0, 1] \tag{5}$$

In other words, given that the classifier generates N predictions all with a confidence score $s = 0.7$, we expect that 70% of these N predictions are correct. From Fig. 3 (red curve) we can see that out of N predictions with score $s = 0.7$ generated by our model, only 40% of these are actually correct, implying an over-confident model.

While an improperly calibrated model does not in general affect its evaluation performance, the classification scores generated by the model would lack interpretability. And if the scores are used for ranking purposes (e.g. ranking videos in terms of their likelihood of containing pornography), or to give an indication to the user or to a downstream component in an application pipeline, it can lead to incorrect or misleading results. It would also create problems if the pornography classifier is used as part of an ensemble of models (for example, an ensemble where each model considers different evidences (Valle et al., 2011); or for example, when fusing video classification with audio classification (Song and Kim, 2020)).

The over-confidence characteristic of a deep learning model can carry over when such a model is applied to a different (but related) domain, or when applied to out-of-distribution (o.o.d.) examples. This can lead to apparent generalisation, i.e., the model still outputting over-confident prediction scores even when faced with more uncertain examples. This is exacerbated if the use of domain adaptation between the source and target domains is not possible or permissible. An example of this situation is a pornography classifier, trained on adult pornography, which is intended for use in detecting CSAM: A commonly-adopted approach for such an application is to combine the adult pornography classifier with an age detector (Jung et al., 2017; Macedo et al., 2018; Islam et al., 2019; Al Nabki et al., 2020; Anda et al., 2020), since training on the target domain data might often be problematic.

For the above reasons, we decided to calibrate our RNN-based pornography classifier. Numerous techniques can be found in the literature addressing model calibration, especially deep learning ones (Guo et al., 2017). A group of techniques are based on either reducing a model's capacity, or increasing its regularisation (Pereyra et al., 2017). While these methods can help improve model calibration, they also tend to suffer from some reduction in accuracy (Guo et al., 2017). More recently, focal loss (Lin et al., 2017) was employed during training instead of cross-entropy loss, since this was found to lead to better calibrated models (Mukhoti et al., 2020). But some re-scaling of the model's output scores will still need to be done after the classification stage.

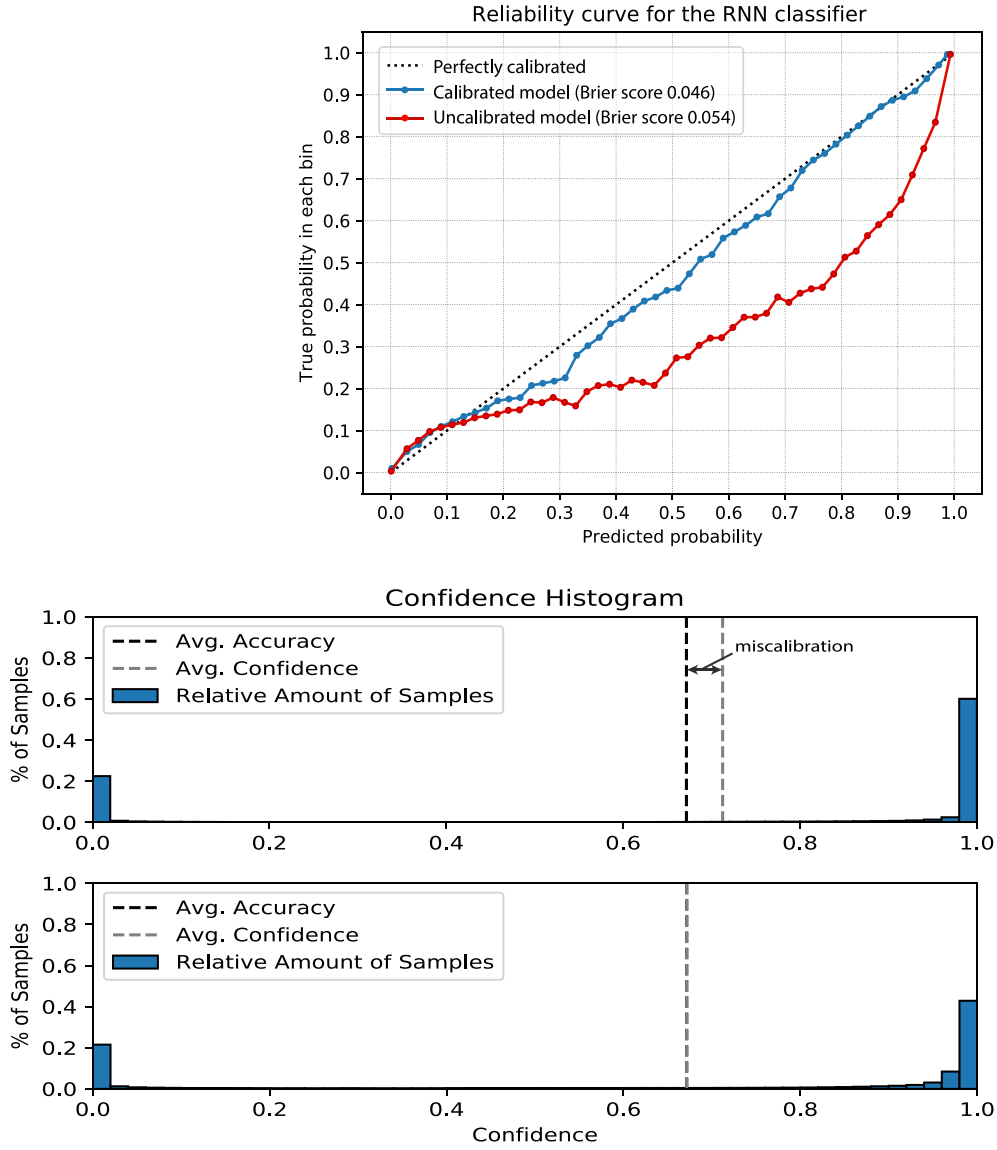

Fig. 3. (Top) Reliability curve of our RNN model, prior (red) and after (blue) model re-calibration, showing a change from over-confident scores to values more in line with the true classification probabilities (diagonal line). (Bottom) The corresponding confidence histogram before and after model calibration. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

In our work we settle for a model re-calibration technique implemented as a post-processing operation, rather than attempting calibration within the deep learning model itself or during its training. We investigated a number of computationally-efficient techniques including temperature scaling (Mozafari et al., 2018), Platt scaling (Platt, 1999), and isotonic regression (Niculescu-Mizil and Caruana, 2005), opting for Platt scaling based on the empirical results obtained (see Fig. 3), and due to its optimality characteristics (Böken, 2021).

Platt scaling (Platt, 1999) operates by essentially fitting a logistic regression to the output scores $\{s_i\}$ of the classifier:

$$p \approx \hat{p} = \frac{1}{1 + e^{As+B}} \tag{6}$$

where p is the true probability, $\hat{p}$ is the re-calibrated score, and A and B are the parameters of the logistic regression. We fit the logistic regression using a hold out set (separate from the training, validation and test sets), resulting in parameter $A = -0.7561$ and $B = 0.6573$. Once the logistic regression model is generated, it is then used to map (scale) the scores of the classifier to yield calibrated values that should now be closer to the true classification probabilities. Fig. 3 shows the model's output confidence scores after performing re-calibration via Platt scaling, while Table 2 gives the reduction in model miscalibration that we obtained, as measured in terms of standard metrics: ECE, MCE, and ACE (Guo et al., 2017; Nixon et al., 2019). Another benefit of a calibrated model is that the optimal threshold obtained from the receiver operating characteristic (ROC) is now closer to 0.5.

4.4. Full video classification

The framewise results $\{y_i\}_1^L$ of the model are then mapped to a video-level classification label by a majority voting scheme. An alternative scheme could be employed, dependent on the application's use case; for example, some law enforcement scenarios might require that a video is classified as pornographic even if a small number of frames are positive.

Table 2
Model re-calibration.

Metric	Uncalibrated model	Re-calibrated model
Expected calibration error (ECE) ↓	0.0471	0.0119
Maximum calibration error (MCE) ↓	0.3314	0.0841
Adaptive calibration error (ACE) ↓	0.1828	0.0302

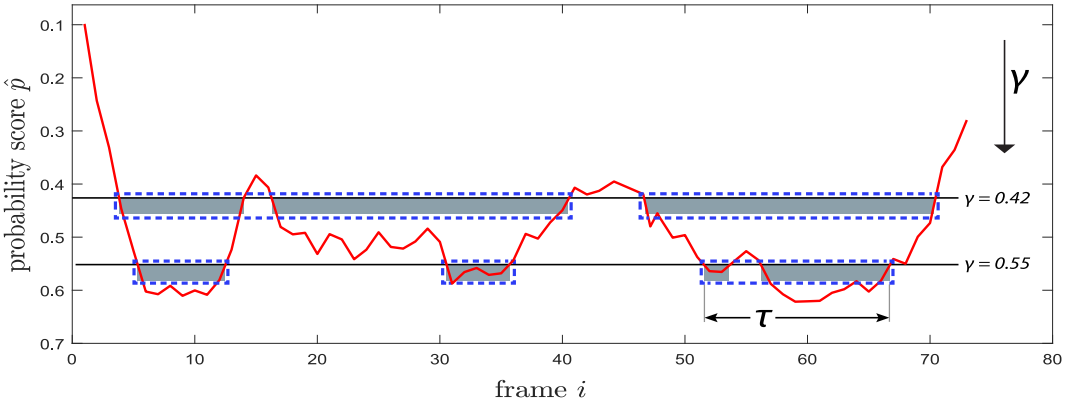

Fig. 4. Temporal action grouping (TAG) method used for the extraction and merging of video segments. In our case we use a single value for γ, determined from the optimal point of the ROC curve of the binary classifier.

4.5. Pornographic video segment extraction

To group the frame-wise results into video segments containing pornographic content, we adopt a method inspired by the temporal actionness grouping (TAG) scheme of Zhao et al. (2017).This method utilises the classical watershed algorithm to identify the video segments in the 1D signal formed by the probability output (re-calibrated scores) of the classifier $\{\widehat{p}_i\}_1^L$. This method is illustrated in Fig. 4: Given a flooding level specified by parameter γ, the corresponding catchment basins (video segments) shown in gray are identified. This is followed by a basin growing (segment merging) technique which joins segments together conditioned on the ratio of their length to the total duration encompassed by the segments being smaller than parameter τ (merged basins outlined in blue).

Zhao et al. (2017) perform a sweep over a range of values for both the flooding parameter γ and the grouping criterion τ, generating action proposals to be then fed to specific action recognisers. In our case, we fix the value of γ and set it to the optimal point of the classifier's ROC curve, since unlike Zhao et al. (2017) we do not have another set of networks to assess the video segments. Then we perform a search for the optimal value of the grouping criterion τ using a validation set. The advantage of adopting this approach, rather than a more rudimentary segment grouping technique, is that minimal changes are required if for future work we add specific pornographic action recognisers to our system.

Once the positive video segments are located, a score for each segment s_j is computed from the framewise results;

$$\text{score}_j = 1 + \frac{\sum_{s_j(0)}^{s_j(1)} \ln(\widehat{p}_i)}{s_j(1) - s_j(0)} \tag{7}$$

where $s_j(0)$ and $s_j(1)$ represent the starting and ending position of video segment j specified in number of frames, and $\widehat{p}_i$ is the calibrated score of frame i.

4.6. Sexual object detection

The third stage of our approach (refer to Fig. 1) operates on the video segments classified as being pornographic by the previous stage. For efficiency reasons, sexual object detection is not performed on each and every frame of the positive video segments; instead, video shot detection is performed first, followed by keyframe selection, and finally sexual object detection is performed only on the keyframes. The assumption behind this approach is that the types of sexual objects should stay reasonably constant within each video shot.[3]

We make use of the TransNet model of Souvcek and Lokovc (2020) for video shot detection. This network achieves efficient processing of large videos via its use of 3×3 dilated 3D CNN cells, with dilation rates of up to 8 in the temporal dimension, leading to an effective receptive field of 97 frames with just six cells.

Once a video segment s_j is partitioned into one or more video shots $\{vs_l\}$, we apply a keyframe selection scheme on each video shot. We evaluated a number of different schemes, eventually opting for a uniform keyframe sampling scheme:

$$\left\{ vs_l(0) + \frac{i}{K} vs_l(1) \right\}_{i=1}^{K-1} \tag{8}$$

where parameter K determines the number of keyframes extracted per video shot and $[vs_l(0), vs_l(1)]$ represents the start and end position of the l^{th} video shot.

Each keyframe is then fed to a sexual object detector; we employ the sexual object detector developed by Tabone et al. (2020), which utilises the YOLO v3 architecture (Redmon and Farhadi, 2018) for object detection and localisation. This detector was trained on a

[3] Here *video shot* is a film editing term, referring to a continuous piece of footage between two edits, transitions or cuts. In contrast the term *video segment* as used in this paper refers to a portion of a video (made up of consecutive video frames) that has been assigned the same classification label, e.g., a benign video segment, or a pornographic video segment. Typically a pornographic video segment can be decomposed into one or more video shots.

Table 3
Sexual object detection.

SOD class label	Severity ranking
Female breast	1
Female buttock	2
Male buttock	2
Female genitalia posing	2
Female genitalia sexually active	3
Male genitalia	3
Sex toys	3
Coitus	4
Anal	4

subset of the APD-2M image corpus (University of Leon), and can recognise the sexual objects listed in Table 3 (first column).

For each video segment, we keep track of the types of sexual objects detected and their occurrence frequencies, as well as their spatial locations in the frames and confidence scores.

4.7. Severity estimation and ranking of pornographic videos

In the final stage of our pipeline (see Fig. 1, lower part), we attempt to capture the semantic content of pornographic video segments based on the presence or absence of sexual objects. And we use this to assign a severity estimate to each segment for the purpose of ranking them in order of severity.

Our idea bears some similarity to Shen et al. (2018)'s use of 'semantic components', but while they used this information to aid the object inference process (via contextual information provided by the co-presence of sexual objects) and applied it solely to pornographic image data, we use the semantic information provided by the sexual objects for the purpose of ranking video data. We also employ a finer-grained sexual object classification, when compared to that of Shen et al. (2018).

There appears to be a lack of consensus in general when it comes to grading pornographic material based on the severity or gravity of its content (excluding the basic categorisation into 'softcore' and 'hardcore'). In the field of social sciences, the problem of ranking adult pornography is more concerned with studying the consumption patterns of such material, its arousal properties, addiction and related effects (Levitt, 1969; Laughton and Rensleigh, 2008). A number of taxonomies have been developed for the purpose of bibliographic access system (Dilevko and Gottlieb, 2002), but these typically lack the element of specifying the severity of pornographic material. And from a legal perspective, one can find works and legislation distinguishing extreme pornography (subject to criminal prosecution) from other types of legal pornography (The Crown Prosecution Service). In particular, in the area of child pornography, one finds the COPINE scale (Taylor et al., 2001) for ranking CSAM.

In this work, we constrain the problem of estimating the severity of the pornographic content of videos to a simple scheme based on the types of detected sexual objects. We adopt the severity scale given in Table 3, ranging from 1 (low severity) to 4 (more severe). We then rank video segments based on a combination of the severity class as described here, together with the video segment score given by Equation (7).

5. Experimental analysis

In this section we describe the experiments performed using our proposed solution, including model selection, training and optimisation. The TensorFlow library (Abadi et al., 2015) is used in the implementation of our DL models.[4]

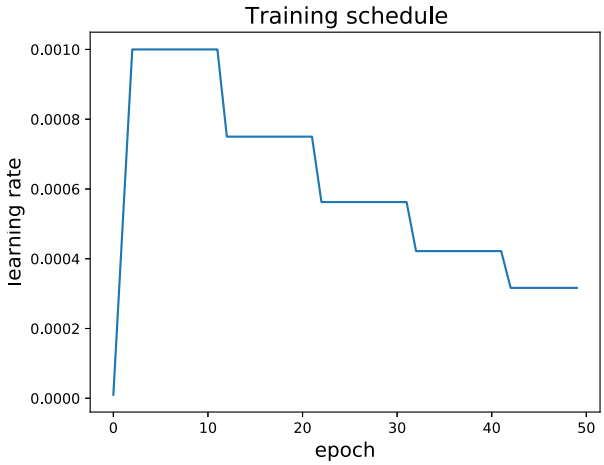

Fig. 5. The adopted training schedule with initial learning rate warm-up, followed by learning rate decay.

5.1. RNN training

The RNN is trained using mini-batch stochastic gradient descent (SGD) with Adam optimiser (Kingma and Ba, 2015). Batch size is set to 16. The learning rate is varied over 50 epochs according to the training schedule shown in Fig. 5. This schedule includes an initial warm-up period: for the first two epochs, we start with a learning rate much smaller than the target learning rate of 0.001, and then increase it linearly during the warm-up period until the target learning rate is reached. This strategy is known to help reduce early over-fitting (Liu et al., 2020). We also make use of early stopping as a form of implicit regularisation: we terminate training if the cross-entropy loss on the validation set does not improve over 10 epochs. The RNN is trained with a total of 15,910 video segments, partitioned into two folds (via stratified sampling), containing 80% (12,728 video segments) for training and 20% (3182 segments) for validation.

5.2. Model selection and hyperparameter tuning

Due to an inherent degree of non-determinism experienced when training deep neural networks, and in order to compare models in a statistically sound manner, multiple training runs need to be performed for each particular model and its set of hyper-parameter values, and then an accuracy measure or loss value determined from these multiple runs. A traditional way of performing this in the machine learning literature is via the use of *N*-fold cross validation, followed by a paired Student t-test for model comparison and selection (Vanwinckelen and Blockeel, 2012; Reimers and Gurevych, 2018).

A problem with this approach is that the overlapping use of the training data in the multiple training runs violates the assumption of each data sample being used only once. A different approach is to split the training data into separate folds, with each fold used for only one experiment. But this tends to become impractical when the dataset is relatively small, since the number of training examples in each fold will not be sufficient for training a deep learning model. As a result, we choose the notion of *stochastic dominance* for determining if a deep learning model is statistically superior to

[4] The authors will make the source code publicly available upon publication.

Table 4
Hyperparameter tuning.

Hyperparameter	Search range	Optimal value	ASO score
RNN hidden units	[16…128]	88	0.046
RNN dropout	{0.1, 0.2, 0.3, 0.4, 0.5, 0.6}	0.3	0.014
fc layer 1 hidden units	[2…48]	32	0.124
fc layer 2 hidden units	[2…24]	16	0.368
fc layer dropout	{0.1, 0.2, 0.3, 0.4, 0.5}	0.3	0.755
SGD batch size	$\{4*2n\}_{n=1}^{8}$	16	0.005

Table 5
Ablation studies.

Component	Training loss	Difference	ASO score
Full model	0.13628	–	–
without Batch normalisation	0.15857	−0.02229 (−16.4%)	1.0
without RNN dropout	0.16012	−0.02384 (−17.5%)	0.935
without fc dropout	0.18536	−0.04908 (−36.0%)	0.999

Table 6
Video-level classification results (APD-VIDEO dataset).

	Precision	Recall	F1 score	support
Benign	98.86%	96.11%	97.46%	175
Pornographic	96.67%	99.02%	97.83%	210
Accuracy			**97.66%**	385

Table 7
Frame-level classification results (APD-VIDEO dataset).

	Precision	Recall	F1 score	support
Benign	92.23%	88.68%	90.42%	984,700
Pornographic	94.57%	96.34%	95.45%	2,013,534
Accuracy			**93.83%**	2,998,234

another (Heathcote et al., 2010; del Barrio et al., 2018). In particular we use Dror et al. (2019)'s stochastic dominance method which is based on stochastic order (SO): They propose a relaxation method, called approximate stochastic order (ASO), which apart from telling whether the performance of a model is statistically superior to that of another (to the required significance level), it can still provide a relative ordering for models even when their performance difference is not statistically significant.

We perform a number of model selection experiments in order to choose amongst the different gated units available for the RNN: GRUs or long short-term memorys (LSTMs), the number of RNN layers, and whether each layer is bidirectional or not. In total we compare 12 different models using the ASO method described above, choosing the 2-layer bidirectional GRU as the best model (which is found to be statistically superior to 8 of the other models with 95% confidence, and statistically better than the 3 remaining models when approximate statistical ordering is considered).

Table 4 summarises the results of hyperparameter tuning, while Table 5 gives the results of the ablation studies we conducted. For both cases, multiple training runs were performed to measure the statistical significance of the optimal hyperparameter value (compared to the other values in the search range) and of the ablation studies. Due to lack of space, only the ASO scores are reported in the respective table. The results of the ablation studies show that dropout (both within the RNN and fully-connected layers), and to a slightly lesser extent batch normalisation, all contribute positively to the overall model.

6. Evaluation

We now evaluate our model against the datasets introduced in §3. Starting with the APD-VIDEO dataset, the results obtained for both video-level and frame-level classification are given in Tables 6 and 7 respectively, while Fig. 6 shows the confusion matrices. As expected, video-level accuracy is higher than frame-level accuracy, since for the former a majority-voting scheme is applied (as described in §4.4), which tends to smoothen the frame-wise results. From both confusion matrices, we can observe that the incidence of false positives (FPs) (1.8%, 3.7%) is slightly higher than that for the false negatives (FNs) (0.5%, 2.5%).

Unfortunately, we were not able to compare our results on the APD-VIDEO dateset with the state of the art, as none of the reviewed papers made their trained models or source code publicly available. We hope that by making our model and the APD-VIDEO dataset publicly available, can help in the evaluation of future systems on a larger and temporally richer dataset than what is currently available.

Moving on to the NPDI dataset, the currently de facto benchmark dataset, we train our model using the official training fold as provided with this dataset in order to perform comparative analysis with the state of the art. We utilise the full videos during training, and then test against the selected key frames forming part of the test fold.The results for the NPDI dataset are given in Table 8. As can be observed, our proposed solution achieves performance on par with the state of the art, with only a marginal 0.1% difference between our system and the best result achieved (Perez et al., 2017). In addition, compared to Perez et al.'s (2017) two-stream model with its roughly 13.6M trainable parameters, our baseline model is more efficient, with just 3.5M parameters. It is also worth pointing out that we train only the RNN portion of our model on the NPDI dataset, without performing any fine-tuning of the CNN layers (fine-tuning the last CNN layers could potentially increase the accuracy). This shows that our model generalises well.

Fig. 7 shows some representative output obtained from our system after performing sexual object detection and video severity ranking. In the case of Fig. 7 (c) some sexual objects are not detected mainly due to partial occlusion, resulting in an incorrect severity class assignment. While no sexual objects are detected in the keyframe of the video segment shown in Fig. 7 (a), and thus correctly assigned a severity class of 0, the segment has a non-zero (albeit small) ranking due to some of the framewise results being labelled as pornographic by the RNN. Overall, despite some sexual object mis-detections and severity class mis-categorisations, it can be seen that the video ranking scores reflect the true severity of the videos.

7. Investigations into model generalisation, memorisation and shortcut learning

7.1. Out-of-distribution (o.o.d.) generalisation test

As alluded to in §4.2, deep neural models have the capacity to memorise large datasets if given the chance (Zhang et al., 2017).

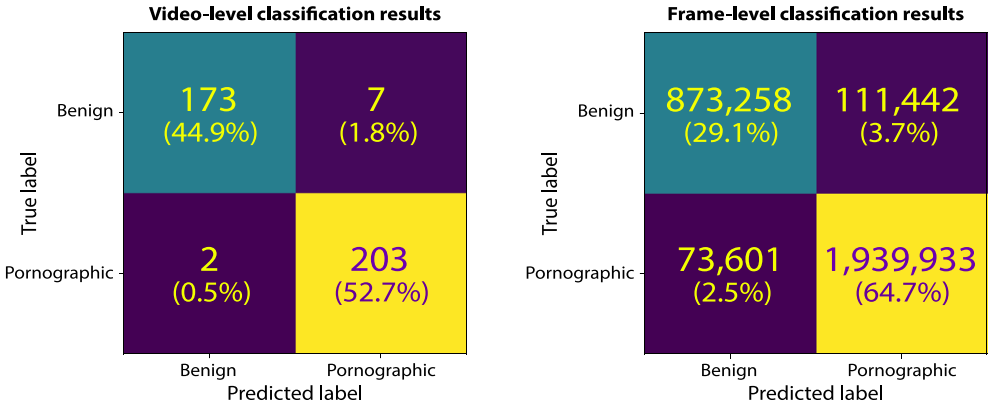

Fig. 6. Confusion matrices of classification results (APD-VIDEO dataset).

Table 8
Classification results (NPDI dataset).

Work	Approach	Params †	Accuracy
Moustafa (2015)	CNN ensemble (AGNet)	(72M)	94.1%
Ou et al. (2017)	Multi-context DL framework (DMCNet)	–	85.3%
Jung et al. (2017)	CNN ensemble	–	94.0%
Wehrmann et al. (2017)	CNN + RNN (LSTM)	(60M)	95.6%
Perez et al. (2017)	Two-stream CNN (image + motion features)	13.6M	**97.9**%
Shen et al. (2018)	CNN ensemble + Bayesian net	(23M)	94.7%
Da Silva and Marana (2019)	Spatiotemporal (VGG-3CD) CNNs	(11.1M)	95.1%
Our proposed solution	**CNN + RNN architecture**	**3.5M**	**97.8%**

† Number of trainable parameters: "—" means not specified in the original paper; figures in brackets are our estimates derived from the number of parameters of the base network used, e.g. the 60M parameters of ResNet-152 used by Wehrmann et al. (2017) as the convolutional base of their ACORDE model.

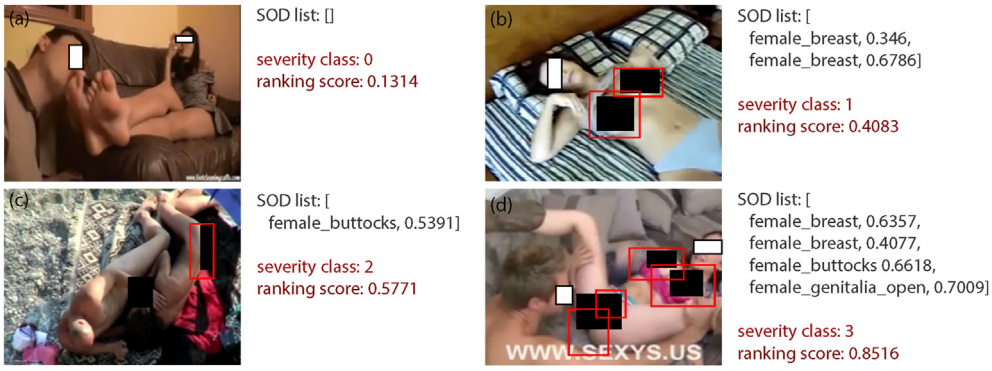

Fig. 7. Example results of sexual object detection and video severity ranking. On the left hand side is a representative keyframe from each video segment, and on the right hand side is the list of sexual objects (and their confidences) detected in the given keyframe (SOD list; also outlined in red). This is followed by the severity class and ranking score assigned to the video segment. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

This is mostly due to their large number of trainable parameters.

Arpit et al. (2017) suggest a spectrum for the learning abilities of deep neural models, ranging from complete memorisation at one end (as demonstrated by the experiments of Zhang et al. (2017) with randomised labels), to strong learning at the opposite end. As regards to the point on this spectrum where a particular deep learning system resides at, Arpit et al. (2017) attribute this to a combination of factors, including: the model architecture used, the training procedure adopted (including the use and types of regularisation), the nature and richness of the training data, and the amount of label noise present in the training data. Jiang et al. (2020) are in agreement with this interpretation, and show how deep learning models can memorise rare and irregular training samples, but then generalise across training samples that share common patterns or structures. In a series of experiments Carlini et al. (2019) try to measure how much models generalise (quantified in terms of model perplexity - how useful a model is), versus how much they memorise of the training data (quantified in terms of model exposure). A main finding of their work is that memorisation is not overfitting: memorisation starts while a model is still actively learning (i.e., it has not yet started to overfit).

While there is ongoing research into trying to come up with metrics or estimates to measure the generalisability of a model (e.g., C-score (Jiang et al., 2020), model perplexity and exposure (Carlini et al., 2019), etc.), Geirhos et al. (2020) propose the use of an out-of-distribution (o.o.d.) dataset as the ideal test for generalisation. This is in contrast to the test set coming from the same dataset, which can be termed as an independent and identically distributed

Table 9
o.o.d. generalisation test results.

Test type	Training set	Test set	Accuracy
i.i.d.	APD-VIDEO	APD-VIDEO	97.76%
o.o.d.	APD-VIDEO	NPDI	94.12% **(-3.54%)**

(i.i.d.) test.

While we can not attest as to the position where our model resides on the memorisation-generalisation continuum, we believe that a number of techniques described in the previous sections, should help minimise the amount of memorisation in our model. These include the use of label smoothing to address noisy labels (one of the factors mentioned by Arpit et al. (2017)), as well as a number of implicit and explicit regularisations used in our model and its training (early stopping, dropout, batch normalisation, etc.).

We also follow Geirhos et al. (2020)'s suggestion of performing an o.o.d. generalisation test. Since to the best of our knowledge there are no standard benchmark o.o.d. datasets in the domain of pornography detection, we employ a scheme whereby we train our model on our APD-VIDEO dataset (§3.2) and then evaluate its performance on the NPDI dataset (§3.1). A cursory glance at these two datasets strongly suggests that there is a distribution shift between the two, thus in our opinion qualifying the NPDI dataset as an o.o.d. test. In particular, the majority of the NPDI benign videos contain text or come from Portuguese-speaking sources, while our pornography dataset is mainly sourced from English-speaking sources.[5] The NPDI benign videos also exhibit a higher incidence of children, while a number of the NPDI positive videos consist of animated movies, something which is missing from our pornography dataset.

From Table 9, we can observe that although the accuracy of the o.o.d. test is lower than that for the i.i.d. test, at 94.12% it is still high. To put this difference in perspective, other studies have reported drops in accuracy of more than 10% where models failed to generalise to external data (Zech et al., 2018; Mårtensson et al., 2020). We can therefore reasonably conclude that our model is able to generalise satisfactorily to o.o.d. data.

Fig. 8 gives some qualitative results, with mis-classified frames highlighted in red. Of particular interest are the results obtained for the animated videos (Fig. 8 (b) and (f)). When presented with these o.o.d. samples, the model is still able to classify several parts of these videos correctly. The probability chart also shows that the confidence level of the model is lower than for samples it has been trained on (e.g. Fig. 8 (d)), which indicates that our model is well-calibrated.

7.2. Shortcut learning tests

We next investigate whether the good performance results of our model are really due to shortcut learning or not. Shortcut learning (Geirhos et al., 2020) is the process via which a deep neural network identifies the simplest solution (a 'shortcut') for classifying the given input data. This often takes the form of 'cheating', i.e., the model exploits unintended signals or confounding information in the data instead of learning the true patterns. When learning in such a manner, a network may demonstrate deceptively good results. But then it can exhibit unintuitive failures when faced with o.o.d. cases.

Shortcut opportunities can come from background cues or scene biases in the dataset being used to recognise the primary objects (He et al., 2016; Zhu et al., 2017; Beery et al., 2018). Cues arising from the source or acquisition/preparation method of the data samples might also creep into the dataset (Dawson et al., 2019; Mårtensson et al., 2020). Alternatively, the model can learn from the presence or absence of ancillary tokens present in images; for example, from the placement of metal tokens in the corners of radiographs (Zech et al., 2018), or learning that skin lesions with a ruler placed next to them are more likely to be malignant (Narla et al., 2018).

A clear shortcut opportunity that exists in pornography datasets is the presence of logos indicating the pornographic sites from where the videos are sourced from. These logos typically appear in a corner of the video frames and remain present throughout the duration of the videos. Such logos appear in both datasets used in our experimental work. We therefore perform a number of investigations to determine whether our model is utilising these unintended cues during classification.

We first extract logos from both the pornographic and benign samples of the APD-VIDEO dataset. We apply a simple automatic procedure to do this, relying on the temporal stability of the pixel values where the logos are located: first a motion energy image (MEI) is accumulated from all the video frames of a video; then an adaptive thresholding method is applied to find the temporally-stable pixels (low MEI values), followed by morphological operations, and finally extracting contiguous regions conditioned on some basic region size filtering. A final manual-based checking is performed to eliminate incorrectly extracted logos. Table 10 gives a breakdown of the number of extracted logos from the videos of our dataset, indicating a prevalence for logos in pornographic videos; and Fig. 9 shows a sample of the extracted logos for both classes.

For our first experiment, we generate new test videos by superimposing the logos extracted from pornographic videos on to the benign videos and vice versa. Fig. 10 shows such an example. The logos are placed in the video frame at the same original position in which they were found. The aim of this experiment is to investigate whether benign video frames are incorrectly classified by our model as being pornographic solely based on the presence of the superimposed logos mormally found in pornographic videos. A large reduction in accuracy would indicate that the model is taking a shortcut by memorising logos.

In our second experiment, we investigate whether the model is basing its classification decision on the spatial placement of the logos in the video frames (rather than, or in conjunction to, memorising the logos). We noticed that there is a certain spatial bias in where logos appear in the video frames depending on whether they are pornographic or benign – this is evident from the heatmaps shown in Fig. 11 (a) and (b). To test whether this bias is being utilised by our model, we create a second set of synthetic videos where this time, we mask out the area where logos are likely to be found – see Fig. 11 (c) for an example.

For both experiments, 200 videos were generated and the original unaltered videos are used as the control experiment. Training of the RNN-based model is conducted as described in §5.1, and the results obtained are given in Table 11.

We can observe that the overall reduction in accuracy is minimal for both experiments. In the case of the second experiment, we attribute the slightly larger decrease in accuracy to the fact that more of the image content is lost by the masking operation. Fig. 12 shows the difference in accuracies for each individual video used in the logo superimposition experiment. We notice that most of the variations occur for videos where the model was not confident in its prediction in the first place. And surprisingly, a number of videos exhibit a net gain in accuracy.

These results lead us to conclude that our model is not relying

[5] While our proposed solution does not use the audio stream of videos, the difference in language is apparent visually in many of the videos in terms of logo texts, titles, sub-titles and captions.

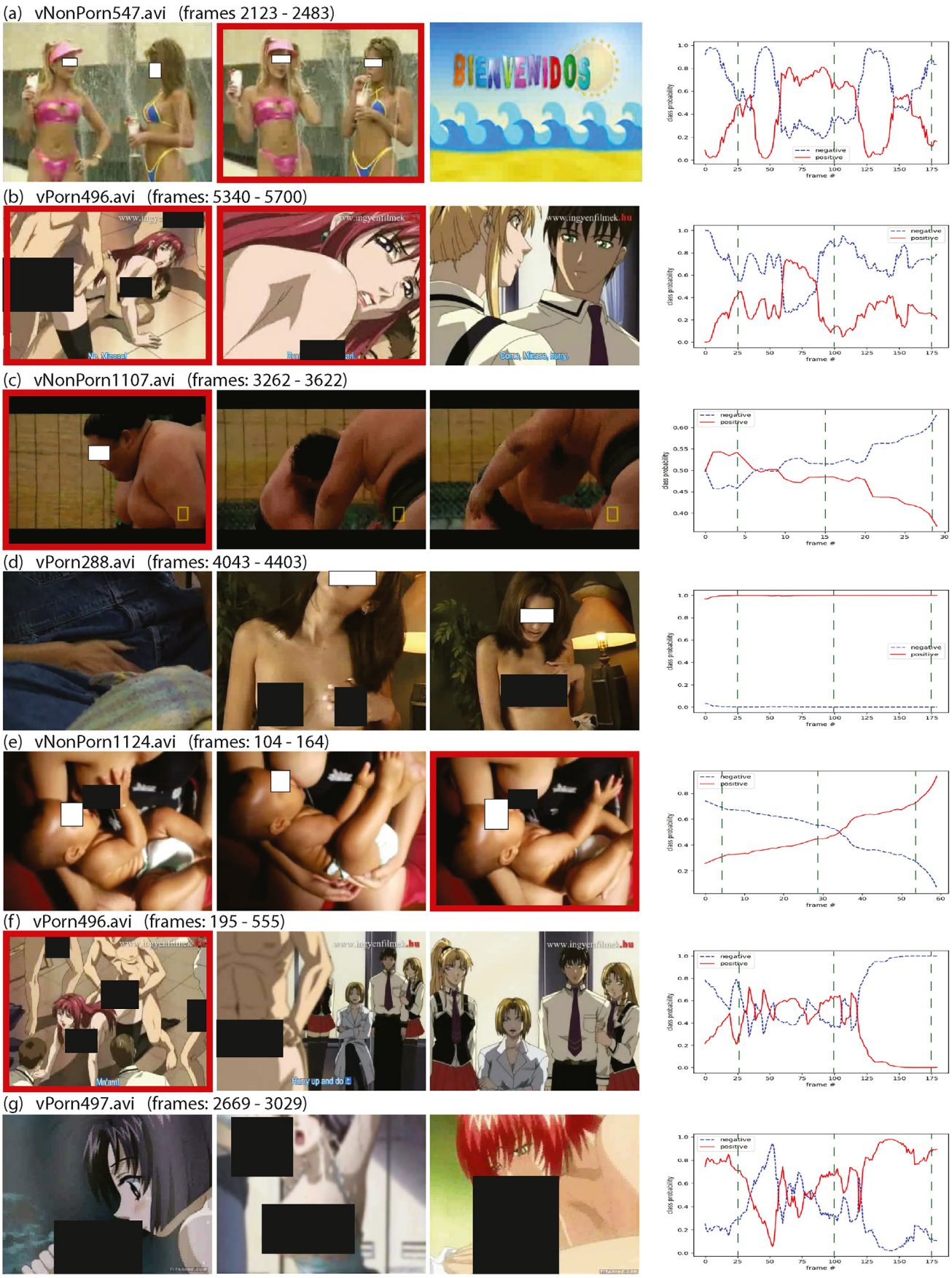

Fig. 8. Result samples from the NPDI o.o.d. test. Three frames are shown for each video segment, while the plots on the right of the frames display the model's probability output over the entire segment, with the position of the three frames indicated by the dashed vertical lines. Mis-labeled video frames are shown with a red border. Several of these video segments (e.g., the animated videos) are considered as o.o.d. samples, since the model was trained with a dataset that lacked such examples. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 10
Distribution of logos in the APD-VIDEO dataset.

Class	Total video segments	Video segments with logos
Pornographic	8201	4943 (60.27%)
Benign	7426	1230 (16.56%)

on the presence of logos or their position in the video frame to classify videos. Possible reasons may include the diversity of logos present in the dataset, and that there are other elements in the videos apart from logos that are temporally stable over long portions of the video such as background elements – we think that this may make logos less of a prominent signal for the network to exploit as a shortcut during its training.

8. Explainability

To help us investigate further our deep neural model's behaviour, we ran a number of explainability tests. More specifically, we apply the occlusion sensitivity method (Zeiler and Fergus, 2014) on the CNN base of our model, in order to determine the region(s) in the video frame that our model is paying attention to when deciding between pornographic and benign material. Occlusion sensitivity uses a rectangular window to hide a portion of the video frame and measure the resulting change in the confidence of the classification result. By sliding this window across the frame, a heatmap is generated, like the examples shown in Fig. 13.

Fig. 13 (a)–(c) highlight the regions our model focuses on to generate correct predictions. In particular, (b) shows the behaviour of our model when dealing with almost fully-clothed persons engaged in sexual activity. Fig. 13 (d)–(f) show incorrectly-classified examples and the regions which influenced the model's decision. In the case of (d) and (e), the model appears to focus mostly on skin and skin-coloured clothing respectively. While in the case of Fig. 13 (f) it is harder to relate the areas of interest with the classification decision taken.

9. Inference speed optimisation

9.1. SOD execution

As mentioned in §4.6, for efficiency reasons SOD is not performed on each video frame, but instead it is run on a small set of

Fig. 9. A sample of logos extracted from (left) pornographic and (right) benign videos.

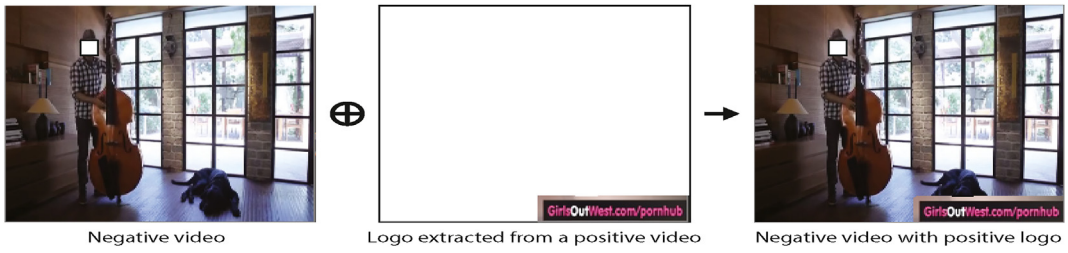

Fig. 10. An example of a synthetic test video created by superimposing a logo belonging to the opposite class.

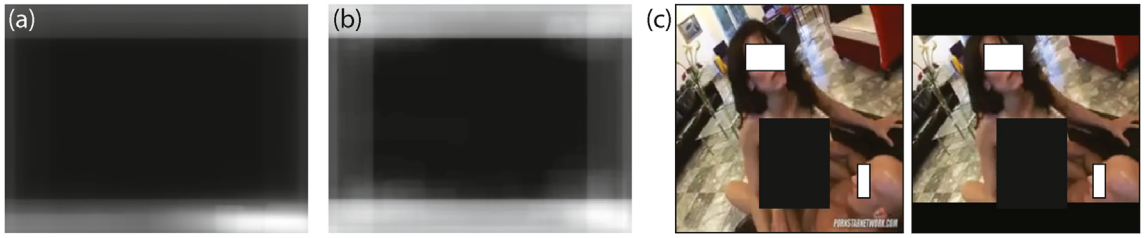

Fig. 11. Heatmaps indicating the likelihood of having a logo at the given position for (a) pornographic and (b) benign videos in the APD-VIDEO dataset, showing a prevalence for logos to be placed in the lower-right corner for pornographic videos. (c) An example of a synthetic video frame generated by masking out the area where logos tend to be found in pornographic videos (left: the original frame with a logo (white text) in the bottom-right corner; right: video frame with masked out areas).

Table 11
Results of shortcut learning investigations.

Experiment	Frame-level accuracy	Diff.	Video-level accuracy	Diff.
Control test (200 videos)	91.62%	–	93.5%	–
Superimposed logos	90.86%	**−0.76%**	92.97%	**−0.53%**
Logo areas masked out	89.53%	**−2.09%**	92.69%	**−0.81%**

keyframes extracted from each video shot of pornographic segments. In our next experiment, we measure how many sexual object detections are missed by our approach, comparing against the number of detections obtained when running against each video frame as a baseline. Fig. 14 shows the results obtained. We can observe that for many of the SOD classes the reduction in detections is not overly significant. To put this into perspective, in these same tests, we determined that on average, SOD is performed on just 9.0% of the video frames (median of 5.3%) which offers a significant speedup in execution.

9.2. Model pruning

Finally we report on investigations into ways of improving the efficiency of model inference. We adopt a model pruning technique (Zhu and Gupta, 2018): during training, based on a pre-determined sparsity connection setting, weights which are less salient than others are gradually reduced to 0, thus enabling their pruning from the model once the model is trained. This pruning technique

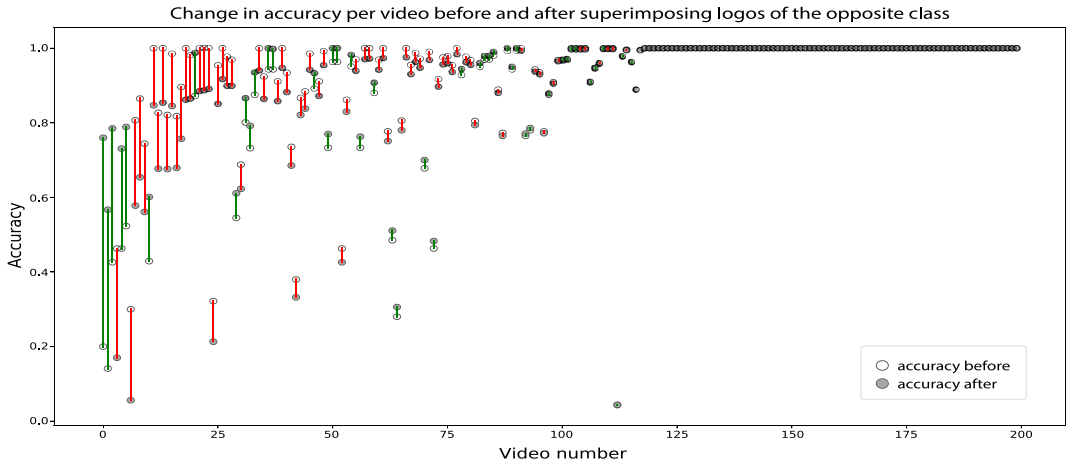

Fig. 12. Differences in accuracy per video for the logo superimposition experiment. Differences shown in green indicate a net gain in accuracy; red indicates a reduction in accuracy. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

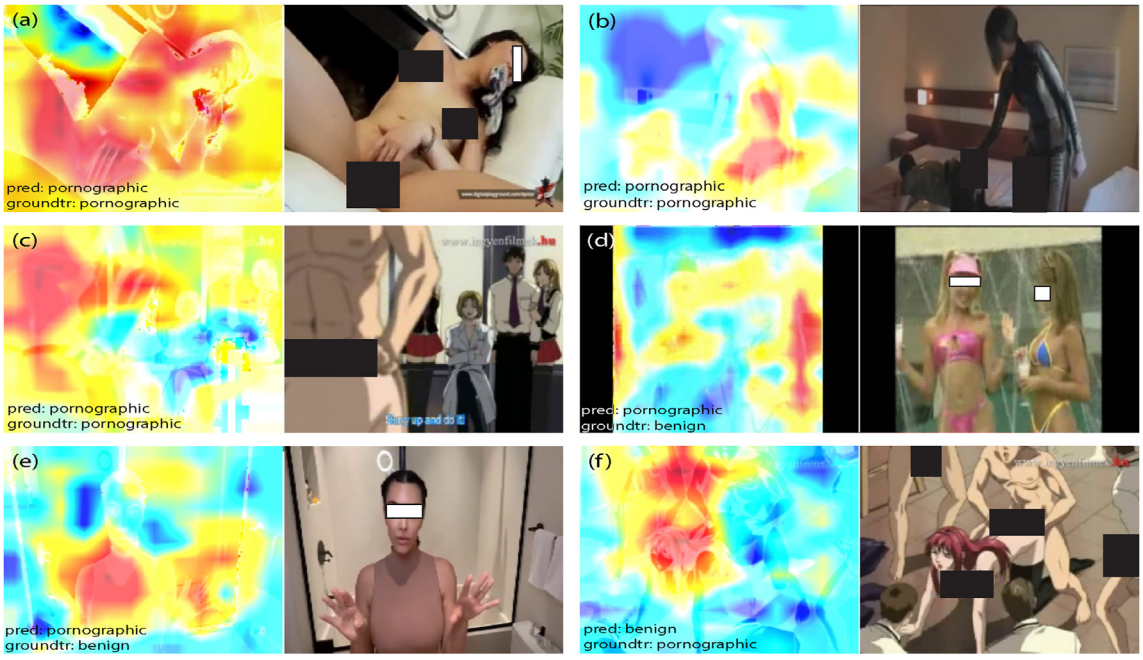

Fig. 13. Some results from the model explainability investigations. (Left) Normalised heatmaps obtained by running the occlusion sensitivity method, with red indicating the areas of most interest to our model; (Right) the original video frame corresponding to the heatmap shown on the left. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

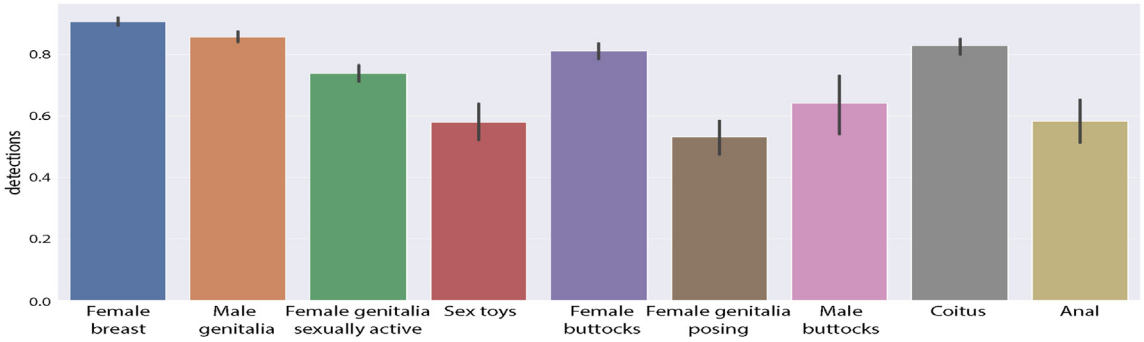

Fig. 14. The reductions in detections when running SOD on key frames, compared to when running SOD on all frames of a video (the baseline, 100%). For a number of classes, the difference is minimal. The vertical black lines indicate variations (95% confidence interval) in the results across the videos used in this test.

Table 12
Model pruning results.

Model	Parameters	Sparsity	Accuracy	Speedup
Baseline model	3,465,378	–	95.45%	–
Pruned model	3,370,772	40%	95.38% (−0.07%)	×2.1
	3,347,121	50%	95.10% (−0.35%)	×2.4
	3,323,469	60%	94.98% (−0.47%)	×2.6
	3,299,764	70%	94.81% (−0.64%)	×2.9
	3,276,164	80%	94.02% (−1.43%)	×3.3

normally gives rise to a more efficient model (less parameters), often accompanied with minimal degradation in performance.

While several different methods exist for determining the saliency of the weight parameters (Zhu and Gupta, 2018; Blalock et al., 2020), we opt for a magnitude-based weight pruning strategy (Janowsky, 1989) mainly due to its computational efficiency. Table 12 shows the resulting speedup and loss of performance for increasing levels of sparsity. We can observe minimal degradation in performance, even for large values of sparsity, while at the same time getting a speed improvement during inference of more than double when compared to the unpruned model.

10. Conclusion

In this paper, we proposed a pornographic detection system consisting of a CNN for automatic feature extraction, followed by a bi-directional GRU RNN. We described how our system can be used for both video-level labelling as well as for localising pornographic content within videos. Given pornographic video segments, we described an efficient method for finding sexual objects within the segments, and how the types of the detected sexual objects can be used to generate an estimate of the severity ('harmfulness') of the pornographic content. This estimate can be utilised for ranking videos based on their severity. We evaluated our proposed system against a benchmark dataset, achieving results on par with the state of the art. Investigations into model generalisability, shortcut learning, and explainability, suggest that our model is using strong learning.

As future work, we plan to investigate the use of multiple modalities (such as the audio stream or optical flow), as well as better and more semantically meaningful ways of estimating the severity of pornographic content.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research has been funded with support from the European Commission under the 4NSEEK project with Grant Agreement 821966. This publication reflects the views only of the authors, and the European Commission cannot be held responsible for any use which may be made of the information contained therein.

References

Abadi, M., et al., 2015. TensorFlow: large-scale machine learning on heterogeneous systems. URL: www.tensorflow.org/.
Al Dahoul, N., Abdul Karim, H., et al., 2021. Transfer detection of YOLO to focus CNN's attention on nude regions for adult content detection. Symmetry 13.
Al Nabki, M., Fidalgo, E., et al., 2020. Evaluating Performance of an Adult Pornography Classifier for Child Sexual Abuse Detection. CoRR.
Anda, F., Le-Khac, N.A., Scanlon, M., 2020. DeepUAge: improving underage age estimation accuracy to aid CSEM investigation. Forensic Sci. Int.: Digit. Invest. 32.
Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B., 2014. 2D Human Pose Estimation: New Benchmark State of the Art Analysis. CVPR.
Arpit, D., Jastrzebski, S., Ballas, N., et al., 2017. A closer look at memorization in deep networks. In: Proc. 34th Int. Conf. On Machine Learning, 70. JMLR, pp. 233–242. PMLR.
Avila, S., Thome, N., Cord, M., Valle, E., de Araújo, A., 2013. Pooling in image representation: the visual codeword point of view. Comput. Vis. Image Understand. 117.
Beery, S., Horn, G.V., Perona, P., 2018. Recognition in terra incognita. In: Computer Vision - ECCV Proc. Springer, pp. 472–489.
Behrad, A., Salehpour, M., Ghaderian, M., et al., 2012. Content-based obscene video recognition by combining 3D spatiotemporal and motion-based features. EURASIP J. Image Video Process. 2012, 1–17. https://doi.org/10.1186/1687-5281-2012-23.
Blalock, D., Ortiz, J.J.G., Frankle, J., Guttag, J., 2020. In: What Is the State of Neural Network Pruning?.
Böken, B., 2021. On the appropriateness of platt scaling in classifier calibration. Inf. Syst. 95, 101641. https://doi.org/10.1016/j.is.2020.101641.
Caetano, C., Avila, S., Guimarães, S., Araújo, A.d.A., 2014. Representing local binary descriptors with bossanova for visual recognition. In: Proc. 29th Annual ACM Symposium on Applied Computing, pp. 49–54. https://doi.org/10.1145/2554850.2555058.
Carlini, N., Liu, C., Erlingsson, U., Kos, J., Song, D., 2019. The secret sharer: evaluating and testing unintended memorization in neural networks. In: Proc. 28th USENIX Conf. On Security Symposium.
Chen, P., Liao, B., Chen, G., Zhang, S., 2019. Understanding and Utilizing Deep Neural Networks Trained with Noisy Labels. CoRR.
Cho, K., van Merrienboer, B., Gulcehre, C., et al., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics.
da Silva, M.V., Marana, A.N., 2019. Spatiotemporal CNNs for pornography detection in videos. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Springer Int. Publ.
Dawson, M., Zisserman, A., Nellåker, C., 2019. From same photo: cheating on visual kinship challenges. In: ACCV. Springer, pp. 654–668.
de Freitas, P.V.A., Mendes, P.R.C., dos Santos, G.N.P., Busson, A.J.G., Livio Guedes, Álan, Colcher, S., Milidiú, R.L., 2019. A Multimodal CNN-Based Tool to Censure Inappropriate Video Scenes arXiv. arXiv:1911.03974.
del Barrio, E., Cuesta-Albertos, J., Matrán, C., 2018. In: Some Indices to Measure Departures from Stochastic Order.

Dilevko, J., Gottlieb, L., 2002. Deep Classification: Pornography, Bibliographic Access, and Academic Libraries, vol. 26. Library Collections, Acquisitions, & Technical Services, pp. 113–139. https://doi.org/10.1080/14649055.2002.10765838.
Dror, R., Shlomov, S., Reichart, R., 2019. Deep dominance - how to properly compare deep neural models. In: Proc. 57th Annual Meeting of the Association for Computational Linguistics, pp. 2773–2785. https://doi.org/10.18653/v1/P19-1266.
Everingham, M., Van Gool, L., Williams, C., Winn, J., Zisserman, A., 2012. The PASCAL Visual Object Classes Challenge (VOC) Results.
Gangwar, A., Fidalgo, E., Alegre, E., González-Castro, V., 2017. Pornography and child sexual abuse detection in image and video: a comparative evaluation. In: Proc. ICDP, pp. 37–42. https://doi.org/10.1049/ic.2017.0046.
Gangwar, A., González-Castro, V., Alegre, E., Fidalgo, E., 2021. AttM-CNN: attention and metric learning based CNN for pornography, age and child sexual abuse (CSA) detection in images. Neurocomputing 445.
Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A., 2020. Shortcut learning in deep neural networks. Nat. Mach. Intell. 2. https://doi.org/10.1038/s42256-020-00257-z.
Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. In: Proc. ICML, JMLR.
He, Y., Shirakabe, S., Satoh, Y., Kataoka, H., 2016. Human Action Recognition without Human. In: ECCV. Springer, Cham, pp. 11–17.
He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., Li, M., 2019. Bag of tricks for image classification with convolutional neural networks. In: Proc. CVPR. IEEE, pp. 558–567. https://doi.org/10.1109/CVPR.2019.00065.
Heathcote, A., Brown, S., Wagenmakers, E., Eidels, A., 2010. Distribution-free tests of stochastic dominance for small samples. J. Math. Psychol. 54, 454–463. https://doi.org/10.1016/j.jmp.2010.06.005.
Islam, M., Watters, P., Mahmood, A., Alazab, M., 2019. Toward Detection of Child Exploitation Material: A Forensic Approach. Springer.
Janowsky, S.A., 1989. Pruning versus clipping in neural networks. Phys. Rev. A 39, 6600–6603. https://doi.org/10.1103/PhysRevA.39.6600.
Jiang, Z., Zhang, C., Talwar, K., Mozer, M.C., 2020. Characterizing Structural Regularities of Labeled Data in Overparameterized Models arXiv.
Jung, J., Makhijani, R., Morlot, A., 2017. Combining CNNs for Detecting Pornography in the Absence of Labeled Training Data.
Karavarsamis, S., Ntarmos, N., Blekas, K., Pitas, I., 2013. Detecting pornographic images by localizing skin ROIs. Int. J. Digital Crime Forensics (IJDCF) 5.
Karimi, D., Dou, H., Warfield, S.K., Gholipour, A., 2020. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Med. Image Anal. 65. https://doi.org/10.1016/j.media.2020.101759.
Kelly, W., Donnellan, A., Molloy, D., 2008. Screening for objectionable images: a review of skin detection techniques. In: MVIP. https://doi.org/10.1109/IMVIP.2008.21.
Kingma, D.P., Ba, J., 2015. Adam: a method for stochastic optimization. In: 3rd Int. Conf. Learning Representations. ICLR.
Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems, pp. 6402–6413.
Laranjeira da Silva, C., Macedo, J., Avila, S., dos Santos, J., 2022. Seeing without looking: analysis pipeline for child sexual abuse datasets. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. ACM, NY, USA, pp. 2189–2205.
Laughton, P., Rensleigh, C., 2008. Investigating on-line pornography at the university of johannesburg. J. Inf. Manag. 10 (2). https://doi.org/10.4102/sajim.v10i2.314.
Lee, H.E., Ermakova, T., Ververis, V., Fabian, B., 2020. Detecting child sexual abuse material: a comprehensive survey. Forensic Sci. Int.: Digit. Invest. 34.
Levitt, E.E., 1969. Pornography: some new perspectives on an old problem. J. Sex. Res. 5, 247–259.
Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: common objects in context. CoRR abs/ 1405, 0312.
Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Int. Conf. Computer Vision (ICCV). https://doi.org/10.1109/ICCV.2017.324.
Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J., 2020. On the variance of the adaptive learning rate and beyond. In: International Conference on Learning Representations.
Macedo, J., Costa, F., dos Santos, J.A., 2018. A benchmark methodology for child pornography detection. In: 31st SIBGRAPI Conference on Graphics, Patterns and Images, pp. 455–462. https://doi.org/10.1109/SIBGRAPI.2018.00065.
Mallmann, J., Santin, A.O., Viegas, E.K., dos Santos, R.R., Geremias, J., 2020. PPCensor: architecture for real-time pornography detection in video streaming. Future Generat. Comput. Syst. 112, 945–955. https://doi.org/10.1016/j.future.2020.06.017.
Mårtensson, G., Ferreira, D., Others, 2020. The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. Med. Image Anal. 66, 101714. https://doi.org/10.1016/j.media.2020.101714.
McKee, A., Byron, P., Litsou, K., Ingham, R., 2020. An interdisciplinary definition of pornography: results from a global delphi panel. Arch. Sex. Behav. 49, 1085–1091.
More, M.D., Souza, D.M., Wehrmann, J., Barros, R.C., 2018. Seamless nudity censorship: an image-to-image translation approach based on adversarial training. In: Int. Joint Conf. Neural Networks (IJCNN).
Moreira, D., Avila, S., Perez, M., Moraes, D., Testoni, V., Valle, E., Goldenstein, S., Rocha, A., 2016. Pornography classification: the hidden clues in video space–time. Forensic Sci. Int. 268, 46–61. https://doi.org/10.1016/j.forsciint.2016.09.010.
Moustafa, M.N., 2015. Applying Deep Learning to Classify Pornographic Images and Videos. PSIVT.
Mozafari, A., Gomes, H., Leão, W., Janny, S., Gagné, C., 2018. Attended Temperature Scaling: A Practical Approach for Calibrating Deep Neural Networks arXiv: Learning.
Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P.H., Dokania, P.K., 2020. Calibrating Deep Neural Networks Using Focal Loss.
Muller, R., Kornblith, S., Hinton, G., 2019. When does label smoothing help? Adv. In: Neural Information Processing Systems (NIPS).
Narla, A., Kuprel, B., Sarin, K., Novoa, R., Ko, J., 2018. Automated classification of skin lesions: from pixels to practice. J. Invest. Dermatol. 138, 2108–2110.
Niculescu-Mizil, A., Caruana, R., 2005. Predicting good probabilities with supervised learning. In: Int. Conf. On Machine Learning. ACM.
Nixon, J., Dusenberry, M., Zhang, L., Jerfel, G., Tran, D., 2019. Measuring Calibration in Deep Learning, 01685. ArXiv:1904.
Oronowicz-Jaskowiak, W., 2018. Classification of seven types of legal pornography using a neural network. Przegl Seks 1.
Ou, X., Ling, H., Yu, H., Li, P., Zou, F., Liu, S., 2017. Adult image and video recognition by a deep multicontext network and fine-to-coarse strategy. ACM Trans. Intell. Syst. Technol. 8. https://doi.org/10.1145/3057733.
Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Łukasz, Hinton, G., 2017. Regularizing Neural Networks by Penalizing Confident Output Distributions. ICLR.
Perez, M., Avila, S., Moreira, D., et al., 2017. Video pornography detection through deep learning techniques and motion information. Neurocomputing 230, 279–293. https://doi.org/10.1016/j.neucom.2016.12.017.
Platt, J.C., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers. MIT Press, pp. 61–74.
Qamar Bhatti, A., Umer, M., et al., 2018. Explicit content detection system: an approach towards a safe and ethical environment. Appl. Comput. Intel. Soft Comput. https://doi.org/10.1155/2018/1463546.
Redmon, J., Farhadi, A., 2018. YOLOv3: an Incremental Improvement arXiv.
Reimers, N., Gurevych, I., 2018. Why Comparing Single Performance Scores Does Not Allow to Draw Conclusions about Machine Learning Approaches arXiv.
Rolnick, D., Veit, A., Belongie, S., Shavit, N., 2018. Deep Learning Is Robust to Massive Label Noise arXiv:1705.10694.
Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C., 2019. MobileNetV2: Inverted Residuals and Linear Bottlenecks arXiv.
Shen, R., Zou, F., Song, J., Yan, K., Zhou, K., 2018. EFUI: an ensemble framework using uncertain inference for pornographic image recognition. Neurocomputing 322, 166–176. https://doi.org/10.1016/j.neucom.2018.08.080.
Short, M.B., Black, L., Smith, A.H., et al., 2012. A review of internet pornography use research: methodology and content from the past 10 years. Cyberpsychol., Behav. Soc. Netw. 15.
Song, K., Kim, Y.S., 2020. An enhanced multimodal stacking scheme for online pornographic content detection. Appl. Sci. https://doi.org/10.3390/app10082943.
Song, H., Kim, M., Park, D., Lee, J.G., 2020. Learning from Noisy Labels with Deep Neural Networks: A Survey, 08199 arXiv:2007.
Souvcek, T., Lokovc, J., 2020. Transnet V2: an Effective Deep Network Architecture for Fast Shot Transition Detection.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: CVPR. https://doi.org/10.1109/CVPR.2016.308.
Tabone, A., Bonnici, A., Cristina, S., Farrugia, R.A., Camilleri, K.P., 2020. Private Body Part Detection Using Deep Learning. ICPRAM.
Tabone, A., Camilleri, K.P., Bonnici, A., Cristina, S., Farrugia, R.A., Borg, M., 2021. Pornographic content classification using deep-learning. In: DocEng '21: Proceedings of the 21st ACM Symposium on Document Engineering.
Taylor, M., Holland, G., Quayle, E., 2001. Typology of paedophile picture collections. Police J. 74, 97–107.
The Crown Prosecution Service. Extreme pornography. URL: https://www.cps.gov.uk/legal-guidance/extreme-pornography.
University of Leon, . Adult Pornography Dataset - 2M (APD-2M). URL: http://gvis.unileon.es/dataset/apd-2m/.
University of Malta, 2021. The Adult Pornography Video Dataset (APD-VIDEO). and Control Engineering, University of Malta online. URL: http://gvis.unileon.es/dataset/the-adult-pornography-video-dataset-apd-video/. Dept. of Systems.
Valle, E., de Avila, S., da Luz Jr., A., de Souza, F., de Miranda Coelho, M., de Albuquerque Araújo, A., 2011. Content-based filtering for video sharing social networks. CoRR abs 1101, 2427.
Vanwinckelen, G., Blockeel, H., 2012. On Estimating Model Accuracy with Repeated Cross-Validation, pp. 39–44.
Vitorino, P., Avila, S., Perez, M., Rocha, A., 2018. Leveraging deep neural networks to fight child pornography in the age of social media. J. Vis. Commun. Image Represent. 50. https://doi.org/10.1016/j.jvcir.2017.12.005.
Wehrmann, J., Simões, G., Barros, R., Cavalcante, V., 2017. Adult content detection in videos with convolutional and recurrent neural networks. Neurocomputing. https://doi.org/10.1016/j.neucom.2017.07.012.
Yousaf, K., Nawaz, T., 2022. A deep learning-based approach for inappropriate content detection and classification of youtube videos. IEEE Access 10,

16283–16298.
Zech, J.R., Badgeley, M.A., Liu, M., et al., 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 15.
Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: Proc. ECCV, 2014.
Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2017. Understanding Deep Learning Requires Rethinking Generalization. ICLR.
Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D., 2017. In: Temporal Action Detection with Structured Segment Networks. ICCV.
Zhelonkin, D., Karpov, N., 2020. Training effective model for real-time detection of nsfw photos and drawings. In: Analysis of Images, Social Networks and Texts. Springer Int. Publ., Cham, pp. 301–312.
Zheng, H., Daoudi, M., 2004. Blocking adult images based on statistical skin detection. ELCVIA - Electron. Lett. Comput. Vis. Image Anal. 4.
Zhu, M., Gupta, S., 2018. To Prune, or Not to Prune: Exploring the Efficacy of Pruning for Model Compression. In: ICLR.
Zhu, Z., Xie, L., Yuille, A., 2017. Object recognition with and without objects. In: Proc. Int. Joint Conf. Artificial Intelligence. IJCAI. https://doi.org/10.24963/ijcai.2017/505.